Multiview Graph Learning for single-cell RNA sequencing data

Abdullah Karaaslanli^{1,*}, Satabdi Saha^{2,*}, Selin Aviyente¹ and Tapabrata Maiti²

¹*Department of Electrical and Computer Engineering,
Michigan State University, East Lansing, MI, USA*

²*Department of Statistics and Probability,
Michigan State University, East Lansing, MI, USA*

Abstract

Characterizing the underlying topology of gene regulatory networks is one of the fundamental problems of systems biology. Ongoing developments in high throughput sequencing technologies has made it possible to capture the expression of thousands of genes at the single cell resolution. However, inherent cellular heterogeneity and high sparsity of the single cell datasets render void the application of regular Gaussian assumptions for constructing gene regulatory networks. Additionally, most algorithms aimed at single cell gene regulatory network reconstruction, estimate a single network ignoring group-level (cell-type) information present within the datasets. To better characterize single cell gene regulatory networks under different but related conditions we propose the joint estimation of multiple networks using multiview graph learning (mvGL). The proposed method is developed based on recent works in graph signal processing (GSP) for graph learning, where graph signals are assumed to be smooth over the unknown graph structure. Graphs corresponding to the different datasets are regularized to be similar to each other through a learned consensus graph. We further kernelize mvGL with the kernel selected to suit the structure of single cell data. An efficient algorithm based on prox-linear block coordinate descent is used to optimize mvGL. We study the performance of mvGL using synthetic data generated with a diverse set of parameters. We further show that mvGL successfully identifies well-established regulators in a mouse embryonic stem cell differentiation study and a cancer clinical study of medulloblastoma.

1. Introduction

Gene expression arises from an underlying network of regulatory interactions between transcription factors, co-factors and signaling molecules [49, 58]. Elucidating the topology of this underlying network is a major goal in modern computational biology, as it will allow researchers to discover pivotal factors that can help differentiate between the phenotypes of varying cell states [15, 41]. Advances in next generation sequencing technologies has made it possible to profile the transcriptomes of individual cells, hence capturing expression levels of thousands of genes at a cellular resolution. This has further enabled scientists to discover natural grouping among cells and define cell-types on the basis of the transcriptome [8, 19, 18]. The variation present between these cell-types holds the key to inferring how genes transcriptionally regulate each other and how their expressions and interactions change across cell subgroups. Accurate inference of cell-type specific GRNs will lead to a deeper understanding of subgroup specific transcriptional regulation, uncover mechanisms that occur when cellular processes are perturbed and enable identification of potential therapeutic targets for diseases [48, 29, 54]. Despite its advantages, single cell RNA

*To whom correspondence should be addressed. Equal contribution.

sequencing (scRNA-seq) datasets is undermined by a series of technical limitations, such as drop-out events (expressed genes undetected by scRNA-seq) and a high level of noise, which makes single cell gene regulatory network inference a difficult problem [15, 6, 1]. Thus, given the high dimensionality and the complex nature of these datasets, there is a pressing need to develop specialized computational methods for the inference of cell-type specific GRN's [7, 43].

In general, scRNA-seq datasets from a single experiment often contain cells belonging to different cellular subtypes [8, 19]. As the cells belonging to the different clusters have originated from the same tissue, their networks may be assumed to share some common structure. Independent network estimation for each subgroup does not allow for incorporation of shared structures between the subgroups. Thus, there is a need to develop joint network estimation strategies that can borrow information across subgroups while retaining between subgroup heterogeneity. Dozens of algorithms have been proposed for reconstruction of gene regulatory networks in scRNA-seq datasets [7, 43, 37]. However, most of these algorithms consider cells to be identically and independently distributed samples which violates the assumption of natural subgroups present within the data. Given our assumption of a clustered dataset, we should be able to apply these algorithms to estimate networks from each cell subgroup separately. However, this approach would not take the commonality across different cells within the subgroup which may lead to information loss. Multiple algorithms have been proposed for joint estimation of networks from high dimensional data. Most of these methods assume that the data has a Gaussian distribution. Seminal papers by [16, 10] paved the way for penalized estimation of multiple Gaussian graphical models (GGM), and demonstrated the use of lasso based penalty functions for better estimation of the sparsity structures across multiple groups. Very few methods have been proposed for joint estimation of multiple networks from scRNA-seq datasets. [35] developed PIPER, a penalized local Poisson graphical model (LPGM) [2] for joint estimation of multiple networks in scRNA-seq datasets. One of the main limitations of PIPER is that the Poisson distribution has one single parameter characterizing both the mean and the standard deviation. Single cell datasets would be much better characterized by a negative binomial distribution which has a separate dispersion parameter or a zero inflated negative binomial distribution which could account for the excess zeroes. To account for the non-gaussian nature of the scRNA-seq datasets, [55] proposed a modification of the joint Gaussian copula graphical model (JGCGM) based on the Gaussian copula transformation proposed in [26]. [11] proposed a three step hybrid joint estimation strategy that relies on (a) integrated application of a Bayesian zero inflated Poisson (ZIP) based model imputation strategy and single cell imputation technique McImpute [22, 33], (b) data Gaussianization [27] and eventually (c) joint estimation of a Gaussian graphical model [10]. Contrary to [35], the last two proposed approaches estimate graphical models for continuous data and rely on a data transformation step for making the data continuous.

Recent work in graph topology inference focuses on ideas developed in graph signal processing (GSP), which extends classical signal processing concepts to data defined on nodes of a graph, i.e. *graph signals* [51]. GSP based graph learning (GL) approaches infer the graph structure from the observed graph signals based on assumptions made about the relation between the signals and the unknown graph G [13]. These assumptions include smoothness, where graph signals are assumed to have low-frequency representation in graph frequency domain, and stationarity, where graph signals are assumed to be stationary with respect to G [31]. Due to explicit representation of graph signals in the graph frequency domain, GSP based GL has more flexibility in modeling signals compared to previous network inference methods, such

as statistical models reviewed above for GRN inference. Therefore, various GL algorithms have been proposed and shown to perform well in different applications [12, 23, 50, 40]. However, most of this prior work focuses on learning a single graph structure and is not suitable for joint learning of multiple related graphs. Recently, a method for joint inference of multiple graphs from stationary graph signals is proposed in [36], where graphs are regularized to have similar sparsity patterns.

In this paper, we present a multiview Graph Learning (mvGL) algorithm for joint inference of graphs from multiple datasets. Based on our previous work [24], where smoothness of graph signals with respect to the underlying topology has been shown to perform well for GRN inference, mvGL learns multiple graphs by deriving an optimization problem using smoothness assumption. Graphs corresponding to the different datasets are ensured to be similar to each other through a learned consensus graph. Compared previous joint graph learning method [36], we focus on smoothness because of following reasons. First, smooth signals have low-frequency and sparse representations in graph frequency domain, thus GL problem correspond to inferring efficient information processing transforms for graph signals. Second, many graph-based machine learning and signals processing tasks are developed based on the smoothness of the graph signals. Finally, many real-world applications are observed to have smooth signals [31]. mvGL is used for joint inference of GRNs across different cell types to unravel regulator genes whose interactions may vary across subgroups. During application to single cell data, mvGL is kernelized as in [24] to deal with the structure of scRNA-seq data. The proposed method has various unique advantages over existing approaches. First, it does not make any specific parametric assumptions about the data. Second, it employs specialized kernels to consider the nature of scRNA-seq data. For instance, it can employ proportionality measures as gene expressions in scRNA-seq reflect relative rather than absolute abundance. Similarly, zero-inflated Kendall’s tau can be used to handle drop-outs [52]. Finally, compared to [36], we focus on the proposed method learns an additional consensus graph, which captures the common structure across all graphs.

2. Background

2.1. Graphs and Smooth Graph Signals

A weighted undirected graph can be denoted as $G = (V, E)$ where V is the node set with $|V| = n$ and E is the edge set. An edge between nodes i and j is represented by e_{ij} and it is associated with a positive weight w_{ij} . Algebraically, a graph is represented by an adjacency matrix $\mathbf{W} \in \mathbf{R}^{n \times n}$ where $W_{ij} = W_{ji} = w_{ij}$ if $e_{ij} \in E$ and 0, otherwise. The combinatorial Laplacian matrix of G is defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$ where $\mathbf{D}$ is the diagonal matrix with node degrees, i.e. $D_{ii} = \sum_{j=1}^n W_{ij}$. The Laplacian matrix is a positive semi-definite matrix, thus its eigendecomposition is $\mathbf{L} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$ where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues. We assume the eigenvalues of $\mathbf{L}$ are ordered such that $0 = \Lambda_{11} \leq \Lambda_{22} \leq \dots, \Lambda_{nn}$. We represent vectors with all ones and zeros as $\mathbf{1}_n$ and $\mathbf{0}_n$, respectively, where the subscript indicates the size of the vector¹. The operator $\text{diag}() : \mathbf{R}^{n \times n} \rightarrow \mathbf{R}^n$ returns the diagonal of the input matrix. The operator $\text{upper}() : \mathbf{R}^{n \times n} \rightarrow \mathbf{R}^M$ returns upper triangular part of the input matrix where $M = n(n-1)/2$. For an $n \times n$ symmetric matrix $\mathbf{A}$, we define the matrix $\mathbf{S} \in \mathbf{R}^{M \times n}$ such that $\text{Supper}(\mathbf{A}) = \mathbf{A}\mathbf{1} - \text{diag}(\mathbf{A})$.

A graph signal is a function $x : V \rightarrow \mathbf{R}$ and can be represented by a vector $\mathbf{x} \in \mathbf{R}^n$ where each x_i is the signal value on node i . Graph Fourier transform (GFT) of $\mathbf{x}$ can be defined using spectrum of $\mathbf{L}$ as

¹The subscript is ignored for the notational clarity in case the size is deducible from the context.

its eigenvalues and eigenvectors provide a notion of frequency, i.e., small eigenvalues correspond to low frequency and larger ones to high frequencies [51]. GFT of $\mathbf{x}$ is then defined as $\hat{\mathbf{x}} = \mathbf{V}^\top \mathbf{x}$ and inverse GFT is $\mathbf{x} = \mathbf{V}\hat{\mathbf{x}}$. Thus, $\mathbf{x}$ is the linear combination of eigenvectors of $\mathbf{L}$ with the coefficients determined by $\hat{\mathbf{x}}$. If most of the energy of $\hat{\mathbf{x}}$ is concentrated in its first entries, $\mathbf{x}$ has a low-frequency representation in graph Fourier domain, i.e. it varies smoothly over the graph. The smoothness of $\mathbf{x}$ can be quantified by the total variation as:

$$\text{tr}(\hat{\mathbf{x}}^\top \mathbf{L} \hat{\mathbf{x}}) = \text{tr}(\mathbf{x}^\top \mathbf{V} \mathbf{L} \mathbf{V}^\top \mathbf{x}) = \text{tr}(\mathbf{x}^\top \mathbf{L} \mathbf{x}), \quad (1)$$

where smaller values indicate a smoother graph signal $\mathbf{x}$ with respect to G .

2.2. Single View Graph Learning

Given a set of graph signals $\{\mathbf{x}_i \in \mathbb{R}^n\}_{i=1}^p$ that are assumed to be smooth over an unknown graph G , the structure of G can be learned by the following optimization problem [12, 23]:

$$\underset{\mathbf{L} \in \mathcal{L}}{\text{minimize}} \quad \text{tr}(\mathbf{X}^\top \mathbf{L} \mathbf{X}) + \alpha_1 \|\text{diag}(\mathbf{L})\|_2^2 + \iota_H(\text{upper}(\mathbf{L})) + \alpha_2 \|\mathbf{L}\|_{\text{F}, \text{off}}^2, \quad (2)$$

where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the data matrix whose columns are $\mathbf{x}_i$'s, $\mathcal{L} = \{\mathbf{L} : L_{ij} = L_{ji} \leq 0 \forall i \neq j, \mathbf{L}\mathbf{1} = \mathbf{0}\}$ is the set of valid Laplacian matrices. The first term in (2) measures the smoothness of the graph signals. The second term regularizes the degree distribution of the graph. The third term is the indicator function for hyperplane $H = \{\mathbf{v} : \mathbf{v}^\top \mathbf{1} = -n\}$ and is included to prevent the trivial solution $\mathbf{L} = \mathbf{0}$. The final term is the Frobenius norm of the off-diagonal entries of $\mathbf{L}$ and controls the density of the learned graph. Hereafter, the problem in (2) is referred to as single view graph learning (svGL).

3. Method

In this section, the proposed method to infer multiple GRNs from multiple cell types is described. The method is developed by deriving an optimization problem that extends svGL to handle multiple datasets. The optimization problem is then solved by prox-linear block coordinate descent (BCD) algorithm.

3.1. Multiview Graph Learning

Assume that we are given N datasets $\{\mathbf{X}^i \in \mathbb{R}^{n \times p_i}\}_{i=1}^N$ where the columns of $\mathbf{X}^i$ are p_i graph signals defined on an unknown graph $G^i = (V, E^i, W^i)$. All G^i 's have the same vertex set V with $|V| = n$. Their respective edge sets E^i 's and associated edge weights W^i are assumed to be different, but similar to each other. Due to this assumption, inference of G^i can be improved by incorporating information from other graphs. For instance, in the analysis of scRNA-seq expressions, the datasets generated from a single tissue comprises of cells belonging to multiple cell-types. As cells having different cell-type assignments originate from a single tissue, their networks G^i may be considered to be related to each other. Learning G^i 's independently from each cell-type does not allow for information borrowing from graphs of other cell types. Thus, the estimation of cell-type specific graphs can be improved by simultaneously learning multiple cell-type graphs. To this end, we propose an optimization problem, where we learn G^i 's simultaneously while regularizing them to be close to a consensus graph G . During the optimization, G is also learned by combining information from G^i 's; thus G^i 's are ensured to be similar to each other through G . Let $\mathbf{L}^i$'s and $\mathbf{L}$ be the Laplacian matrices of G^i 's and G , respectively, then the corresponding

optimization problem is:

$$\begin{aligned} & \underset{\mathbf{L}^1, \dots, \mathbf{L}^N, \mathbf{L}}{\text{minimize}} \sum_{i=1}^N \left\{ \text{tr}(\mathbf{X}^i \mathbf{L}^i \mathbf{X}^i) + \alpha_1 \|\text{diag}(\mathbf{L}^i)\|_2^2 + \mathbf{1}_H(\text{upper}(\mathbf{L}^i)) + \alpha_2 c(\mathbf{L}^i, \mathbf{L}) \right\} + \alpha_3 r(\mathbf{L}) \\ & \text{subject to } \mathbf{L}^i \in \mathbb{L} \text{ and } \mathbf{L} \in \mathbb{L}, \end{aligned} \quad (3)$$

where the first term in the summation measures the smoothness of $\mathbf{X}^i$ over G^i , the second and third terms have the same meaning as in svGL. The function $c(\mathbf{L}^i, \mathbf{L})$ is included to ensure that each $\mathbf{L}^i$ is close to $\mathbf{L}$. Finally, $r(\mathbf{L})$ is a regularizer that controls the sparsity of $\mathbf{L}$. While different choices for $c(\cdot, \cdot)$ and $r(\cdot)$ can be considered, in this work, $c(\mathbf{L}^i, \mathbf{L}) = \|\alpha_4 \mathbf{L}^i - \mathbf{L}\|_{F, \text{off}}^2$ and $r(\mathbf{L}) = \|\mathbf{L}\|_1$ where α_4 is a free variable introduced to control the sparsity of G^i 's. (see Section 3.3 for further details about the selection of hyperparameters). Finally, (3) can be kernelized to exploit various (nonlinear) relations between graph signals. Kernelization is important for GRN inference as it is unclear which association measure between gene expressions is best performing for various scRNA-seq data analysis [52]. Therefore, following [24], we kernelize (3) by noting $\text{tr}(\mathbf{X}^i \mathbf{L}^i \mathbf{X}^i) = \text{tr}(\mathbf{X}^i \mathbf{X}^i \mathbf{L}^i)$ and changing $\mathbf{X}^i \mathbf{X}^i \mathbf{L}^i$ with any positive semi-definite kernel matrix $\mathbf{K}^i$. In this paper, we consider three kernels: correlation coefficient, r , proportionality measure ρ [45] and zero-inflated Kendall's tau (τ_{zi}) [42]. These kernels are considered because r is a commonly used for network inference, ρ is found as the best performing measure in [52] and τ_{zi} can handle of dropouts in scRNA-seq.

3.2. Optimization

The problem in (3) can be written in vectorized form, where one learns the upper triangular parts of $\mathbf{L}^i$'s and $\mathbf{L}$. Let $\mathbf{k}^i = \text{upper}(\mathbf{K}^i)$, $\mathbf{d}^i = \text{diag}(\mathbf{K}^i)$, $\ell^i = \text{upper}(\mathbf{L}^i)$ and $\ell = \text{upper}(\mathbf{L})$. The vectorized form of (3) is:

$$\begin{aligned} & \underset{\ell^1, \dots, \ell^N, \ell}{\text{minimize}} \sum_{i=1}^N \left\{ \langle \mathbf{k}^i - \mathbf{S}^\top \mathbf{d}^i, \ell^i \rangle + \alpha_1 \|\mathbf{S}^\top \ell^i\|_2^2 + \alpha_2 \|\alpha_4 \ell^i - \ell\|_2^2 \right\} + \sum_{i=1}^N \left\{ \mathbf{1}_H(\ell^i) \right\} + \alpha_3 \|\ell\|_1 \\ & \text{subject to } \ell^i \leq 0, \forall i \text{ and } \ell \leq 0. \end{aligned} \quad (4)$$

This problem can be solved using block coordinate descent (BCD) as its objective function is the sum of a smooth term and a separable non-smooth term. At each iteration of BCD, the objective function is minimized cyclically over each block, i.e. ℓ^i 's and ℓ , while fixing the other blocks [56]. The subproblems considered at each iteration of BCD can be solved exactly or inexactly. When minimizing with respect to ℓ^i 's, we perform an inexact minimization using prox-linear update, as it results in an easy-to-solve subproblem and has fast convergence when extrapolation is employed. When minimizing with respect to ℓ , we perform an exact minimization. More specifically, let $F(\ell^1, \dots, \ell^N, \ell)$ be the the first term in (4) and $\nabla_{\ell^i} F(\ell^1, \dots, \ell^N, \ell)$ be its gradient with respect to ℓ^i . At iteration t , ℓ^i is updated as follows:

$$\ell^{i(t)} = \underset{\ell^i \leq 0}{\text{argmin}} \langle \hat{\mathbf{g}}^i, \ell^i - \hat{\ell}^i \rangle + \frac{L^i}{2} \|\ell^i - \hat{\ell}^i\|_2^2 + \mathbf{1}_H(\ell^i) = \underset{\ell^i \leq 0}{\text{argmin}} \|\ell^i - \hat{\ell}^i + \frac{L^i}{2} \hat{\mathbf{g}}^i\|_2^2 + \mathbf{1}_H(\ell^i) \quad (5)$$

where $\hat{\ell}^i = \ell^{i(t-1)} + w(\ell^{i(t-1)} - \ell^{i(t-2)})$ is the extrapolated point with w being the extrapolation weight, $\hat{\mathbf{g}}^i = \nabla_{\ell^i} F(\ell^1, \dots, \ell^{i-1(t)}, \hat{\ell}^i, \ell^{i+1(t-1)}, \dots, \ell^{(t-1)})$ and L^i is the Lipschitz constant of $\nabla_{\ell^i} F(\ell^1, \dots, \ell^N, \ell)$.

The solution of this problem is the projection of $\hat{\ell}^i + -\frac{L^i}{2}\hat{g}^i$ onto the negative simplex [14]. After all ℓ^i 's are updated, the update of ℓ is obtained by solving (4) with respect to ℓ :

$$\ell^{(t)} = \underset{\ell \leq 0}{\operatorname{argmin}} \sum_{i=1}^N \left\{ \alpha_2 \|\alpha_4 \ell^i - \ell\|_2^2 \right\} + \alpha_3 \|\ell\|_1 = \underset{\ell \leq 0}{\operatorname{argmin}} \left\| \frac{\alpha_4}{N} \sum_{i=1}^N \ell^i - \ell \right\|_2^2 + \frac{\alpha_3}{\alpha_2 N} \|\ell\|_1, \quad (6)$$

which is the proximal operator of ℓ_1 norm evaluated at $\frac{\alpha_4}{N} \sum_{i=1}^N \ell^i$ [39]. The convergence of this approach for solving (4) can be found in [56].

3.3. Hyperparameter Selection

There are four parameters that need to be selected for mvGL. α_1 controls the degree distribution of the learned graphs such that larger values of α_1 make all nodes have very similar degrees. Since degree distribution of real world networks is generally heterogeneous, we set α_1 to the smallest possible value. However, if α_1 is too small, the learned graph may become disconnected. Thus, we set α_1 to the smallest possible value which ensures heterogeneous degree distribution and connectedness of the learned graphs. α_4 controls the edge density of G^i 's, i.e., larger values of α_4 give denser graphs. We set it to a value that provides a desired edge density. Similarly, α_3 controls the density of G such that as α_3 gets larger G becomes sparser. Thus, we determine α_3 in the same way as α_4 . Finally, α_2 controls how close G^i 's are to G , thus it can be set to a value which results in a desired correlation between G and G^i 's. In the results section, the correlation between G and G^i 's are determined so that there will be a weak correlation between G^i 's, which means views are related to each other but view-specific features are also allowed to be learned.

4. Results

In this section, the performance of mvGL is evaluated on simulated data and two real scRNA-seq datasets. For simulated data, learned graphs are compared to ground truth networks to quantify the performance of mvGL. We employed area under precision recall curve (AUPRC) as the performance metric during this analysis. The performance of mvGL is benchmarked against three GRN algorithms and svGL on simulated datasets. We consider GENIE3, GRNBOOST2 and PIDC methods as these are found to be the best performing GRN inference algorithms for scRNA-seq data [43]. These methods and svGL can only learn a single graph from each dataset at a time². Therefore, they are applied to each X^i separately and learned graphs are compared to ground truth G^i 's. Additionally, they are applied to concatenation of X^i as a way to infer G , i.e. the consensus graph. Results for three different kernels are reported for mvGL and svGL: proportionality (ρ), correlation (r) and zero-inflated Kendall's tau (τ_{zi}). Hyperparameters of mvGL are set as described in Section 3.3. α_1 and α_2 for svGL control the degree distribution and density of the learned graph, respectively. Their values are set the same way as α_1 and α_4 are determined in mvGL.

For the analysis of real datasets, we first note that GRN can include both activating and inhibitory edges. In [24], it is observed that two genes that are connected by an activating edge have similar expressions. On the other hand, if they are connected with an inhibitory edge, their expressions are dissimilar. From GSP perspective, the former observation implies that the graph signal, i.e. gene expressions in a cell,

²Joint GRN inference algorithms are fairly new and their toolboxes are not publicly available. Therefore, we could only benchmark against the methods that learn from single dataset.

has a low-frequency representation over the graph defined by activating edges, while the latter implies the graph signal has a high-frequency representation over the graph defined by inhibitory edges. mvGL as formulated in (3) only learns activating edges, because it assumes that graph signals have low-frequency representation over the unknown GRN. However, it can also be used to infer inhibitory edges. In particular, if the smoothness of the signals is maximized with respect to L^i 's and L , the learned edges will be inhibitory. Thus, by changing K^i with $-K^i$, we are able to learn inhibitory edges of the GRN. Based on this observation, results for both activating and inhibitory edges are reported in real data analysis.

4.1. Simulated Data

To validate the performance of mvGL, we simulate gene expression data from a multivariate zero-inflated negative binomial (ZINB) distribution. The ZINB distribution has been shown to accurately capture the characteristics of single cell datasets in several published studies [46, 17]. Given a known graph structure, we generate synthetic datasets using an algorithm developed by [57] and illustrated in [22, 24]. Two graph structures are considered for creating the baseline graph G : random graphs following an Erdős-Rényi model with an edge density of 0.1 and hub graphs following a Barabási-Albert model and having degree distribution that follows a power-law function. Since many real-world biological networks are known to possess scale free properties, hub networks present a more realistic representation of the single cell graph structures. Next, we generate $K = 5$ individual networks $\{G_i\}_{i=1}^K$ by adding ηM individual edges to the baseline graph G where M is the number of edges in G and $\eta = 0.1$. The ZINB simulator is then used to generate datasets $\{X_i\}_{i=1}^K$, having $n = 100$ genes and $p_i = 400$ cells from the underlying graphs $\{G_i\}_{i=1}^K$. The three parameters of the ZINB distribution; λ , k and ω , which control its mean, dispersion and degree of zero-inflation, respectively were determined using a real scRNA-seq dataset [21]. This dataset has been previously used for estimating simulation parameters for a Bayesian zero inflated poisson model in [11]. To investigate the effects of dropouts in graph reconstruction accuracy, we generate datasets by varying levels of zero-inflation between (0.26 – 0.36)%. Each simulated dataset include 10 random realizations and the average performance over realizations is reported.

Performance of methods on all datasets are shown in Figure 1. Since most of the GRN algorithms learn fully connected networks, we set desired density for mvGL and svGL to a large value (around 0.5). For views, average AUPRC over all views and realizations are reported. Irrespective of which kernel is employed, mvGL is performing better than svGL. This indicates that the proposed method shares valuable information across views, which improves the performance. Existing state of the art GRN algorithms are observed to perform poorly, which is inline with findings in previous joint GRN inference works [55]. When comparing the performance of different kernels, proportionality measure is found to be the best performing kernel. This finding is inline with [52], where different association metrics are compared based on their performance on various learning problems from scRNA-seq data. We also report results for how well consensus graphs match with the baseline graph G . As mentioned before, consensus graphs of benchmarking methods is found by concatenating X^i 's. As concatenation leads to increased cell numbers, AUPRC values of benchmarking methods increase compared to values for views. Even in this case, mvGL emerges as the best performing algorithm. Finally, performance of most of the methods drops with increasing zero-inflation level (Although there are cases where this is not the case, we believe this is due to randomized nature of algorithms and datasets.) Performance drop in mvGL with increasing dropouts is less than the drop observed in other algorithms. This indicates sharing information across views may

Figure 1: Performance of mvGL and svGL with three different kernels on simulated data. Top plots indicate AUPRC values for G^i 's, where average over views and 10 realizations is reported. Bottom plots show AUPRC values for G , where average over realizations is reported.

bring robustness to dropouts.

4.2. Analysis of scRNA-seq data from mouse embryonic stem cell differentiation

Central to differentiation process and many other cellular processes is the expression of right combination of genes or modules of genes. Accurate characterization of the co-expression networks for progenitor and multiple cell types can help in understanding the cascade of cellular state transitions [32]. In this real data example, we study the differentiation process of mouse embryonic stem cells (mESC) using single cell RNA sequencing datasets [25]. This data was generated using high-throughput droplet-microfluidic approach and was primarily used to study differentiation in mESC before and after leukemia inhibitory factor (LIF) withdrawal. Since LIF maintains pluripotency of mESC, LIF withdrawal is considered to initiate the differentiation process. The dataset contains cells sampled from 4 states (or natural subgroups): before leukemia inhibitory factor (LIF) withdrawal, day 0 and after the withdrawal LIF for 2, 4, 7 days respectively. The subgroups contain 933, 303, 683 and 798 cells, respectively. This dataset has been previously analysed for studying joint graphical estimation in [35, 55] and similar to them we only consider the 72 stem cell markers for our example [44].

As per our workflow, we first estimated the subgroup specific and the consensus graphs, which are learned with activating and inhibitory edge densities set to 0.1. Based on the results of simulated data, we study graphs learned by proportionality measure. Next, we learned and ranked the hub genes for each of the learned graphs. Since gene networks are known to contain a handful of hub nodes that play very important roles via their many diverse connections, analysis of these hub nodes should highlight genes

Figure 2: Hub genes in learned GRNs by mvGL for mESC dataset. To determine hub genes, degrees of the genes are sorted in descending order and the genes in the top-25th percentile are identified as hub genes. G^1 , G^2 , G^3 and G^4 correspond to Day 0, Day 2, Day 4 and Day 7, respectively.

that are key to maintaining the cell differentiation process [3]. Figure 2 reports the hub genes learnt by mvGL. Our analysis confirms the importance of regulator genes Nanog, Sox2, Pou5f1, Zfp42, Utf1 in early stages of differentiation. Sox2, Nanog and Pou5f1 are known to play a fundamental role in the self-renewal and pluripotency of mouse embryonic stem cells [59] and therefore have been correctly identified as hub genes in the early stages of differentiation. Zfp42 and Utf1, both markers of pluripotency [30, 53] are also correctly identified as hub genes in Day 0 and Day 2 respectively. Reduction in expression of Nanog has been shown to be correlated with the induction of genes GATA4 which initiates differentiation of pluripotent cells [20] and therefore it has been correctly identified as hub genes in Day 2.

4.3. Analysis of scRNA-seq data from medulloblastoma

Medulloblastoma (MB) is a highly malignant cerebellar tumor mostly affecting young children [38]. Several studies have been done to pinpoint the genetic drivers in each of its four distinct tumour subgroups: WNT-pathway-activated, SHH-pathway-activated, and the less-well-characterized Group 3 and Group 4 [38]. Among these subgroups, Group 3 and Group 4 tumors account for majority of the MB diagnoses, with Group 3 MB having a metastatic diagnosis of approximately 50%. Transcription factors (TFs) *Myc2* and *Otx2* have commonly been identified as key oncogenic TFs in Group 3 and 4 tumorigenesis. [13] used their proposed joint single cell network algorithm to study the roles of *Myc* and *Otx2* utilizing the MB scRNA-seq data set (GSE119926) by [21]. Using the same selected samples from a subset of 17 individuals that were grouped into three subsets Group 3, Group 4 and an intermediate cell type, we estimate the joint gene regulatory network for the three groups for ~ 750 genes among which most are enzyme-related genes from mammalian metabolic enzyme database [9].

Subgroup specific networks along with the consensus graph was estimated using proportionality kernel. The desired activating and inhibitory edge densities are set to 0.1. Figure 3 shows that Myc had a higher degree density for Group 3 in comparison to Group 4 and the intermediate subgroup. This confirms the major role MYC plays in initiation, maintenance, and progression in Group 3 tumors [47]. ALDH3A2 and ALDH1A2, the top ranking hub nodes were found to be downregulated in all the tumor subgroups confirming their role in cancer resistance [34, 5]. mvGL was also able to detect relationships between Myc and metabolic genes PAICS, PPAT, IMPDH1, ATIC and GART only in Group 3 tumors, however they did not make it to the top 25% of the MYC connections reported in Figure 3. These genes related to the human purine biosynthesis pathways have been previously reported to be induced and directly bound by Myc [28]. PRPS2, which emerged as one of strongest hubs in the OTX2 network analysis

Figure 3: Neighbors of MYC and OTX2 genes. Marker size is proportional to connection weights. Orange and blue indicate that the connection is activating and inhibitory, respectively. Only connections that fall into top 25 percentile are reported. G¹, G² and G³ correspond to Intermediate, Group 3 and Group 4, respectively. At the right most part of the figure, total degrees of two genes in each graph are also shown.

been known to promote increased nucleotide biosynthesis in Myc-transformed cells. This confirms that Otx2 is functionally cooperating with Myc to regulate gene expression in medulloblastoma [47]. mvGL also reported unique sets of hub genes for the OTX2 network confirming the unique role of OTX2 in MB tumorigenesis [11]. All of the results obtained indicate that mvGL was able to correctly identify the important functional interaction between OTX2 and MYC in regulating gene expression in medulloblastoma [4].

5. Conclusion

In this paper we presented a multiview Graph Learning (mvGL) algorithm for joint inference of graphs from multiple scRNA-seq datasets. mvGL, a robust non-parametric approach, borrows ideas from kernel based graph signal processing for estimating the sparse high-dimensional structure of the underlying gene regulatory networks. Using simulation studies, we demonstrated the superior performance of mvGL over single view learning for random and hub graph topologies. Additionally the performance of mvGL remained robust to increase in dropout levels. We further analyzed two real single cell datasets and mvGL was shown to identify the key regulators in each of the differentiation and disease studies.

Future work can be aimed at inferring the multiple graph structures with a combination of kernel choices. That would ease the step of choosing a kernel suited to the particular dataset. Another limitations of mvGL is hyperparameter selection, which remains as an open problem in graph learning literature. In this work, we set them by manually searching hyperparameter space for values that provide graphs with desired densities and a desired correlation across views. Since, there is a monotonic relation between α_3 (α_4) and edge densities of the learned graph, this search can be performed automatically, which we leave as a future work. Incorporating combinational kernel strategies with improved hyperparameter learning could potentially further improve the accuracy of the estimated graphs.

References

- [1] Kyle Akers and TM Murali. Gene regulatory network inference in single cell biology. *Current Opinion in Systems Biology*, 2021.
- [2] Genevera I Allen and Zhandong Liu. A local poisson graphical model for inferring networks from sequencing data. *IEEE transactions on nanobioscience*, 12(3):189–198, 2013.
- [3] Albert-László Barabási and Eric Bonabeau. Scale-free networks. *Scientific american*, 288(5):60–69, 2003.
- [4] Jens Bunt, Nancy E Hasselt, Danny A Zwiijnenburg, Jan Koster, Rogier Versteeg, and Marcel Kool. Joint binding of otx2 and myc in promotor regions is associated with high gene expression in medulloblastoma. *PloS one*, 6(10):e26058, 2011.
- [5] Peter Mu-Hsin Chang, Che-Hong Chen, Chi-Chun Yeh, Hsueh-Ju Lu, Tze-Tze Liu, Ming-Huang Chen, Chun-Yu Liu, Alexander TH Wu, Muh-Hwa Yang, Shyh-Kuan Tai, et al. Transcriptome analysis and prognosis of aldh isoforms in human cancer. *Scientific reports*, 8(1):1–10, 2018.
- [6] Geng Chen, Baitang Ning, and Tielu Shi. Single-cell rna-seq technologies and related computational data analysis. *Frontiers in genetics*, 10:317, 2019.
- [7] Shuonan Chen and Jessica C Mar. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC bioinformatics*, 19(1):1–21, 2018.
- [8] Tabula Muris Consortium et al. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562(7727):367–372, 2018.
- [9] Callan C Corcoran, Cameron R Grady, Trairak Pisitkun, Jaya Parulekar, and Mark A Knepper. From 20th century metabolic wall charts to 21st century systems biology: database of mammalian metabolic enzymes. *American Journal of Physiology-Renal Physiology*, 312(3):F533–F542, 2017.
- [10] Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 76(2):373, 2014.
- [11] Meichen Dong, Yiping He, Yuchao Jiang, and Fei Zou. Joint gene network construction by single-cell rna sequencing data. *bioRxiv*, 2021.
- [12] Xiaowen Dong, Dorina Thanou, Pascal Frossard, and Pierre Vandergheynst. Learning laplacian matrix in smooth graph signal representations. *IEEE Transactions on Signal Processing*, 64(23):6160–6173, 2016.
- [13] Xiaowen Dong, Dorina Thanou, Michael Rabbat, and Pascal Frossard. Learning graphs from data: A signal representation perspective. *IEEE Signal Processing Magazine*, 36(3):44–63, 2019.
- [14] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279, 2008.
- [15] Mark WEJ Fiers, Liesbeth Minnoye, Sara Aibar, Carmen Bravo González-Blas, Zeynep Kalender Atak, and Stein Aerts. Mapping gene regulatory networks from single-cell omics data. *Briefings in functional genomics*, 17(4):246–254, 2018.
- [16] Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.
- [17] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome biology*, 20(1):1–15, 2019.
- [18] Xiaoping Han, Ziming Zhou, Lijiang Fei, Huiyu Sun, Renying Wang, Yao Chen, Haide Chen, Jingjing Wang, Huanna Tang, Wenhao Ge, et al. Construction of a human cell landscape at single-cell level. *Nature*, 581(7808):303–309, 2020.

- [19] Shuai He, Lin-He Wang, Yang Liu, Yi-Qi Li, Hai-Tian Chen, Jing-Hong Xu, Wan Peng, Guo-Wang Lin, Pan-Pan Wei, Bo Li, et al. Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs. *Genome biology*, 21(1):1–34, 2020.
- [20] Shelley R Hough, Ian Clements, Peter J Welch, and Kristin A Wiederholt. Differentiation of mouse embryonic stem cells after rna interference-mediated silencing of oct4 and nanog. *Stem cells*, 24(6):1467–1475, 2006.
- [21] Volker Hovestadt, Kyle S Smith, Laure Bihannic, Mariella G Filbin, McKenzie L Shaw, Alicia Baumgartner, John C DeWitt, Andrew Groves, Lisa Mayr, Hannah R Weisman, et al. Resolving medulloblastoma cellular architecture by single-cell genomics. *Nature*, 572(7767):74–79, 2019.
- [22] Bochao Jia, Suwa Xu, Guanghua Xiao, Vishal Lamba, and Faming Liang. Learning gene regulatory networks from next generation sequencing data. *Biometrics*, 73(4):1221–1230, 2017.
- [23] Vassilis Kalofolias. How to learn a graph from smooth signals. In *Artificial Intelligence and Statistics*, pages 920–929. PMLR, 2016.
- [24] Abdullah Karaaslanli, Satabdi Saha, Selin Aviyente, and Tapabrata Maiti. scsgl: Signed graph learning for single-cell gene regulatory network inference. *bioRxiv*, 2021.
- [25] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.
- [26] Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.
- [27] Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(10), 2009.
- [28] Yen-Chun Liu, Feng Li, Jesse Handler, Cheng Ran, Lisa Huang, Yan Xiang, Nicola Neretti, John M Sedivy, Karen I Zeller, and Chi V Dang. Global regulation of nucleotide biosynthetic genes by c-myc. *PloS one*, 3(7):e2722, 2008.
- [29] Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019.
- [30] Shinji Masui, Satoshi Ohtsuka, Rika Yagi, Kadue Takahashi, Minoru SH Ko, and Hitoshi Niwa. Rex1/zfp42 is dispensable for pluripotency in mouse es cells. *BMC developmental biology*, 8(1):1–12, 2008.
- [31] Gonzalo Mateos, Santiago Segarra, Antonio G Marques, and Alejandro Ribeiro. Connecting the dots: Identifying network structure via graph signal processing. *IEEE Signal Processing Magazine*, 36(3):16–43, 2019.
- [32] Hirotaka Matsumoto, Hisanori Kiryu, Chikara Furusawa, Minoru SH Ko, Shigeru BH Ko, Norio Gouda, Tetsutaro Hayashi, and Itoshi Nikaido. Scode: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. *Bioinformatics*, 33(15):2314–2321, 2017.
- [33] Aanchal Mongia, Debarka Sengupta, and Angshul Majumdar. Mcimpute: matrix completion based imputation for single cell rna-seq data. *Frontiers in genetics*, 10:9, 2019.
- [34] Jan S Moreb, Dagmara Muhoczy, Blanca Ostmark, and James R Zucali. Rnai-mediated knockdown of aldehyde dehydrogenase class-1a1 and class-3a1 is specific and reveals that each contributes equally to the resistance against 4-hydroperoxycyclophosphamide. *Cancer chemotherapy and pharmacology*, 59(1):127–136, 2007.
- [35] Sumit Mukherjee, Alberto Carignano, Georg Seelig, and Su-In Lee. 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)*, pages 5034–5040. IEEE, 2018.
- [36] Madeline Navarro, Yuhao Wang, Antonio G Marques, Caroline Uhler, and Santiago Segarra. Joint inference of multiple graphs from matrix polynomials. *arXiv preprint arXiv:2010.08120*, 2020.
- [37] Hung Nguyen, Duc Tran, Bang Tran, Bahadır Pehlivan, and Tin Nguyen. A comprehensive survey of regulatory network inference methods using single cell rna sequencing data. *Briefings in bioinformatics*, 22(3):bbaa190, 2021.

- [38] Paul A Northcott, Giles W Robinson, Christian P Kratz, Donald J Mabbott, Scott L Pomeroy, Steven C Clifford, Stefan Rutkowski, David W Ellison, David Malkin, Michael D Taylor, et al. Medulloblastoma. *Nature reviews Disease primers*, 5(1):1–20, 2019.
- [39] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.
- [40] Bastien Padeloup, Vincent Gripon, Grégoire Mercier, Dominique Pastor, and Michael G Rabbat. Characterization and inference of graph diffusion processes from observations of stationary signals. *IEEE transactions on Signal and Information Processing over Networks*, 4(3):481–496, 2017.
- [41] Raphael Petegrosso, Zhuliu Li, and Rui Kuang. Machine learning and statistical methods for clustering single-cell rna-sequencing data. *Briefings in bioinformatics*, 21(4):1209–1223, 2020.
- [42] Ronald S Pimentel, Magdalena Niewiadomska-Bugaj, and Jung-Chao Wang. Association of zero-inflated continuous variables. *Statistics & Probability Letters*, 96:61–67, 2015.
- [43] Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and TM Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods*, 17(2):147–154, 2020.
- [44] Laralynne M Przybyla and Joel Voldman. Attenuation of extrinsic signaling reveals the importance of matrix remodeling on maintenance of embryonic stem cell self-renewal. *Proceedings of the National Academy of Sciences*, 109(3):835–840, 2012.
- [45] Thomas P Quinn, Mark F Richardson, David Lovell, and Tamsyn M Crowley. propr: an r-package for identifying proportionally abundant features using compositional data analysis. *Scientific reports*, 7(1):1–9, 2017.
- [46] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. Zinb-wave: A general and flexible method for signal extraction from single-cell rna-seq data. *bioRxiv*, page 125112, 2017.
- [47] Martine F Roussel and Giles W Robinson. Role of myc in medulloblastoma. *Cold Spring Harbor perspectives in medicine*, 3(11):a014308, 2013.
- [48] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.
- [49] Guido Sanguinetti and Vân Anh Huynh-Thu. *Gene regulatory networks*. Springer, 2019.
- [50] Santiago Segarra, Antonio G Marques, Gonzalo Mateos, and Alejandro Ribeiro. Network topology inference from spectral templates. *IEEE Transactions on Signal and Information Processing over Networks*, 3(3):467–483, 2017.
- [51] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine*, 30(3):83–98, 2013.
- [52] Michael A Skinnider, Jordan W Squair, and Leonard J Foster. Evaluating measures of association for single-cell transcriptomics. *Nature methods*, 16(5):381–386, 2019.
- [53] Vincent van den Boom, Susanne M Kooistra, Marije Boesjes, Bart Geverts, Adriaan B Houtsmuller, Koshiro Monzen, Issei Komuro, Jeroen Essers, Loes J Drenth-Diephuis, and Bart JL Eggen. Utl1 is a chromatin-associated protein involved in es cell differentiation. *The Journal of cell biology*, 178(6):913–924, 2007.
- [54] Monique GP Van Der Wijst, Dylan H de Vries, Harm Brugge, Harm-Jan Westra, and Lude Franke. An integrative approach for building personalized gene regulatory networks for precision medicine. *Genome medicine*, 10(1):1–15, 2018.
- [55] Nuosi Wu, Fu Yin, Le Ou-Yang, Zexuan Zhu, and Weixin Xie. Joint learning of multiple gene networks from single-cell gene expression data. *Computational and structural biotechnology journal*, 18:2583–2595, 2020.
- [56] Yangyang Xu and Wotao Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013.

- [57] Inbal Yahav and Galit Shmueli. On generating multivariate poisson data in management science applications. *Applied Stochastic Models in Business and Industry*, 28(1):91–102, 2012.
- [58] Wencheng Yin, Luis Mendoza, Jimena Monzon-Sandoval, Araxi O Urrutia, and Humberto Gutierrez. Emergence of co-expression in gene regulatory networks. *PloS one*, 16(4):e0247671, 2021.
- [59] Qing Zhou, Hiram Chipperfield, Douglas A Melton, and Wing Hung Wong. A gene regulatory network in mouse embryonic stem cells. *Proceedings of the National Academy of Sciences*, 104(42):16438–16443, 2007.