MuDCoD

Multi-Subject Community Detection in Dynamic Gene Networks

Ali Osman. Berk Şapcı[1], Shan Lu[2], Oznur Tastan[1*], Sündüz Keleş[2,3*]

1 : **Faculty of Engineering and Natural Sciences, Sabancı University, Turkey;**
2 : **Department of Statistics, University of Wisconsin, Madison, WI, USA;**
3 : **Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI, USA.**

Problem Definition

Jerber-2021 Dataset

- scRNA-seq count matrices.
- Gene-gene networks for each donor at each time point based on Pearson correlation matrices.
  - **After preprocessing:**
    - 2 days: day-32 and day-50.
    - 3 cell types: DA, Sert, and Epen1.
    - 16, 22, and 8 donors: respectively for cell types.

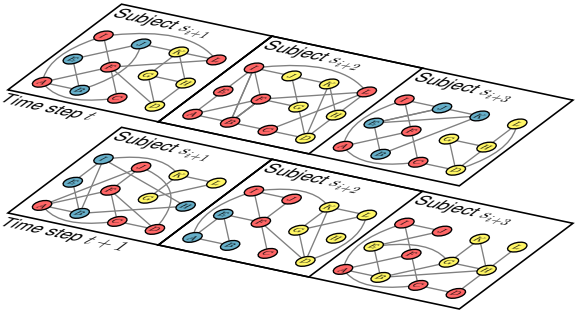
Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation

Julie Jerber[1,2,10], Daniel D. Seaton[3,10], Anna S. E. Cuomo[3,10], Natsuhiko Kumasaka[2], James Haldane[2], Juliette Steer[2], Minal Patel[2], Daniel Pearce[2], Malin Andersson[2], Marc Jan Bonder[3], Ed Mountjoy[1], Maya Ghoussaini[1], Madeline A. Lancaster[4], HipSci Consortium*, John C. Marioni[2,3,5] ✉, Florian T. Merkle[6,7] ✉, Daniel J. Gaffney[2,11] ✉ and Oliver Stegle[2,3,8,9,11] ✉

Problem Definition Cont'd.

Problem

Given a multi-subject dynamic gene co-expression network, we aim to infer the *communities* for each time point and subject.



A Baseline Method: Spectral Clustering

We have multiple time series (unweighted) gene co-expression networks; $\mathcal{G}_{s0}, \ldots \mathcal{G}_{s(T-1)}$ for each subject $s = 0, \ldots, S-1$.

$$L = D^{-1/2} A D^{-1/2} \text{ where } D_{i,j} = \begin{cases} \deg(v_i) & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $\deg(v_i)$ is the degree of node i and A is the adjacency matrix of the $\mathcal{G}$.

- Define $V_{st} \in \mathbb{R}^{G \times K}$ as a matrix with columns corresponding to the K leading eigenvectors of L_{st}.
- Find communities separately at each snapshot of time step and for each individual by clustering on V_{st}.

MuDCoD Formulation

Let $U_{st} = V_{st} V_{st}^T$. In order estimate $\overline{U}_{st}$, we propose the following optimization problem:

$$\min_{\substack{\overline{U}_{st} \\ s=0,\dots,S-1 \\ t=0,\dots,T-1}} \sum_{t=0}^{T-1} \left(\|U_{st} - \overline{U}_{st}\|_F^2 + \beta \|\overline{U}_{st} - \bar{U}_t\|_F^2 \right) + \sum_{t=0}^{T-2} \alpha \|\overline{U}_{st} - \overline{U}_{s(t+1)}\|_F^2 \tag{2}$$

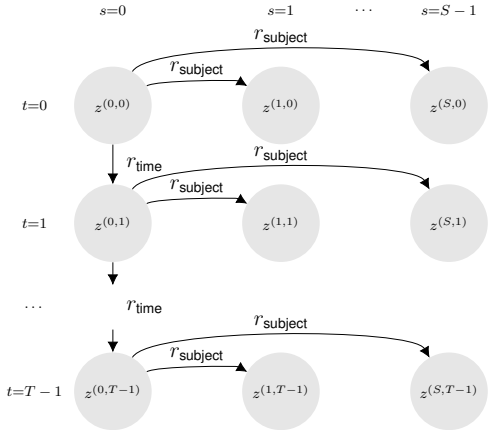
$$\text{subject to } \overline{U}_{st}, \in \left\{ VV^T \ : \ V \in \mathbb{R}^{G \times K}, \ V^T V = I \right\} \ \forall s, \forall t.$$

$\alpha \|\overline{U}_{st} - \overline{U}_{s(t+1)}\|_F^2$ enforces smoothness over the time dimension.

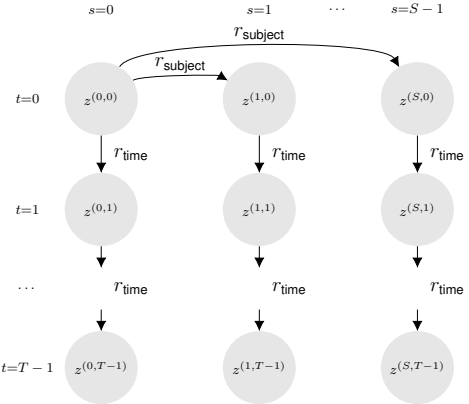
$\beta \|\overline{U}_{st} - \bar{U}_t\|_F^2$ constrains subject-specific variation from the mean time-dependent projection matrix $\bar{U}_t$:

$$\left[\bar{U}_t\right]_{ij} = \frac{1}{S} \sum_{s=0}^{S-1} \left[\overline{U}_{st}\right]_{ij}.$$

Simulation Experiments: Data Generation

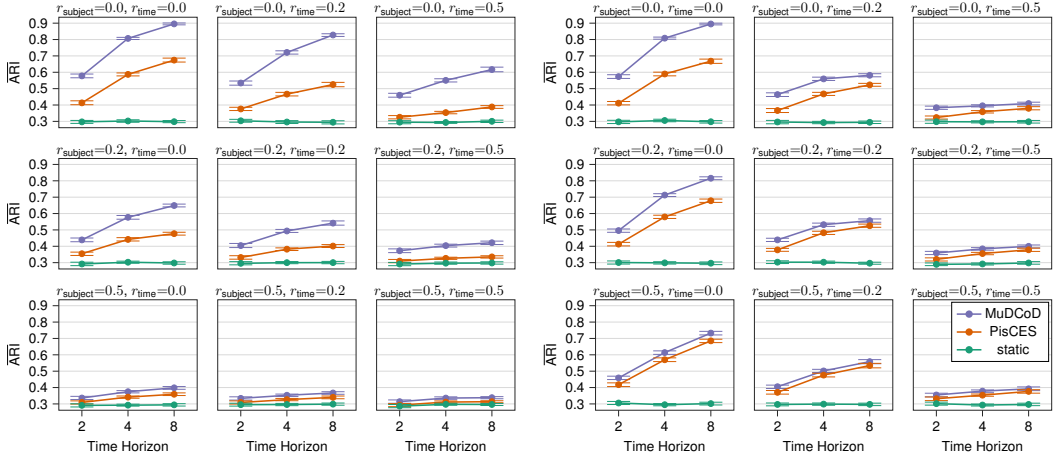


(a) *SSoS setting:* subjects evolve from a common ancestor at each time step t; only the ancestor's evolution over time is parameterized.



(b) *SSoT setting:* subjects evolve from a common ancestor at $t{=}0$; and then they evolve independently over time.

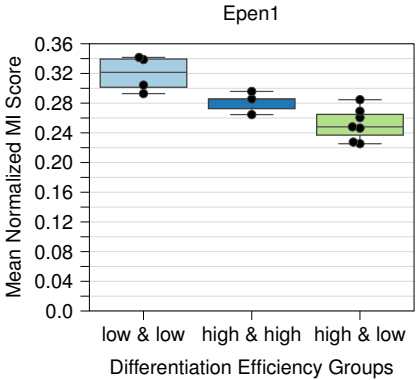
Simulation Experiments: Results



(a) *The SSoS setting.* (b) *The SSoT setting.*

Application to Jerber-2021 Data - 1

MuDCoD discovers revealing gene modules.

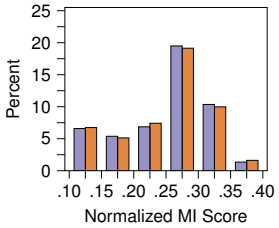


Jerber-2021 Dataset

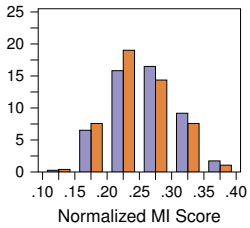
Consistent with the differentiation dynamics, we observed relatively higher heterogeneity within the high group compared to the low group.

Application to Jerber-2021 Data - 2

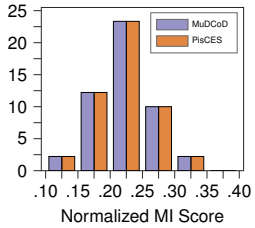
MuDCoD tends to yield higher normalized MI scores between subjects.
MuDCoD displays a comparable heterogeneity to other methods across the time points.



(a) NMI between subjects on day-30.



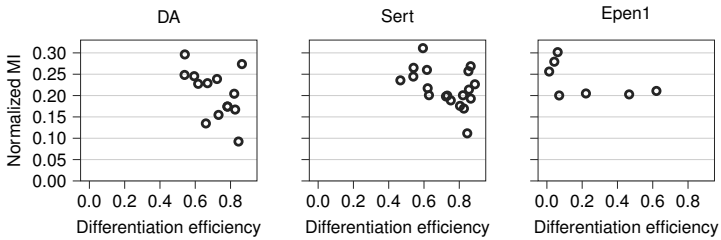
(b) NMI between subjects on day-52.



(c) NMI between day-30 and day-52 for each subject.

Application to Jerber-2021 Data - 3

Normalized MI scores between consecutive time points tends to decrease with increasing differentiation efficiency.



Future Directions

- Considering dissimilar subgroups present in the data.
- Extending this framework for irregularly sampled time series.
- scRNA-seq datasets are noisy and sparse?
  Further investigation for biological interpretation and implications.