Methods of Developing Causal Gene Regulatory Networks of Dynamic Single Cell Data

Kiley Hewitt[1], Tyler Lovelace[2], Takis Benos[2]

[1]TECBio REU @ Pitt, Dept. of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15260

[2] Dept. of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15260

Abstract

Causal inference has important potential applications in learning the gene regulatory networks of dynamic cell data. These applications are typically in finding the putative driver genes responsible for disease progression or tissue differentiation. However, the existing methods for learning causal models are not well-suited to the dynamic, nonlinear interactions in gene regulatory networks. Developing methods to learn these networks thus requires two assumptions made by existing methods to be challenged. First, the stationarity assumption assumes that the variables learned in a network do not change over time. Second, the linearity assumption assumes that all interactions between a network's variables are linear.

The method is modified to challenge these assumptions by incorporating time as a variable in the dataset so that any non-stationary variables are explained as its descendants, and by finding genes that have nonlinear, monotonic connections using nonparanormal data transformation. An ODE simulated dataset is used that simulates the cGAS pathway of 40,000 cells stochastically exposed to a virus. In each cell, the molecular expression levels, as well as the given time point, were then recorded to the dataset. Nonparanormal data transform challenges the assumption of linearity in the methods used to find the initial undirected graph by finding the linear relationships between the cumulative distribution functions of variables. Causal inference is then found by using the algorithms of PC-Stable, PC-Max, CPC-Stable, and FGES.

Under these challenged assumptions, the PC algorithms had higher asymptotic success while FGES had higher success with a low sample size. The PC algorithm's stronger balance between precision and recall caused it to perform better than FGES asymptotically. However, the FGES algorithm performed better when learning the graph with low sample sizes, having much higher adjacency recall than the PC algorithms with only slightly worse precision.

Introduction/Background

The causal inference of networks produced by dynamic single cell data has high potential in the etiology of cell networks impacted by dynamic processes such as differentiation or disease progression. In the causal inference process, the linear correlations between different variables in the data are first established and these correlations are made into an undirected graph. Causal inference algorithms are then run that orient the conditional independence relationships between correlated genes found.

Assumptions that prevent existing methods from learning dynamic, nonlinear networks:

Assumption 1: Assumption of stationarity
Challenge to assumption: Incorporating time as a variable so that non-stationary effects are modeled as a descendent of time

Assumption 2: Assumption of linearity
Challenge to assumption: Use nonparanormal transform to learn linear relationships between cumulative distribution functions and thus accommodate any smooth, monotonic interactions

Challenging Causal Inference Assumptions

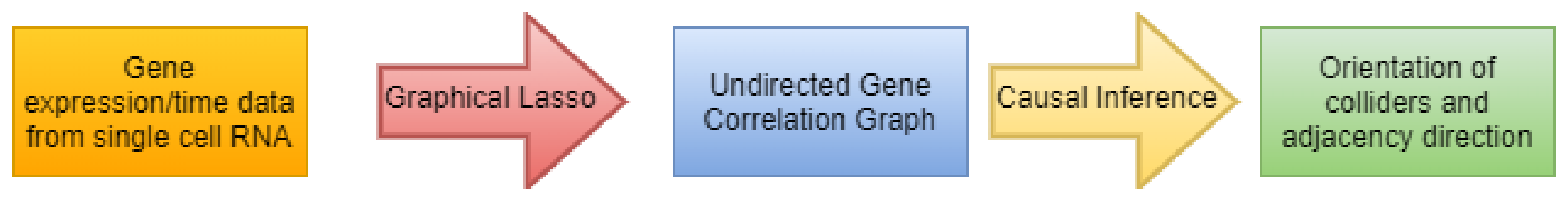

Gene expression/time data from single cell RNA
Graphical Lasso
Undirected Gene Correlation Graph
Causal Inference
Orientation of colliders and adjacency direction

Figure 1. Path of dynamic single cell RNA data to directed graph.

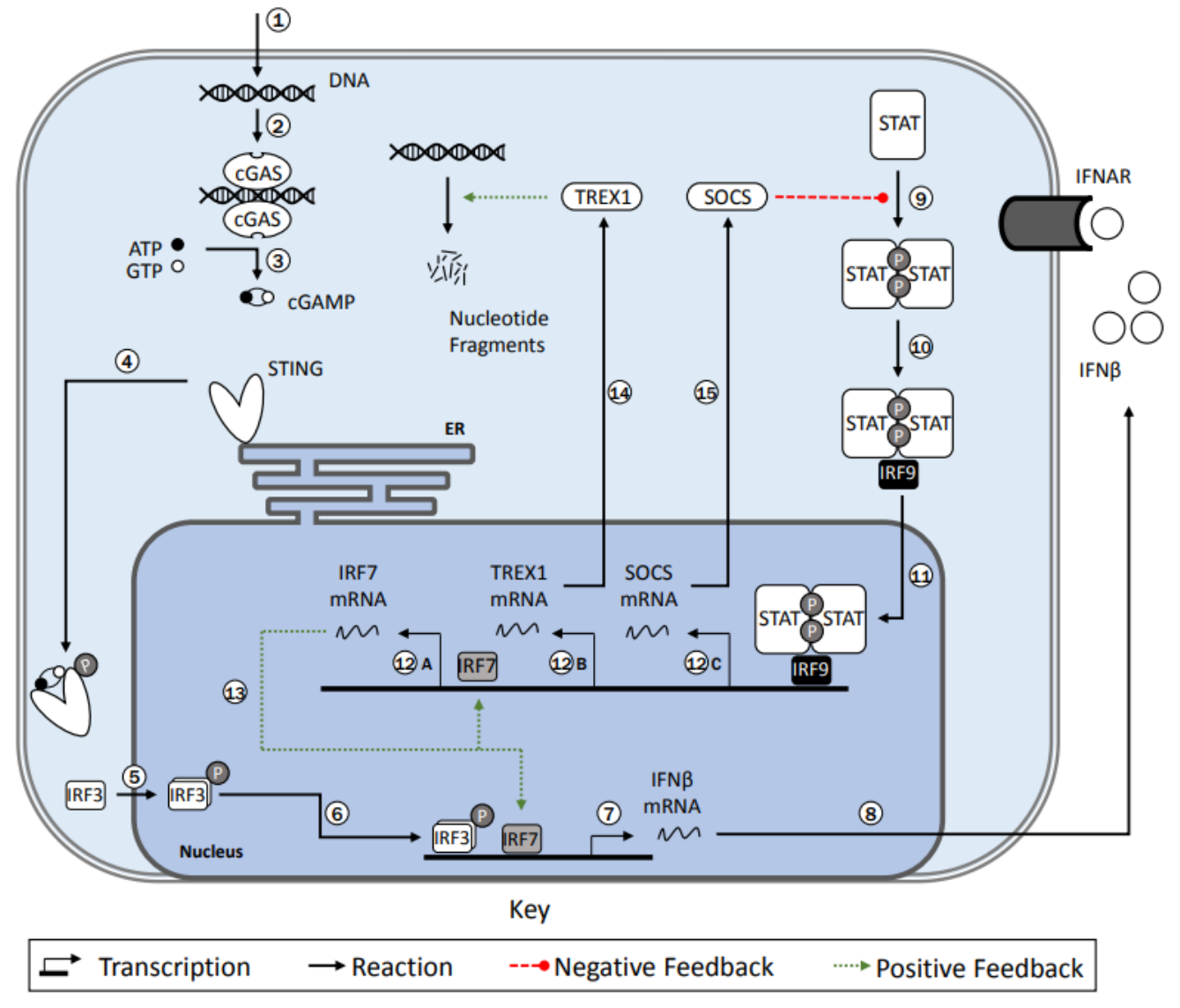


Figure 2. The cGAS Pathway. A cell signaling pathway that the simulated data is based upon.

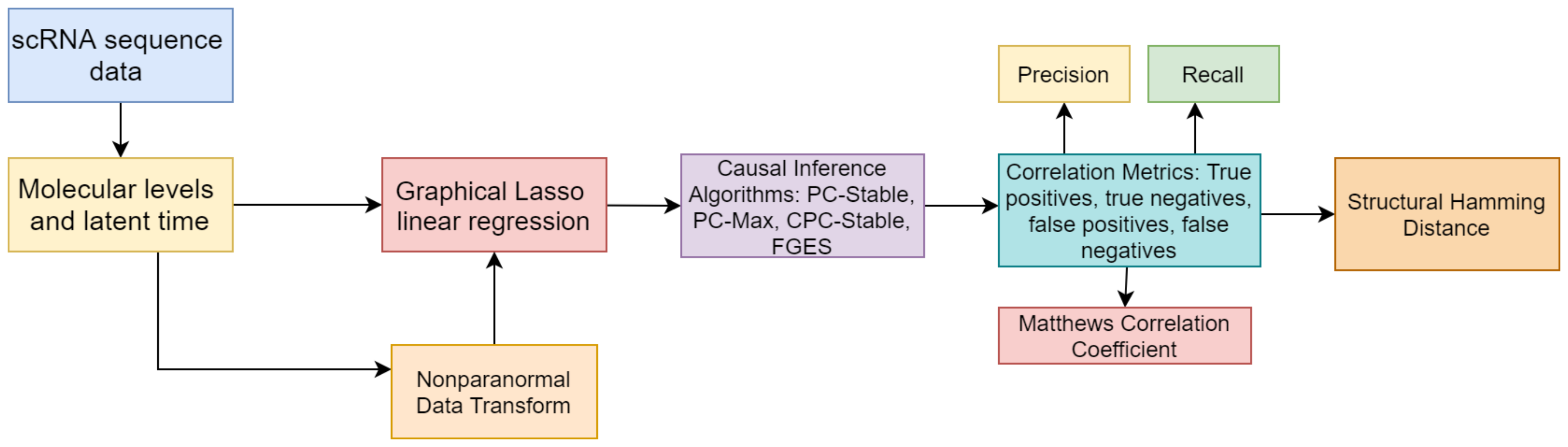
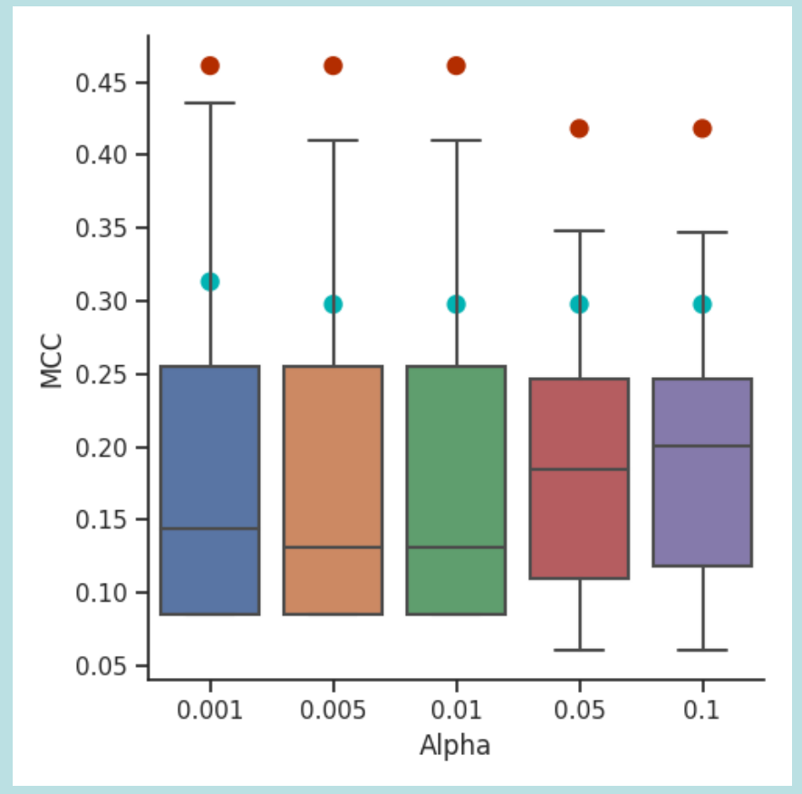

scRNA sequence data
Molecular levels and latent time
Graphical Lasso linear regression
Nonparanormal Data Transform
Causal Inference Algorithms: PC-Stable, PC-Max, CPC-Stable, FGES
Correlation Metrics: True positives, true negatives, false positives, false negatives
Precision
Recall
Matthews Correlation Coefficient
Structural Hamming Distance

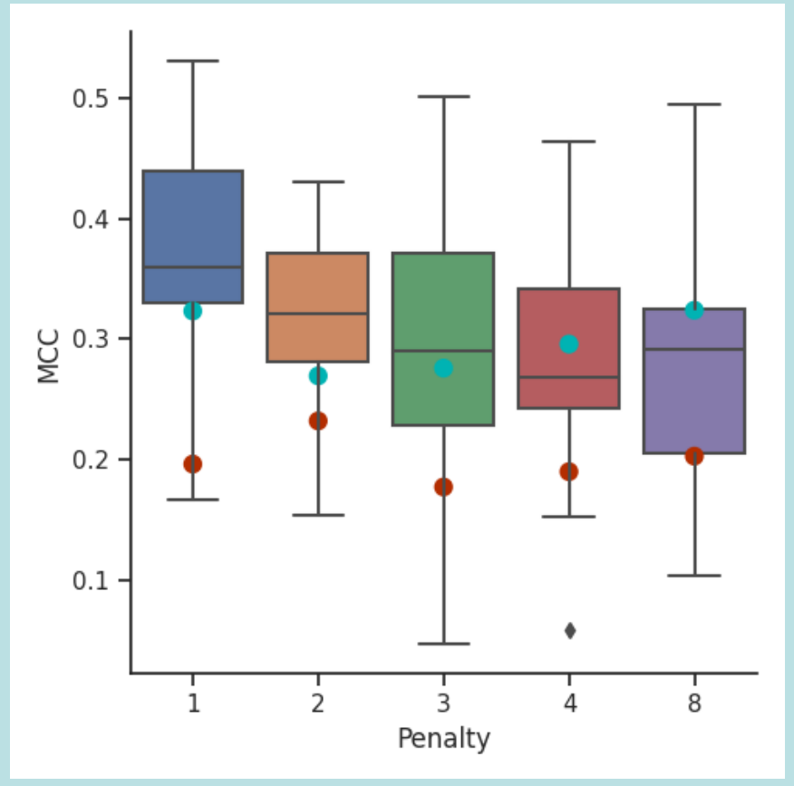
Figure 3. Method output obtainment and measurement process on simulated data.

Causal Graph Accuracy

Adjacency Matthews Correlation Coefficient (MCC)



PC algorithms

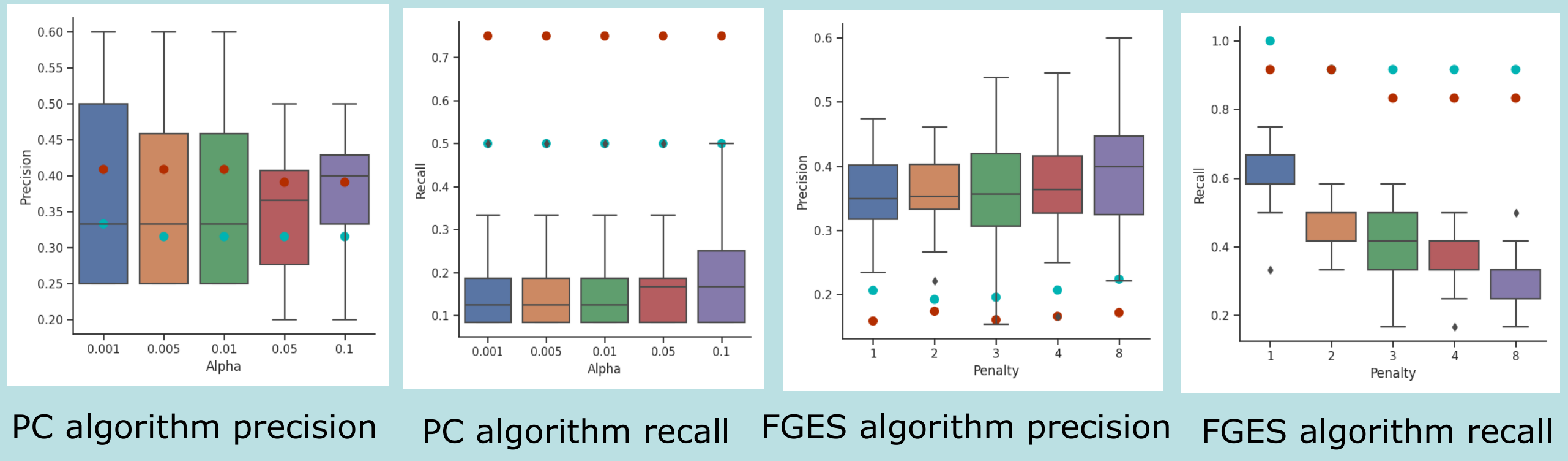


FGES

Figure 4. Adjacency MCC of simulated dataset's subsamples of size 100 and asymptotic data after being learned by PC-Stable, PC-Max, and CPC-Stable.

Figure 5. Adjacency MCC of simulated dataset's subsamples of size 100 and asymptotic data after being learned by FGES.

Adjacencies Precision/Recall

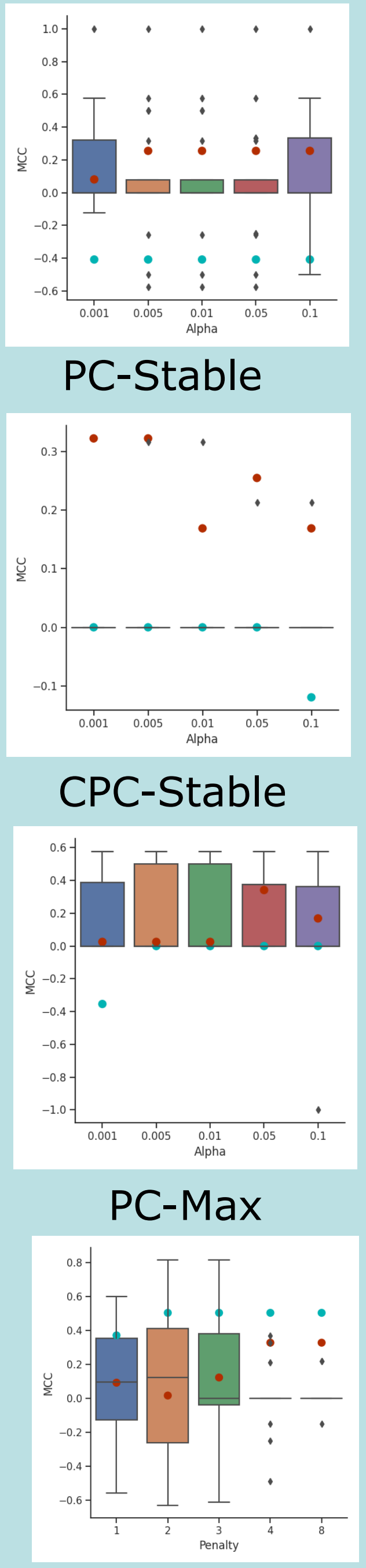


PC algorithm precision PC algorithm recall FGES algorithm precision FGES algorithm recall

Figure 6. Adjacency Precision/Recall of simulated dataset's subsamples of size 100 and asymptotic data after being learned by various algorithms and regression models.

● Asymptotic without nonparanormal data transform
● Asymptotic with nonparanormal data transform

Orientation MCC



PC-Stable

CPC-Stable

PC-Max

FGES

Figure 7. Orientation MCC of simulated dataset's subsamples of size 100 and asymptotic data after being learned by various algorithms and regression models.

Conclusions

The PC algorithms were found to be most successful on asymptotic data while the FGES algorithms were most successful on the subsampling. The FGES algorithms have higher recall with lower precision, resulting in many false positives being found alongside these true positives. It, therefore, performs better on subsamples of the data that don't contain as many undirected adjacencies in their undirected graph to potentially learn for the causal graph. The more balanced trade-off between precision and recall in the PC algorithms causes them to be better at learning asymptotic data while not as effective in detecting enough correct edges from subsamples.

Future Directions

Further research would include the application of these methods to predict the gene regulatory networks of non-simulated gene pathways. When time allows, the application of the methods to these pathways would enhance understanding of the algorithms' abilities to learn asymptotic case gene regulation networks. This would be attempted with CPC-Stable and PC-Max because they are the most effective algorithms for asymptotic case data.

Acknowledgements

TECBio REU @ Pitt is supported by the National Science Foundation under Grant DBI-1659611.

Major thank you to Tyler Lovelace and the Benos lab for the excellent mentorship in Computational Biology research and tireless work in guiding me throughout this project.

References

Bergen V., Lange M., Peidli S., Wolf F.A, & Theis F.J. (2019, Oct. 28). Generalizing RNA velocity to transient cell states for dynamic modeling. *BioRxiv*.
Friedman J., Hastie T., & Tibshirani R. (2007). Sparse inverse covariance estimation with graphical lasso. *Biostatistics, 9(3):432-41*. 10.1093/biostatistics/kxm045.
Gregg, R. W., Sarkar, S. N., & Shoemaker, J. E. (2019). Mathematical modeling of the cGAS pathway reveals robustness of DNA sensing to TREX1 feedback. *Journal of theoretical biology*, 462, 148–157. https://doi.org/10.1016/j.jtbi.2018.11.001
Zhao, T., Liu, H., Roeder, K., Lafferty, J., & Wasserman, L. (2012). The huge Package for high-dimensional undirected graph estimation in R. *Journal of machine learning research : JMLR*, *13*, 1059–1062.