Imputing Optimal Transport Barycenters of Patient Manifolds

Alexander Tong and Smita Krishnaswamy
Yale University Computer Science

Problem Statement

Single-cell technologies are expensive but form relatively smooth manifolds.

The goal is to reduce cost by inferring single-cell measurements based on a set of existing measurements.

This gives a continuous set of distributions from a set of coarse-grained samples.

Main Idea

Interpolate single-cell samples based on barycenters on an inferred sample manifold.

Background

Wasserstein Barycenters [1] Generalizes averaging of points to averaging of distributions based on a ground distance between points.

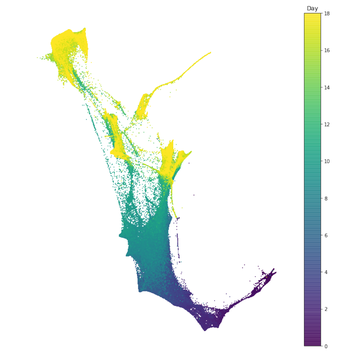
Allows interpolation of a distribution from a weighted set of distributions.

Optimal Transport in Single-cell Datasets [2] A ground distance is extremely important in single cells this is widely taken as the cosine distance between log or sqrt normalized distributions.

Tree distances give an alternate distance between distributions that is fast and generalizes pseudo-bulking, phEMD, and other commonly used comparisons.

Fig 7: Continuous normalizing flow architecture. A CNF models the instantaneous change at every cell state / time. This defines a flow over time from the initial density to t_1

Results



Task 1: Approximate 1D Time Manifold

Fig. 1: Paths learned from a gaussian to S distribution (a-b) for three methods: (c) a discrete OT optimization (d) an unregularized continuous normalizing flow (CNF) and (e) a regularized CNF using R_{energy}. Adding R_{energy} straightens paths by minimizing average kinetic energy over the paths.

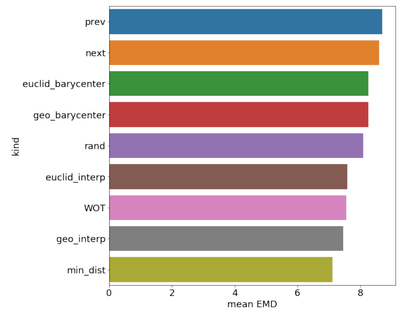


Fig. 2: For a simple 1d manifold embedded in 2d adding $R_{density}$ encourages following the manifold over time.

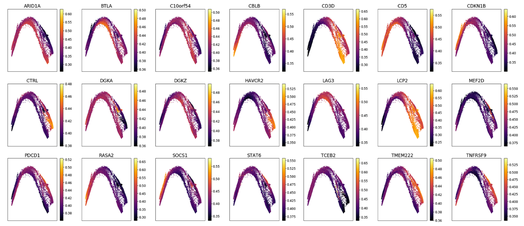


Fig. 4: Embryoid body data [3] with RNA velocity provides an instantaneous time direction of change at each cell. We regularize the cosine angle between RNA velocity and our model at all observed cells / timepoints.

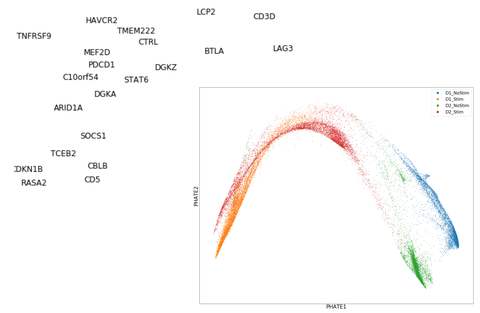


Fig. 5: TrajectoryNet models a density over time. At any time that density is tractable and at measured timepoints matches the measured distribution.

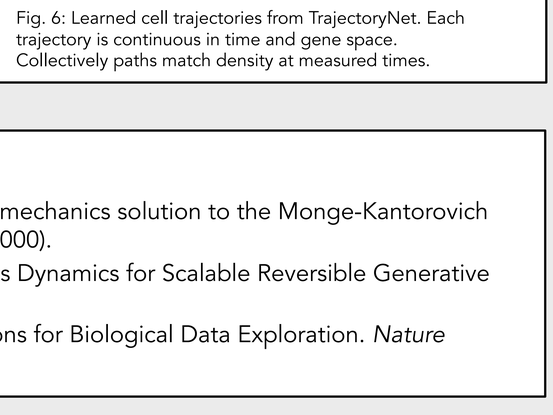


Fig. 6: Learned cell trajectories from TrajectoryNet. Each trajectory is continuous in time and gene space. Collectively paths match density at measured times.

Method

We approximate dynamic optimal transport over gene space with continuous normalizing flows.

We add the following regularizations to build in priors on cellular dynamics.

R_{energy} – Minimize kinetic energy and particle time derivative terms encouraging more Euclidean optimal and energy efficient paths.

$R_{density}$ – At random time interpolations minimize distance to K nearest neighbors for each point preferring trajectories close to the manifold of observed cells.

$R_{velocity}$ – Minimize angle between RNA velocity data and flow at every measured cell building in knowledge of RNA splicing data.

Finally, we learn a growth function **g** for modeling the cell growth rate using unbalanced optimal transport (see Fig. 3).

Conclusions

Our results relate continuous normalizing flows to dynamic optimal transport and create a flexible model for single cell population modeling.

Our model allows for interpolation of trajectories to unobserved cell states inferring future and past states of individuals from population data.

Further exploration is needed in efficient learning of *stochastic* and *unbalanced* models of cell populations.

References

[1] Benamou, J.-D. & Brenier, Y. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik* 84, (2000).

[2] Grathwohl, W. *et al.* FFJORD: Free-form Continuous Dynamics for Scalable Reversible Generative Models. in *ICLR* (2019).

[3] Moon, K. R. *et al.* Visualizing Structure and Transitions for Biological Data Exploration. *Nature Biotech.* (2019).

Further information

Email: alexander.tong@yale.edu

Code: github.com/KrishnaswamyLab

Supported by the Chan-Zuckerberg Initiative