A Diffusion-based Condensation Process for the Multiscale Analysis of Single Cell Data

Tobias Welp[1], Guy Wolf[2], Matthew Hirn[3], and Smita Krishnaswamy[1]

[1] *Dept. of Genetics, School of Medicine, Yale University, New Haven, CT, USA*
[2] *Applied Math Program, Dept. of Mathematics, Yale University, New Haven, CT, USA*
[3] *Dept. of Computational Mathematics, Science & Engineering, Dept. of Mathematics, Michigan State University, East Lansing, MI, USA*

1 Introduction

Clustering algorithms play an integral role in exploratory analysis of single cell data. The aim of these analyses is to find population structures that often exist at multiple scales. Previous approaches are limited to finding clusters in data, and these clustering methods come with serious limitation for exploration of this structure. For instance, they require commonly missing knowledge about the structure of the data such as the granularity or even the number of clusters. We propose a novel scalable diffusion-based data contracting process, that learns the shape of the data and contracts datapoints iteratively towards a diffusion-based data potential function. In each iteration, datapoints move to the center of gravity of their local neighbors as defined by a graph diffusion process, and this in turn alters the graph diffusion to reflect the new data positions. This process is carried forward by a non-homogenous Markov process representing the changing affinities between datapoints, along with changing granularities, and eventually collapses all data to a single point. However, each intermediate step in this process is a clustering at a particular granularity or abstraction level. Therefore, this process can naturally uncover a continual hierarchy of clusters.

We estimate particularly good points in the process to extract clusters using a novel nuclear-norm based technique that identifies metastable states in this contracting process.

We apply our algorithm on data of a clinical trial investigating the effects of anti-PD-1 immunotherapy in patients suffering from glioblastoma multiforme and present preliminary evidence that the algorithm finds meaningful structure in the dataset, i.e. it identifies rare, differentiated cell populations and a hierarchy of these populations, indicating that the algorithm will be of general use in biomedical research.

2 Condensation Process

Algorithm 1 gives an overview of the condensation process: Initially, a Markov affinity matrix is computed by applying a kernel (e.g. Gaussian or MGC [1]) to the pairwise Euclidean distances and by normalizing each row such that it becomes row-stochastic. The application of the kernel is impacted by parameter σ where a larger value for σ causes that a larger neighborhood is taken

Algorithm 1 Contraction Loop

repeat
  Compute Markov-diffusion affinities for each datapoint using σ.
  Diffuse (raise Markov-matrix to a power)
  Contract points towards the gravity center of their nearest neighbor in diffusion space.
  Compare volume of contracted space to that of previous iteration
  if metastable **then**
    determine cluster assignment by spectral clustering with number of clusters determined by high eigenvalues.
    increase σ.
  end if
until all samples are assigned to one cluster

into account. Next, this process is used to diffuse for t steps by powering the Markov matrix by t. Each datapoint is subsequentely moved towards the center of gravity of its neighbors by calculating a weighted sum of the current locations of the data points and the product of the diffused Markov matrix with the current locations. The stability of the process at any point is computed by comparing the nuclear norm of the diffused Markov matrix from the current iteration with that from the previous iteration. The actual cluster assignment can be obtained using a variety of clustering algorithms. At each metastable point the sigma or width of the Gaussian kernel is increased to look for cluster-structures at a higher level of granularity. As the distances between points of the same natural cluster assignment are very close in comparison to those between clusters, the “calling of clusters” is easier than without our condensation process. In our experimentation, we used spectral clustering with k equal to the number of eigenvalues close to one [3].

Figure 1 illustrates the condensation sequence on a data set with 1000 samples generated using a Gaussian mixture model. The upper row of figures shows how the samples are contracted to local density maxima that are iteratively merged. The bottom row of figures shows cluster assignments at the chosen iterations in the uncondensed space. Observe how the algorithm naturally finds a continuous hierarchy of clusters.

Figure 1: Illustration of a Condensation Sequence

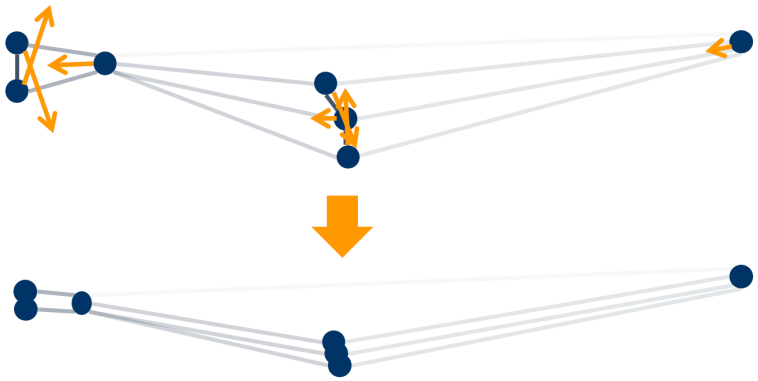

Figure 2: Illustration of the Condensation Move

Intuitively, the condensation move in each iteration can be interpreted as an interplay of gravity tension and inertia. The strength of the gravity force between samples depends on an applied kernel (e.g. Gaussian or MGC [1]). The application of the kernel assures that the condensation does not take place in open space but on the local manifold and as such reduces the impact of noise. Inertia balances gravity forces to prevent that all samples are immediately collapsed in the center of gravity. The idea of the condensation step is illustrated in Figure 2, where a choice of gravity forces between the samples is represented as gray lines (the darkness of lines is proportional to the strength of the gravity force) and the sums of gravity forces are represented by orange arrows.

3 Case Study: Glioblastoma

Glioblastoma multiforme (GBM) is the most malignant and aggressive brain cancer and has an incidence of two to three per 100,000 adults per year. The median survival time of a patient after

diagnosis is 15 months with standard treatment (total resection of tumor, chemo/radiotherapy) and less than three months otherwise [2]. GBM is thought to benefit from an immunosuppressive environment. One mechanism of T cell suppression involves the PD-1/PD-L1 immunomodulatory pathway. Anti-PD-1 treatment aims at blocking the pathway via antibodies.

We ran the condensation process on blood samples from patients undergoing anti-PD-1 immunotherapy. The left two plots in Figure 3 show the cluster results obtained at iterations 340, 389, and 450 during the condensation process where samples are displayed using tSNE [4]. Note how clusters of different granularity are found. The plot on the left of Figure 4 shows the development of the eigenvalues of the Markov matrix during the condensation process. The red lines indicate metastable iterations. Note how the increase in σ after the metastable points leads subsequently to a more effective condensation. The heat map on the right shows the expression of CyTOF channels for the found clusters and allows both for the identification of known cell populations such as cytotoxic lymphocytes (cluster 1) and T helper cells (cluster 2) as well as unique cell populations such as the potential regulatory B cells (cluster 3).

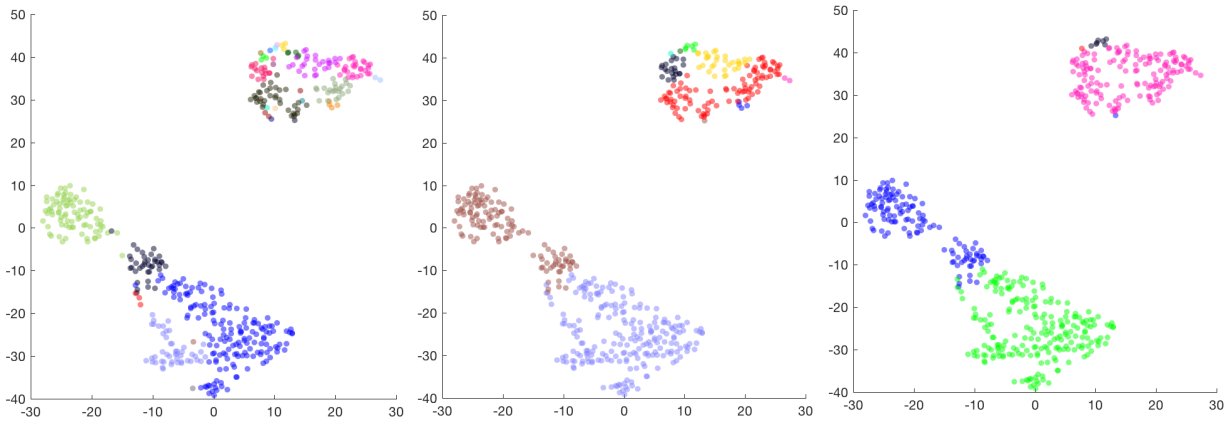

Figure 3 Cluster Assignments at Iterations 340, 389, and 450 in tSNE-Plot

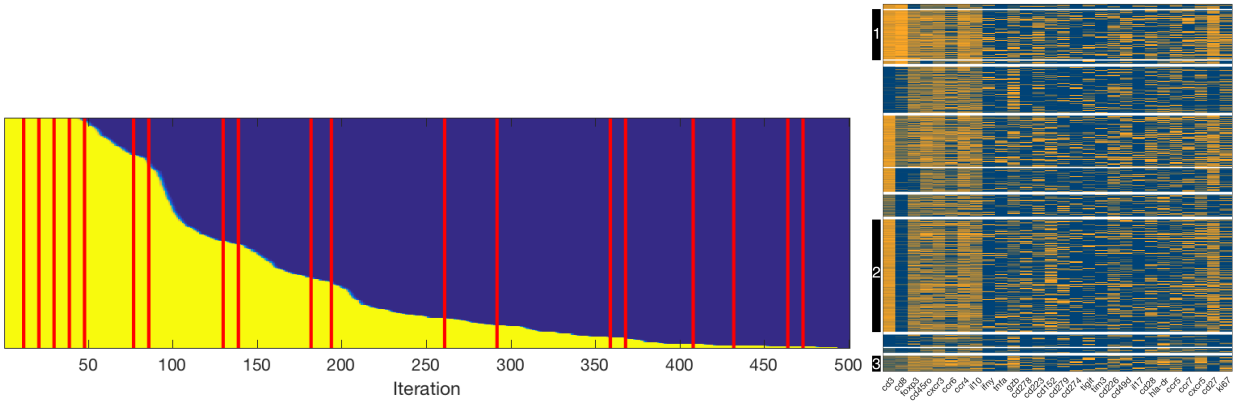

Figure 4 Development of Eigenvalues during the Condensation Process and heat map displaying identified clusters

References

[1] Amit Bermanis, Guy Wolf, and Amir Averbuch. Diffusion-based Kernel Methods on Euclidean Metric Measure Spaces. Applied and Computational Harmonic Analysis, 41(1): 190-213, 2016.

[2] Eric C. Holland. Glioblastoma multiforme: The Terminator. Proceedings of the National Academy of Sciences of the United States of America, 97(12):6242–6244, 2000.

[3] Ulrike von Luxburg. A Tutorial on Spectral Clustering. Statistics and Computing, 17(4):395–416, 2007.

[4] Laurens van der Maaten, Geoffrey Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9: 2579-2605, 2008.