MODELING LATENT FLOWS ON SINGLE-CELL DATA USING THE HODGE DECOMPOSITION

A PREPRINT

Kazumitsu Maehara

Division of Transcriptomics
Medical Institute of Bioregulation
Kyushu University
Fukuoka, Japan
kazumits@bioreg.kyushu-u.ac.jp

Yasuyuki Ohkawa

Division of Transcriptomics
Medical Institute of Bioregulation
Kyushu University
Fukuoka, Japan
yohkawa@bioreg.kyushu-u.ac.jp

ABSTRACT

Single-cell analysis is a powerful technique used to identify a specific cell population of interest during differentiation, aging, or oncogenesis. Individual cells occupy a particular transient state in the cell cycle, circadian rhythm, or during cell death. An appealing concept of pseudo-time trajectory analysis of single-cell RNA sequencing data was proposed in the software Monocle, and several methods of trajectory analysis have since been published to date. These aim to infer the ordering of cells and enable the tracing of gene expression profile trajectories in cell differentiation and reprogramming. However, the methods are restricted in terms of time structure because of the pre-specified structure of trajectories (linear, branched, tree or cyclic) which contrasts with the mixed state of single cells.

Here, we propose a technique to extract underlying flows in single-cell data based on the Hodge decomposition (HD). HD is a theorem of vector fields on a manifold which guarantees that any given flow can decompose into three types of orthogonal component: gradient-flow (acyclic), curl-, and harmonic-flow (cyclic). HD is generalized on a simplicial complex (graph) and the discretized HD has only a weak assumption that the graph is directed. Therefore, in principle, HD can extract flows from any mixture of tree and cyclic time flows of observed cells. The decomposed flows provide intuitive interpretations about complex flow because of their linearity and orthogonality. Thus, each extracted flow can be focused on separately with no need to consider crosstalk.

We developed ddhodge software, which aims to model the underlying flow structure that implies unobserved time or causal relations in the hodge-podge collection of data points. We demonstrated that the mathematical framework of HD is suitable to reconstruct a sparse graph representation of diffusion process as a candidate model of differentiation while preserving the divergence of the original fully-connected graph. The preserved divergence can be used as an indicator of the source and sink cells in the observed population. A sparse graph representation of the diffusion process transforms data analysis of the non-linear structure embedded in the high-dimensional space of single-cell data into inspection of the visible flow using graph algorithms. Hence, ddhodge is a suitable toolkit to visualize, inspect, and subsequently interpret large data sets including, but not limited to, high-throughput measurements of biological data.

The beta version of ddhodge R package is available at:
<https://github.com/kazumits/ddhodge>

Keywords Hodge decomposition · Single-cell RNA-seq · Trajectory analysis

1 Methods

1.1 Hodge decomposition on graph

Our ddhodge is built on a mathematical framework of the Hodge decomposition (HD) that provides an analogy of gradient, curl, and harmonic (cyclic) flows of the vector field. Jiang et. al proposed HodgeRank [1] as an application of HD to reconstruct global rankings from the pairwise relations of user ratings on Netflix movies. We developed the R package ddhodge which implemented the HD framework with the coboundary operators on a graph (grad, div, curl, and Laplacians) as introduced in HodgeRank and the introductory notes of Hodge Laplacians on graphs [2]. We briefly explain the results of HD and some notations used in this article.

The cochains C^0 , C^1 , and C^2 are the spaces of functions that assign a real value to the cartesian product of vertices in V ; $V \rightarrow \mathbb{R}$, $V \times V \rightarrow \mathbb{R}$ and $V \times V \times V \rightarrow \mathbb{R}$, respectively. The coboundary operators $\delta_0 = \text{grad}$ and $\delta_1 = \text{curl}$ are the maps between the cochains $C^k \rightarrow C^{k+1}$, and their adjoints $C^k \rightarrow C^{k-1}$ are $\delta_0^* = -\text{div}$ and $\delta_1^* = \text{curl}^*$ which are defined under an inner product $\langle \delta_k \xi, \eta \rangle_{k+1} = \langle \xi, \delta_k^* \eta \rangle_k$ where $\xi \in C^k$ and $\eta \in C^{k+1}$. They hold the relation $\delta_1 \circ \delta_0 = \delta_0^* \circ \delta_1^* = 0$. HD states that the orthogonal decomposition of C^k exists as:

$$\text{im} \delta_{k-1} \oplus \ker \Delta_k \oplus \text{im} \delta_k^*, \quad (1)$$

where $\Delta_k = \delta_{k-1} \circ \delta_{k-1}^* + \delta_k^* \circ \delta_k$ is Hodge Laplacian. If $k = 1$, the left term is the space of the gradient flow, and the central and right terms are those of divergence-free ($\text{div} X = 0$ iff $X \in \text{im}(\text{grad})^\perp$) harmonic and curl flow, respectively. If $k = 0$, $\Delta_0 = \delta_0^* \circ \delta_0 = -\text{div} \circ \text{grad}$ is known as graph Laplacian in graph theory, and C^0 is decomposed into $\ker \delta_0$ and $\text{im} \delta_0^*$. We illustrate the result of HD in Figure 1.

Figure 1: The scheme of Hodge decomposition.

The cochains C^0 , C^1 , and C^2 can be viewed as vector spaces of weighted nodes, edges, and triangles in graph G . Accordingly, the operators of grad, div, and curl are represented as matrices (linear maps between the weights of nodes, edges, and triangles; see [1, 2]). Using the HD setup on the graph, some flow decomposition problems of G are formulated as a simple least square problem, such as $\min_b \|\mathbf{y} - \text{grad} \mathbf{b}\|^2$ where $\mathbf{y}$ is the given flow (weighted edges). A solution $\mathbf{b}$ is obtained by the orthogonal projection of $\mathbf{y}$ onto the image of grad, and is interpreted as a potential field that produces gradient flows in $\mathbf{y}$. Then, the residual, consisting of the curl and harmonic flow, is divergence-free as explained above.

1.2 Designing edges through diffusion maps

The setup of HD was completed as above, but the problem of what kind of edges (direction and weights) would be appropriate for cells still remained. The Waddington epigenetic landscape [3] is a famous picture of cell lineage specification and has interested researchers who explore the kinetics that determine cell fates [4, 5]. This assumes that a stochastic process along the energy potential of a biochemical system is a natural way to model cell fate. We employed the diffusion process as a candidate of underlying dynamics that produces the time structure in single-cell (sc)RNA-seq data because it is a simple (requires few assumptions) and straightforward way to model the development of the state of cells (e.g., the gene expression profile) as entropy increases.

Diffusion maps [6, 7] are a suitable method of nonlinear dimension reduction that accounts for random walk on data points and was also applied for scRNA-seq data [8, 9]. In diffusion maps, a low-dimensional representation of the data

is obtained through the spectral decomposition of transition probability matrix M . M is obtained by normalizing the affinity matrix (each row sums to 1), which is constructed from a defined similarity of data points. We used Gaussian affinity $\exp(\|\mathbf{x}_i - \mathbf{x}_j\|^2 / \sigma_i \sigma_j)$ with the local scaling parameter σ_i that is determined by the distance to the n -th nearest neighbor from x_i [10]. Local scaling equalizes the different scales of within-cluster variations, which are seen in the mix of slow and fast transitions of states.

The vector field which corresponds to the diffusion process at a fixed time $t = s$ is a potential field that contains only gradient flow. We designed the edges as the gradient arose from a given potential (concentration) at time $t = s$. The initial condition at $t = 0$ (concentrations of starting cells) is given by any prior data knowledge, such as the identification of a cell population with marker genes, known cell groups before differentiation stimuli occur, or cell groups determined by clustering methods if an unbiased manner is preferable. The potential $\mathbf{u}$ at $t = s$ is calculated as:

$$\mathbf{u} = (1/m, 1/m, \dots, 1/m, 0, \dots, 0) M^s, \quad (2)$$

where the first m elements are selected as the starting cell. The edges are defined by $\text{grad} \mathbf{u}$ (i.e., the differences: $u_i - u_j$). We refer to this as the diffusion graph. The directed acyclic graph (DAG) can be constructed as above, though visualization and inspection of the graph to understand the diffusion dynamics remain difficult because it is a fully-connected complete graph (N^2 edges for N cells). We therefore keep only the edges of k -nearest neighbors (NN) on the diffusion coordinate. k -NN graph construction is frequently used to obtain sparse representations of graphs [11, 12]. However, pruning by k -NN criteria breaks several good properties in the original diffusion graph such as divergence and potential. We therefore appended the additional step after edge pruning to keep these properties while maintaining the sparsity of the graph as explained below.

2 Recovering the diffusion nature using a sparse network

The HD framework is intuitive, so helps us to consider several extensions such as placing constraints on the qualitative characteristics of graph G . To preserve the original diffusion, we therefore chose to keep divergence because this tells us the source and sink information of a flow. We constructed the objective function to minimize the loss of original divergence using the pruned graph as:

$$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{y} - \text{div} \mathbf{w}\|^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2, \quad (3)$$

where $\mathbf{y}$ is the divergence of the original diffusion graph, and div is the divergence operator of the pruned k -NN graph. Finally, we obtained the edges of the sparse graph consisting of pure gradient flow (DAG) by the orthogonal projection of the solution $\mathbf{w}$ onto $\text{im}(\text{grad})$ by:

$$-\text{grad} \Delta_0^\dagger \text{div} \mathbf{w}, \quad (4)$$

where $\Delta_0^\dagger$ is the Moore–Penrose pseudoinverse of the graph Laplacian. The residual parts of $\mathbf{w}$ consisting of curl- and harmonic-flow can be safely ignored because they do not contribute to divergence. We also note that once $\mathbf{w}$ is fixed, the recovered potential $-\Delta_0^\dagger \text{div} \mathbf{w}$ is not guaranteed to be close to the original one simply because the graph structure (Δ_0) is different in the k -NN graph.

Instead of the regression as above, the generalized lasso [13] formulation can be incorporated to directly solve the problem of the sparse construction of gradient flows. However, the computational cost may be infeasible.

2.1 10 steps for sparse diffusion graph construction

We summarize our construction of the sparse diffusion graph in 10 steps as follows:

1. Normalize the count matrix using medians of ratio normalization [14]
2. Perform variance-stabilizing transformation using $\log_2(\text{normalized count} + 0.5)$
3. Reduce to p dimensions using principal component analysis
4. Construct a transition probability matrix using Gaussian affinity with local scaling
5. Select starting cells and proceed the diffusion for s step (eq:2)
6. Construct a fully-connected diffusion graph equipped with gradient flow as edges
7. Calculate the divergence of the original diffusion graph
8. Prune the network using k -NN criteria at specified k on diffusion coordinates
9. Recover the original divergence using ridge regression (eq:3)
10. Construct a sparse diffusion graph using reconstructed gradient flow (eq:4)

Figure 2: Visualization with group labels (A), divergence (B), potential (C), and potential-ordered representative gene expression (D) with smoothed curves. The force-directed layout was used.

3 Results

We first present the visualization performance of ddhodge using scRNA-seq in [15]. The data are a collection of single-cell gene expression profiles collected at several time points in the direct reprogramming of mouse embryonic fibroblasts (MEF). Figure 2A shows visualization by the divergence-recovered sparse diffusion graph. Our ddhodge with diffusion successfully clustered the cell lineages without using the information of group labels (colored circles).

We also visualized the recovered divergence in the same layout (Figure 2B). The divergence informs us that the "source" cells that we provided as start cells are in red (divergence > 0) as expected. The negative divergence (blue circles) indicated the existence of corresponding "sink" cell populations at the terminal points of the path to the myocyte and neuron groups from MEFs. The gray cells (around zero divergence) can be reasonably interpreted as waypoints. The potential, which corresponds to the proportion of cells in each state at time $t = s$, can be used as "pseudo-time" information (Figure 2C, D). A human-friendly tree layout is accessible because the graph is purely DAG (Figure 3A).

As described above, it is straightforward to follow the path of reprogramming along the arrows in the sparse diffusion graph. The visualizations show multiple paths from MEFs to neurons and myocytes. The maximum flow of the graph can be used to extract a path that maximizes the flow between specified node pairs with given capacities of edges (gradients) (Figure 3B). The flow can be seen as a representative path in the reprogramming.

4 Discussion

We developed ddhodge as a framework to dissect the latent flows embedded in a collection of single cells. Our analysis of non-linear trajectories of high-dimensional single-cell data was transformed into the traversing of sparse diffusion graphs. The graphs enable the unveiling of directions in differentiation, which are visualized as simplified flows. These are useful to identify the path to a specific differentiation or the appearance of a particular cell population such as the stem cell state, intermediate state, or terminally-differentiated state.

Figure 3: Tree layout (A) and max-flow between a specified source and sink pair (B; large graph layout).

Monocle [16] has become a widely-used tool for scRNA-seq to trace lineage-specific differentiation. Several alternative approaches have been proposed for the trajectory analysis [11, 8, 17, 12], and these are comprehensively reviewed in [18, 19, 20]. Current approaches for trajectory analysis assume that the data contain either branched (generally a tree) or cyclic structures [21, 22, 23]. However, the only requirement for HD is that the graph is directed; it also has the unique advantage of the flexibility of handling any vector field.

The development of HD-based single-cell trajectory analysis is still in its infancy, although it has succeeded in extracting the flows in differentiation paths from scRNA-seq data. However, understanding the more realistic nature of temporal regulation of cells, including circadian rhythms and cell cycles, is still challenging because the information of space and time is not contained in single-cell profiles. Although the methodology uses the similarity of measured data points, this does not deduce causal directions among cells. Designing edges using emerging technologies of statistical causal inference [24, 25, 26, 27] combined with single-cell omics data other than transcriptomes, e.g., epigenomes [28] could pave the way for the reconstruction of causal flows including cyclic relations.

References

- [1] X. Y. Jiang, L. H. Lim, Y. A. Yao, and Y. Y. Ye. Statistical ranking and combinatorial hodge theory. *Mathematical Programming*, 127(1):203–244, 2011.
- [2] Lek-Heng Lim. Hodge laplacians on graphs. *arXiv preprint arXiv:1507.05379*, 2015.
- [3] C. H. Waddington. *The strategy of the genes; a discussion of some aspects of theoretical biology*. Allen & Unwin, London,, 1957.
- [4] J. E. Ferrell. Bistability, bifurcations, and waddington’s epigenetic landscape. *Current Biology*, 22(11):R458–R466, 2012.
- [5] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Y. Liu, S. Lin, P. Berube, L. Lee, J. Chen, J. Brumbaugh, P. Rigollet, K. Hochedlinger, R. Jaenisch, A. Regev, and E. S. Lander. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming (vol 176, pg 928, 2019). *Cell*, 176(6), 2019.
- [6] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.
- [7] Boaz Nadler, Stéphane Lafon, Ronald R Coifman, and Ioannis G Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21(1):113–127, 2006.

- [8] L. Haghverdi, M. Buttner, F. A. Wolf, F. Buettner, and F. J. Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13(10):845–, 2016.
- [9] Kevin R. Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel Burkhardt, William Chen, Antonia van den Elzen, Matthew J Hirn, Ronald R Coifman, Natalia B Ivanova, Guy Wolf, and Smita Krishnaswamy. Visualizing transitions and structure for biological data exploration. *bioRxiv*, page 120378, 2018.
- [10] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Proceedings of the 17th International Conference on Neural Information Processing Systems, NIPS’04*, pages 1601–1608, Cambridge, MA, USA, 2004. MIT Press.
- [11] S. C. Bendall, K. L. Davis, E. D. Amir, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan, and D. Pe’er. Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell*, 157(3):714–725, 2014.
- [12] M. Setty, M. D. Tadmor, S. Reich-Zeliger, O. Ange, T. M. Salame, P. Kathail, K. Choi, S. Bendall, N. Friedman, and D. Pe’er. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature Biotechnology*, 34(6):637–645, 2016.
- [13] R. J. Tibshirani and J. Taylor. The solution path of the generalized lasso. *Annals of Statistics*, 39(3):1335–1371, 2011.
- [14] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10), 2010.
- [15] B. Treutlein, Q. Y. Lee, J. G. Camp, M. Mall, W. Koh, S. A. Shariati, S. Sim, N. F. Neff, J. M. Skotheim, M. Wernig, and S. R. Quake. Dissecting direct reprogramming from fibroblast to neuron using single-cell rna-seq. *Nature*, 534(7607):391–5, 2016.
- [16] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Q. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–U251, 2014.
- [17] H. Matsumoto and H. Kiryu. Scoup: a probabilistic model based on the ornstein-uhlenbeck process to analyze single-cell expression data during differentiation. *Bmc Bioinformatics*, 17, 2016.
- [18] R. Cannoodt, W. Saelens, and Y. Saeys. Computational methods for trajectory inference from single-cell transcriptomics. *European Journal of Immunology*, 46(11):2496–2506, 2016.
- [19] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. *bioRxiv*, page 276907, 2018.
- [20] L. Kester and A. van Oudenaarden. Single-cell transcriptomics meets lineage tracing. *Cell Stem Cell*, 23(2):166–179, 2018.
- [21] N. Leng, L. F. Chu, C. Barry, Y. Li, J. Choi, X. M. Li, P. Jiang, R. M. Stewart, J. A. Thomson, and C. Kendzierski. Oscop identifies oscillatory genes in unsynchronized single-cell rna-seq experiments. *Nature Methods*, 12(10):947–950, 2015.
- [22] Z. H. Liu, H. Z. Lou, K. K. Xie, H. Wang, N. Chen, O. M. Aparicio, M. Q. Zhang, R. Jiang, and T. Chen. Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. *Nature Communications*, 8, 2017.
- [23] A. Scialdone, K. N. Natarajan, L. R. Saraiva, V. Proserpio, S. A. Teichmann, O. Stegle, J. C. Marioni, and F. Buettner. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, 85:54–61, 2015.
- [24] PO. Hoyer, D. Janzing, JM. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems 21*, pages 689–696, Red Hook, NY, USA, June 2009. Max-Planck-Gesellschaft, Curran.
- [25] M. H. Maathuis, M. Kalisch, and P. Buhlmann. Estimating high-dimensional intervention effects from observational data. *Annals of Statistics*, 37(6a):3133–3164, 2009.
- [26] S. Shimizu, P. O. Hoyer, A. Hyvarinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.
- [27] Xiaohai Sun, Dominik Janzing, Bernhard Schölkopf, and Kenji Fukumizu. A kernel-based causal learning algorithm. In *Proceedings of the 24th International Conference on Machine Learning, ICML ’07*, pages 855–862, New York, NY, USA, 2007. ACM.
- [28] A. Harada, K. Maehara, T. Handa, Y. Arimura, J. Nogami, Y. Hayashi-Takanaka, K. Shirahige, H. Kurumizaka, H. Kimura, and Y. Ohkawa. A chromatin integration labelling method enables epigenomic profiling with lower input. *Nat Cell Biol*, 21(2):287–296, 2019.