The Persistent Rayleigh Quotient for Feature Selection in Single Cell Transcriptomics

arXiv:2206.07760

Otto Sumray, Renee S. Hoekzema, Lewis Marsh, Xin Lu, Helen M. Byrne, Heather A. Harrington

Mathematical Institute, University of Oxford
Ludwig Institute for Cancer Research, University of Oxford

otto.sumray@ludwig.ox.ac.uk

UNIVERSITY OF OXFORD

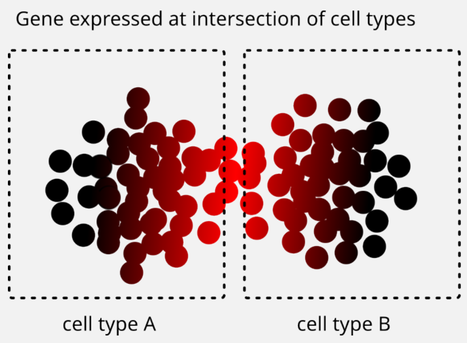
Background

The data: gene expression on a single cell level

- The *gene expression* of a gene is an measurement of how much each gene is being used.
- There are around 20,000 genes in the human genome. The gene expression of all these genes is the *transcriptome*.
- We can measure the transcriptome for thousands of cells in a sample simultaneously.
- This data can be thought of as thousands of points in $\mathbb{R}^{20000}$.

The problem: gene selection

- In biology, we want to understand this data in terms of genes.
- Which genes describe a particular cell type? Which genes drive differences in expression within a sample?
- The usual solution is to cluster the data into cell types and then use a statistical A/B test to find statistically significant genes that differentially expressed between clusters.
- This fails when the data has a high degree of continuity. What if a gene is expressed on the boundary of two clusters? What if the cell states do not naturally form clusters?



The solution: graph Laplacians and Rayleigh quotients

- Modelling the data as a graph (or simplicial complex etc.) allows us to not draw arbitrary boundaries between cell states. We need then a way to understand functions on the graph with respect to the topology.
- Given a graph $G = (V, E)$ the *graph Laplacian* L is defined as

$$L = D - A$$

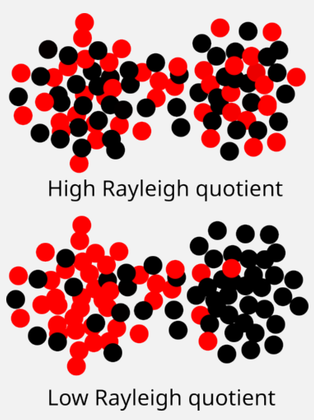
where D and A are the degree and adjacency matrices of the graph. The Laplacian is a discrete analogue of the Laplace-Beltrami operator on manifolds. Applying L to a function $g : V \to \mathbb{R}$ computes the difference between the value of g on a node and the average of the value of g on its neighbours.

- The *Rayleigh quotient* of the function g given L is defined as

$$R_L(g) = \frac{g^T L g}{g^T g} = \frac{\sum_{u \sim v} \|g(u) - g(v)\|^2}{\|g\|^2}.$$

The Rayleigh quotient is non-negative and measures how 'smooth' g is with respect to the graph G, with a smaller value $R_L(g)$ being smoother.

- Viewing each gene g as a real-valued function on V we can rank each gene by how smooth g with smoother genes being more interesting. This idea for feature selection was introduced in [HCN05] and applied to single cell data in [GYC19].



References

[DB13] Florian Dorfler and Francesco Bullo. "Kron Reduction of Graphs With Applications to Electrical Networks". In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 60.1 (Jan. 2013), pp. 150–163. ISSN: 1558-0806.

[GYC19] Kiya W. Govek, Venkata S. Yamajala and Pablo G. Camara. "Clustering-Independent Analysis of Genomic Data Using Spectral Simplicial Theory". In: *PLOS Computational Biology* 15.11 (22nd Nov. 2019), e1007509. ISSN: 1553-7358.

[HCN05] Xiaofei He, Deng Cai and Partha Niyogi. "Laplacian Score for Feature Selection". In: *Proceedings of the 18th International Conference on Neural Information Processing Systems*. NIPS'05. Vancouver, British Columbia, Canada: MIT Press, 5th Dec. 2005, pp. 507–514.

[Mu+20] Tianhao Mu et al. "Embryonic Liver Developmental Trajectory Revealed by Single-Cell RNA Sequencing in the Foxa2eGFP Mouse". In: *Communications Biology* 3.1 (1 3rd Nov. 2020), pp. 1–12. ISSN: 2399-3642.

[MWW21] Facundo Mémoli, Zhengchao Wan and Yusu Wang. "Persistent Laplacians: Properties, Algorithms and Implications". 26th July 2021. arXiv: 2012.02808 [cs, math].

[WNW19] Rui Wang, Duc Duy Nguyen and Guo-Wei Wei. "Persistent Spectral Graph". 11th Dec. 2019. arXiv: 1912.04135 [math].

[Yan+17] Li Yang et al. "A Single-Cell Transcriptomic Analysis Reveals Precise Pathways and Regulatory Mechanisms Underlying Hepatoblast Differentiation". In: *Hepatology (Baltimore, Md.)* 66.5 (Nov. 2017), pp. 1387–1401. ISSN: 1527-3350. pmid: 28681484.

Methods

Extending the Rayleigh quotient: introducing time

- Often biological data has an associated time component. Cell differentiation is the process in which one cell type (e.g a blood stem cell) becomes more specialised (e.g. a white blood cell).
- We want to be able to describe how gene expression relates to this time direction in our topology.

Kron reduction and the Persistent Laplacian

- We can view time data as a filtration on our graph and we would like to reduce our graph based on the superlevel sets of this filtration.
- For subsets $\alpha, \beta \subseteq V$ let $L[\alpha, \beta]$ be the submatrix of L with rows indexed by α and columns indexed by β. Under an appropriate reordering of the node labels, the Laplacian L has block form

$$L = \begin{bmatrix} L[\alpha, \alpha] & L[\alpha, \alpha^c] \\ L[\alpha^c, \alpha] & L[\alpha^c, \alpha^c] \end{bmatrix},$$

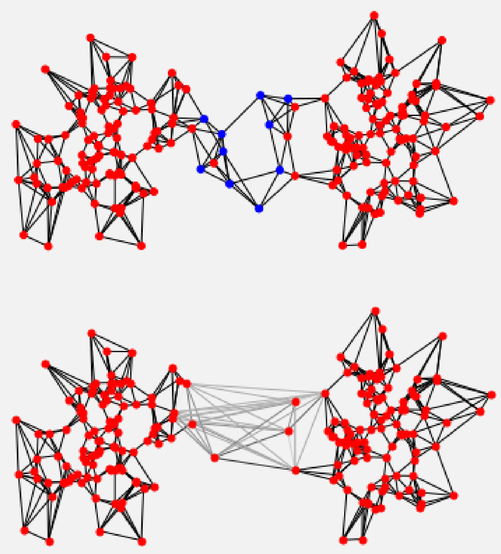
where $\alpha^c = V \setminus \alpha$ is the complement of α in V.

- The *Kron reduced Laplacian* [DB13] (or 0-degree *persistent Laplacian* [MWW21; WNW19]) of L with respect to α is the matrix

$$L_\alpha = L[\alpha, \alpha] - L[\alpha, \alpha^c] L[\alpha^c, \alpha^c]^{-1} L[\alpha^c, \alpha],$$

which is also known as the Schur complement $L/L[\alpha^c, \alpha^c]$.

- The Kron reduced Laplacian L_α is a *bona fide* Laplacian in the sense that there exists a weighted graph G_α with Laplacian L_α.



above: bottom graph is the Kron reduction of the top graph by the red nodes

The persistent Rayleigh quotient

- Suppose we have a filtration f on the nodes of the graph G, a function $f : V \to \mathbb{Z}$ with

$$\alpha(i) = \{v \in V : f(v) \leq i\}$$

being the sublevel set for each $i \in \mathbb{Z}$. For $i, j \in \mathbb{Z}$ with $i \leq j$ define

$$L_i^j = \left(L^{\alpha(i)}\right)_{\alpha(j)}$$

the (i, j)-persistent Laplacian, where $L^{\alpha(j)}$ is the Laplacian of the induced subgraph of G with nodes in $\alpha(j)$.

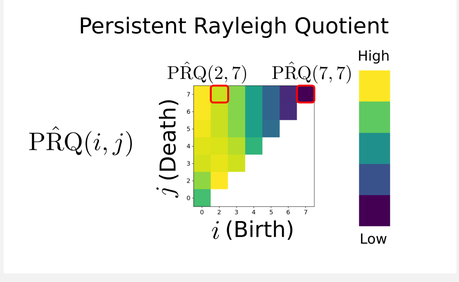
- We define the *persistent Rayleigh quotient* for a graph signal $g : V \to \mathbb{R}$ as

$$\mathrm{PRQ}(i, j)(g) = R_{L_i^j}(g) = \frac{\langle g, L_i^j g \rangle}{\langle g, g \rangle},$$

which is the Rayleigh quotient using the (i, j)-persistent Laplacian.
We further define the *normalised persistent Rayleigh quotient* to be

$$\widehat{\mathrm{PRQ}}(i, j)(g) = \frac{\langle g, L_i^j g \rangle}{\langle g, D_i^j g \rangle},$$

where D_i^j is the degree matrix of the graph corresponding to L_i^j.



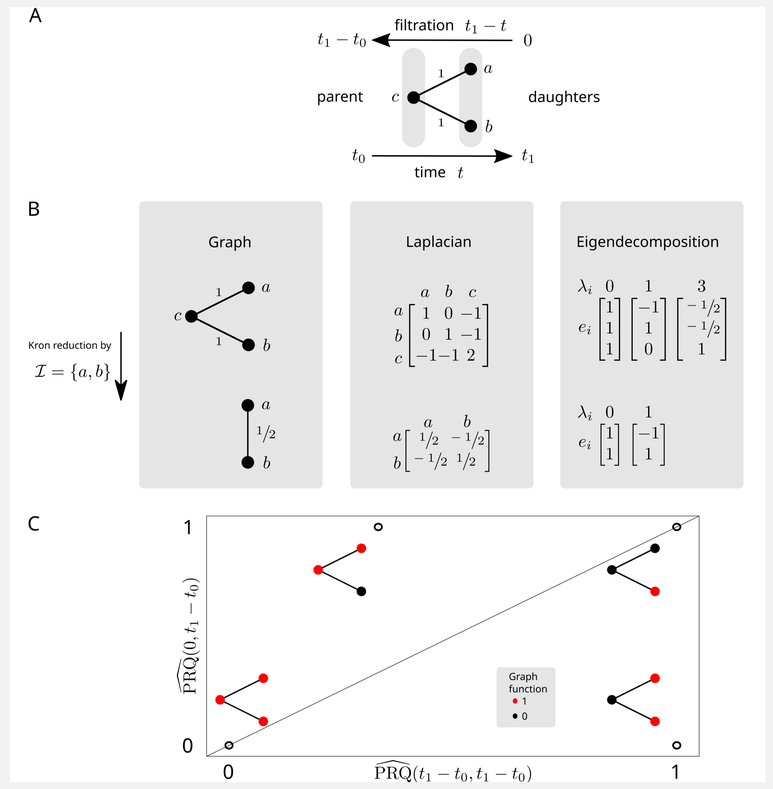
above: the PRQ gives a 2-dimensional non-negative value for each graph signal

Results

For cell differentiation

We apply the PRQ to cell differentiation processes.

- A) The model for the bifurcating differentiation process. A parent cell type c bifurcates over time to daughter cell types a and b.
- (B) The effects on the graph and graph Laplacian after applying the Kron reduction process to the daughter cells.
- (C) The normalised Rayleigh quotients of (x-axis) full Laplacian $L_{t_1-t_0}^{t_1-t_0}$ and (y-axis) persistent Laplacian $L_0^{t_1-t_0}$ for binary functions on the graph, representing genes, separates these based on relevance to the bifurcation.



On mouse foetal liver cells

We apply the PRQ on public data obtained from sampling developing mouse liver cells over the course of 8 days [Yan+17].

- (C) We plot these values for each gene for $(i = 7, j = 7)$ on the x-axis and $(i = 2, j = 7)$ on the y-axis.
- Selected for display (A,B,D,E) are top differentially expressed genes from [Mu+20].
- Genes *Tubb5*, *Mdk*, and *Igfbp1* are expressed in parent and one daughter cell lineage, hepatoblast to (A) cholangiocyte or (B) hepatocyte and lie above the diagonal.
- Genes *Aldob* and *Mt2* are expressed in both daughter cell types but not in the parent cell type (D), and lie below the diagonal.
- Genes *Ahsg* and *Fabp1* are only expressed in one daughter cell type (E) and lie on the diagonal.