

Topological Methods for Single Cell Data



Otto Sumray

New College

University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy in Clinical Medicine

Trinity 2024

Statement of Originality

Chapters [1](#) and [2](#) are based on, and expanded upon, the paper [\[40\]](#) that was published in *Entropy* in 2022, with collaborators Renee Hoekzema, Lewis Marsh, Thomas Carroll, Helen Byrne, and two of my supervisors Xin Lu and Heather Harrington. My main focus within the paper was the persistent Rayleigh quotient and its application.

Chapters [3](#), [4](#) and [5](#) are based on my 2024 preprint [\[83\]](#) with my supervisors Vidit Nanda and Heather Harrington.

Chapter [6](#) is also my original work.

Acknowledgements

I give my great thanks to my supervisors Xin Lu, Heather Harrington, and Vidit Nanda for their support, advice, and inspiration throughout my DPhil. I thank Ludwig Cancer Research for providing the funding and facilities to complete my DPhil. I would like to thank the members of the Lu Lab, ASB group, and the Centre for Topological Data Analysis for being such friendly, funny, and curious people who have been a pleasure to be around. To all my friends in Oxford and beyond, thank you for making life so enjoyable. I particularly want to thank Duncan for keeping me within the realms of sanity throughout these four years. To Wally: 'woof'. To my parents and my sister thank you for being part of my life, and all the love you have given me. To Renee, you are my rock. To the little one, I hope you will have such a love of learning as I do.

The contour eludes me.

– Paul Cézanne

Contents

Introduction	1
1 The Rayleigh Quotient for Feature Selection	21
1.1 Background on Graph Laplacians	21
1.2 The Laplacian Score	26
1.2.1 Graph Structure from Data	26
1.2.2 Feature Selection	27
1.3 Eigenscores	30
1.4 Multiscale Laplacian Score	33
1.5 Persistent Rayleigh Quotient	36
1.5.1 Kron Reduction	38
1.5.2 Persistent Rayleigh Quotient	43
Code Availability	45
2 Gene Selection	46
2.1 Single-Cell Transcriptomics	46
2.2 Standard Analysis of scRNA-seq	48
2.2.1 Cell classification	48
2.2.2 Differential Gene Expression	49
2.2.3 Limitations of Current Analyses	50
2.2.3.1 Models of Cell Type	50

2.3	Eigenscores and MLS Applied to PBMC and T Cell Data	52
2.3.1	Datasets and Preprocessing	52
2.3.1.1	Previous Results on T Cell Data	54
2.3.2	Eigenscores	55
2.3.2.1	The Geometry of PBMC Genes via Eigenscores	55
2.3.2.2	Eigenscores for Analysing Data with Continuous Structure: T Cells	58
2.3.3	Multiscale Laplacian Score	60
2.3.3.1	Multiscale Laplacian Score of PBMC Data	61
2.3.3.2	Multiscale Laplacian Score of T Cell Data	62
2.4	Persistent Rayleigh Quotient	64
2.4.1	Application to Cell Bifurcation	64
2.4.2	PRQ Applied to Mouse Foetal Liver Cell Data Set	66
3	Quiver Laplacians	76
3.1	Quiver Representations	77
3.1.1	Global Sections	79
3.2	The Quiver Laplacian	81
3.3	Approximate Sections	83
3.4	Spectral Stability of Quiver Representations	84
3.4.1	Stability of Subspace Arrangements	90
3.5	Eigenvalue Inequalities for Quiver Operations	91
3.5.1	Sheaves over Quivers	91
3.5.2	Operations on the Quiver	93
4	Feature Selection and Compatibility	101
4.1	Introduction	101
4.2	Compatible Features	103

4.2.1	Overview	103
4.2.2	From Feature Selection to Quiver Representations	105
4.2.3	Compatibility and Sections	108
4.2.4	Local Compatibility	109
4.2.5	Bicompatible features	111
4.2.6	Opposite Representations	114
4.2.7	Reduction of the Feature Selector Problem	115
4.2.8	Remarks on computation	118
	Code Availability	119
5	Single-Cell Chromatin Accessibility	120
5.1	Chromatin accessibility	120
5.2	Measuring Chromatin Accessibility: ATAC-seq	121
5.3	Analysing scATAC-seq with Quiver Laplacians	123
5.3.1	Pre-processing the data	124
5.3.2	Generating the quiver	125
5.3.3	Building the representation	125
5.4	Results	126
6	Quiver Laplacians and Persistence Modules	130
6.1	Background on Persistent Homology	130
6.1.1	Multiparameter persistence	133
6.2	Harmonic Persistent Homology	134
6.3	Indecomposables of Unitary Quiver Representations	137
6.4	The Spectra of Persistence Modules	139
6.4.1	Multiparameter persistence	142
A	Appendix	145
A.1	Removing Edges and Vertices	145

A.2	Naturalisers	147
A.2.1	Proof of Lemma A.2.4	151
A.3	Classical Results on Eigenvalues of Hermitian Matrices	156

List of Figures

1	Pipeline for finding consistent features from feature selection.	14
2	Application of feature selection pipeline to single-cell chromatin accessibility data.	20
1.1	Two graph signals	29
1.2	Demonstration of eigenscores on toy graph.	32
1.3	Variation of information at different scales.	35
1.4	Two signals on a tree I	36
1.5	Two signals on a tree II	38
2.1	Differential gene expression on two graph signals.	51
2.2	Clustering of T cell data	54
2.3	Gene space of PBMC data via eigenscores	58
2.4	Eigenscores compared to differential gene expression	59
2.5	Gene space of T Cell data via eigenscores	60
2.12	Graph from mouse foetal liver cell dataset	67
2.6	Multiscale Laplacian score on PBMC data	70
2.7	Multiscale Laplacian score on T cell data	71
2.8	Seurat clusters on PBMC data	72
2.9	Eigenscore ranks for PBMC data	73
2.10	Expression of relevant marker genes in the T cell data set.	73

2.11	The persistent Rayleigh quotient on bifurcation model	74
2.13	The persistent Rayleigh quotient applied to mouse foetal liver cell data	75
3.1	An example of the reduction of a quiver Q with $V = \{v\}$	98
4.1	Two features defined on a dataset equipped with a decomposition into two overlapping subsets.	103
4.2	Example of different compatibility conditions.	109
4.3	An illustration of the floret based at u and its representation.	111
4.4	Illustration of the merged quiver	113
5.1	Quiver Representation from Chromatin Accessibility data	128
5.2	Approximate compatible features after peak-calling	129
6.1	The functions ψ_1 and ψ_2	140
6.2	Harmonic representatives of persistence	141
6.3	Comparison of the spectra of persistence modules	143
A.1	The floret functor	152

Introduction

This thesis lies at the confluence of two growing fields: single-cell sequencing analysis and applied topology. The former was brought about by recent technological developments allowing the high-throughput sequencing of genomics, transcriptomics, proteomics, epigenetics, and other ‘omics’ of individual cells from a sample, giving biologists an unprecedented view into what makes cells tick. This technology is producing large, high-dimensional datasets, presenting significant data science challenges, but also challenging how we model biology: we can no longer view cells as belonging to static cell types, but rather as part of a continuum of dynamic cell states. Applied topology on the other hand has emerged as a field of mathematics in the 21st century, taking ideas from long-standing fields in pure mathematics, such as algebraic topology and differential geometry, and applying them to understand complex non-linear data, inspired by related fields such as network science.

This thesis is concerned with one particular story in applied mathematics, the discrete Laplacian. The Laplace operator is fundamental in many areas of mathematics, from describing diffusion, wave, and heat equations in physics, to algebraic geometry and topology via Hodge theory, to the signal processing that much of our telecommunication relies on. Discrete versions of the Laplacian began with the simplest of topological models, the graph, and in recent years have been extended to more general objects, such as hypergraphs [13], simplicial complexes [30],

and sheaves [37], as well as finding application in dimensionality reduction [6] and the growing field of geometric deep learning. The discrete Laplacian connects topological or combinatorial properties of objects, e.g. the number of connected components, with more approximate notions, such as the ‘bottleneckness’ of a graph, making it attractive when working with data, where everything is approximate.

In application to single-cell data, this thesis promotes the use of more sophisticated topological models, such as filtered graphs and quiver representations, to model the continuous nature of cell states, and analyse the behaviour of cells.

This thesis is organised as follows: Chapters 1 and 2 are based on the paper [40] published in *Entropy* in 2022. Chapter 1 introduces classical and novel methods based on the graph Laplacian to rank graph signals by coherence with the graph topology. Chapter 2 applies these methodologies to single-cell transcriptomics datasets where the graph signals are the expressions of genes on a cell state network. Chapters 3, 4 and 5 are based on the 2024 preprint [83]. Chapter 3 introduces the quiver Laplacian, a new generalisation of the graph Laplacian to quiver representations, in order to approximate global sections of the representation and to define the spectrum of a quiver representation. The chapter further proves several spectral stability results under transformations of the representations, and natural operations on the underlying quiver. Chapter 4 describes the problem of feature selection on data whose constituent points fall into groups. The chapter proposes a solution using the formalism from Chapter 3 to find features that are approximately consistently relevant across the groups. Chapter 5 applies the methodology from Chapter 4 to single-cell chromatin accessibility data in order to group accessible genomic regions that are either locally or globally relevant across cell types. Finally, Chapter 6 proposes using the spectrum of the quiver Laplacian as an invariant for (multi-)persistence modules with inner product, to gain richer information from the topology and geometry of data.

Part I: Graphs and Single-Cell Transcriptomics

Single-cell transcriptomics is a relatively new technique in biology and bioinformatics in which transcripts of genes, also known as reads, are measured for many individual cells. This technique allows one to estimate how much each gene is being expressed in individual cells, which us about cell type, cell state and cellular activity. Typically there are upwards of 20,000 genes considered, and thousands to tens of thousands of cells are sequenced from a single sample. Many of these 20,000 genes will not be biologically relevant for the research questions that we want to pose in the analysis of the particular sample, where we often want to uncover biological heterogeneity leading to an understanding of biological function of observed expression patterns. One therefore applies feature selection to select the genes that are most biologically relevant or interesting for the given dataset.

Currently the standard method for selection of genes applied in single-cell transcriptomics is differential gene expression. In this method, the data is first split into subpopulations, interpreted as cell types, by means of a clustering algorithm. These cell types are labelled using known marker genes. Subsequently, genes that vary significantly between the subpopulations are ranked as more relevant.

However, in many biological situations, it turns out that cell types are not completely discrete. For example in the case of cell differentiation, cells can be in a transitory state in the process of differentiating from a more stem-cell like type to a more mature cell type. The advent of single-cell transcriptomics indeed is highlighting that the discrete cell type paradigm presents a too restrictive point of view. By using a discrete model of cell types and states and applying differential gene expression, we cannot consider the more continuous and transitory situations or capture the gene expression patterns relevant to them. Thus, there is a need for methods of gene selection that can also deal with continuous patterns of cell states.

In single cell analysis, networks (graphs) are commonly used as a model for

underlying connectivity between data points, for example in building a visualisation of the data (e.g., by means of the non-linear dimension reduction algorithm UMAP), when calculating so-called pseudotime or performing lineage tracing. Indeed, the clustering algorithms used to establish cell types are often based on community detection from networks constructed from the single-cell data. Hence it is natural to take a network as the underlying topological model on which we wish to perform gene selection. The next step is then to formulate appropriate methods for feature selection and analysis of signals on networks that can cope with continuous and discrete structure in the network on an equal footing, and be applied in the situation of single-cell transcriptomics as a replacement and enhancement of the classical method of differential gene expression. This is what is done in Part I of the thesis. Chapter 1 introduces three novel methods for feature selection of signals on networks based on spectral graph theory and working with the Laplacian. In Chapter 2 these methods are shown to recover biologically relevant genes when applied to published single-cell transcriptomics datasets.

Chapter 1: The Rayleigh Quotient for Feature Selection

A graph signal is a real-valued function from the vertices of a graph or network. For example, in the case of single-cell transcriptomics, if the nodes of the network represent individual cells, the expression pattern of a single gene is a graph signal.

Given a graph signal f , the Dirichlet energy measures the total variation of f with respect to the underlying graph. The Dirichlet energy can be measured using the graph Laplacian, which is a vertex-by-vertex matrix defined as the difference of the degree matrix and the adjacency matrix of the graph. The Rayleigh quotient is a normalisation of the Dirichlet energy, measuring again how varying the signal is with respect to the underlying graph.

In [39], the Rayleigh quotient was further modified to form the Laplacian score,

which was proposed as a method for feature selection. This feature selection method was applied to single-cell transcription (scRNA-seq) data in [32], where it was further extended to higher order notions via simplicial complexes. While this method was able to sort gene signals based on their coherence with respect to the underlying graph and without dependence on a subdivision into clusters or cell types, it did not differentiate between genes expressed, for example, uniformly on the whole dataset versus coherently on different components of the graph. In practice, what genes are interesting depends on the scale within the dataset. It is therefore interesting to be able to give multirankings of relevant genes, where scale and location within the dataset can be taken into account without predefining a clustering on the data. To this end, we present three novel spectral methods for giving multirankings of graph signals.

The method of *eigenscores* splits the Rayleigh quotient into individual contributions in the different low-frequency eigenvector directions. This presents a way of sorting the signals both for coherence with respect to the graph as well as frequency, corresponding more or less to size of the coherent patches. Eigenscores can furthermore be used to build a visualisation of graph signals.

The *multiscale Laplacian score (MLS)* takes as an input the random walk Laplacian, which we can use to define a scale-dependent version of the Laplacian score. Natural scales at which to assess and compare the MLS of a graph signal are indicated by considering variation of information, which measures similarity of graph communities found by community detection based on continuous time random walks.

The third method, the *persistent Rayleigh quotient (PRQ)*, can be applied to situations in which the dataset comes equipped with a filtration. In applications this may for example be the case if measurements were taken at different time points. The graphs representing the dataset at different filtration levels and the associated

Laplacian matrices can be compared to each other by means of Kron reduction, also known as the persistent Laplacian. Using the persistent Laplacian in the Rayleigh quotient gives rise to the PRQ, which measures the coherence of graph signals with respect to the underlying graph as well as the filtration, and can separate signals that play a different role with respect to the time-varying graph.

Chapter 2: Gene Selection

In this chapter the three novel Laplacian-based feature selection techniques for ranking graph signals are successfully applied to the problem of gene selection in several public scRNA-seq datasets.

Eigenscores and MLS are applied to two datasets, one of peripheral blood mononuclear (PBMC) cells that is commonly used as a reference and another dataset of T cells that was previously used in [32]. Standard preprocessing techniques are applied to construct a network from the datasets. The eigenscores gene ranking on the PBMC data recovers results from differential gene expression (DGE) while also highlighting new biologically meaningful genes. These are relevant either to larger regions of the dataset, consisting of multiple clusters in the clustering assumed in DGE, or they are highly expressed on smaller regions not marked as clusters for the DGE. In the T cell dataset, which has more continuous structure, eigenscores pick up for example a gene that is highly expressed on a trajectory between two clusters of cells, and is not picked up by DGE. Similarly, by comparing the MLS scores at different resolutions, one can compare genes for how highly they are expressed at different scales, and find genes that are expressed on a sub-community in a larger community but not on the whole.

The persistent Rayleigh quotient is applied to a mouse foetal liver cell dataset, measured over several days of cell differentiation, which endows the dataset with the structure of a filtration. The PRQ is able to separate out genes that play a

different role in the differentiation process, separating genes that are activated in the parent as well as one of the daughter cell types from those that are expressed in both daughter cell types or in only one.

Part II: Feature Selection and Chromatin Accessibility

Feature selection is an integral part of pipelines in bioinformatics, and data analysis in general. In Chapter 2 we show feature selection being needed for the more than 20,000 genes in the human genome, however, it is even more necessary in the next biological dataset we consider where the features potentially number in the billions.

Feature selection algorithms typically expect features on a single input dataset and then return the features that are relevant to those data points. However, when studying single-cell data, we are interested in the properties of *subpopulations* of the data, often rare or abnormal subpopulations. Now, the features that are relevant for the whole dataset may not be relevant for these subpopulations of the data, and vice versa.

A possible workaround, as exhibited in [71], is to first break up the data into natural groupings via clustering or community detection, with each group corresponding to a cell type. Feature selection is then applied to each group. However this presents a problem of recombining the results on subpopulations to a coherent feature selection output on the entire dataset. In the paper [71], features are collated across groups by selecting the most significant feature from each set of similar features.

Problems of combining data processing outputs on different subsets such as this are common in single-cell bioinformatics. Often we wish to process data that consists of many heterogeneous subpopulations, such as in the above feature se-

lection problem, or when removing technical batch effects, or combining multiple datasets from different modalities. Current methods for combining results on substructures often involve applying various combined techniques with many parameters without an underlying mathematical foundation or formalism, and generally few guarantees on how different parameters affect the output. For the case of the feature selection on data with subpopulations one might ask how the result is affected by

1. different parameters of the feature selection algorithm,
2. different clustering methods, or
3. different parameters for combining features.

This part of the thesis attempts to provide a mathematical foundation for these types of problems and answer some of the stability questions listed above. We use the mathematical structure of a quiver representation to represent vector spaces of selected features attached to data points, which will correspond to subpopulations of cells (and the intersections and unions thereof) in the eventual application on chromatin accessibility data. We then frame the problem of finding features that are relevant across different subpopulations as the question of finding approximate global sections of a quiver representation. In order to compute these approximate sections, we define a Laplacian for quiver representations and analyse its spectrum. Chapter 3 sets out the mathematics of the novel quiver Laplacian and proves stability results for the spectrum with respect to changes in the underlying quiver or transformations of the quiver representation. Chapter 4 formulates a theoretical framework for the problem of feature selection on data with subpopulations. Chapter 5 applies this framework and methodology to a case study of chromatin accessibility data, recovering biologically relevant features as approximate sections.

Chapter 3: Quiver Laplacians

The aim of this chapter is to set out the mathematical foundations for finding approximate global sections of quiver representations by means of defining the quiver Laplacian and proving results about its spectrum.

The quiver representation (valued in finite dimensional inner product spaces) will describe the data structure we obtain after applying a feature selection algorithm to different parts of a dataset. We formulate the process of combining these features across the different parts of the dataset as finding global sections of the quiver representation. However, since we are working with data, we can never exactly find these sections, indeed they might not even exist. We therefore introduce a notion of an approximate section of a quiver representation:

Definition (3.3.1). For a representation $\mathbf{A}_\bullet$ of quiver Q , an ε -approximate section of $\mathbf{A}_\bullet$ is a unit-norm vector in the total space of $\mathbf{A}_\bullet$ such that its Dirichlet energy is no greater than ε .

In order to compute approximate sections, we introduce a generalisation of the graph Laplacian for quiver representations:

Definition (3.2.2). For a representation $\mathbf{A}_\bullet$ of a quiver Q the quiver Laplacian is a self-adjoint positive semi-definite operator on the total space of $\mathbf{A}_\bullet$ whose null-space is the space of global sections of $\mathbf{A}_\bullet$.

The Dirichlet energy of an eigenvector of this quiver Laplacian is the corresponding eigenvalue, thus approximate global sections of the representation can be found by computing eigenvectors of the associated Laplacian with small eigenvalues.

The remainder of this chapter addresses two questions: how does the spectrum of the quiver Laplacian change under

1. transformations of the quiver representation and
2. certain natural operations on the underlying quiver.

In the direction of the first question we prove the following results.

Theorem (3.4.4). *Let $\mathbf{A}_\bullet$ and $\widehat{\mathbf{A}}_\bullet$ be representations of a quiver Q . Each transformation $\tau : \mathbf{A}_\bullet \rightarrow \widehat{\mathbf{A}}_\bullet$ gives a bound on the 1-Wasserstein distance between the spectra of $\mathbf{A}_\bullet$ and $\widehat{\mathbf{A}}_\bullet$.*

This bound will be smaller if τ is closer to an isometric isomorphism. The quiver representations we construct in the subsequent Chapters 4 and 5 are an incarnation of what we term a subspace arrangement of a quiver. Applying Theorem 3.4.4 to transformations arising as orthogonal projections between subspace arrangements provides the central theorem of the chapter; a stability result for the spectrum of subspace arrangements of quivers.

Theorem (Corollary 3.4.5). *Suppose $\mathbf{S}$ and $\mathbf{T}$ are two subspace arrangements of a quiver Q in an inner product space. If, for each vertex v of Q , the respective subspaces $\mathbf{S}_v$ and $\mathbf{T}_v$ are within $\epsilon < \pi/2$ in the Grassmannian metric, then the 1-Wasserstein distance between their spectra is bounded by*

$$\frac{|Q_1|}{m} \left[f(\epsilon) \cdot c + g(\epsilon) \cdot (m - c) \right]$$

for particular trigonometric polynomials f and g . Here, $|Q_1|$ is the number of edges of Q , m the maximum of the total dimensions of $\mathbf{S}$ and $\mathbf{T}$, and c is the sum of the differences in dimension of $\mathbf{S}_v$ and $\mathbf{T}_v$ over all vertices v of Q .

Note that $c = 0$ if and only if the spaces for each vertex for $\mathbf{S}$ and $\mathbf{T}$ are equidimensional. In this case, as $g(0) = 0$, the first term in our bound vanishes and the second term approaches 0 as $\epsilon \rightarrow 0$.

For the second question, on how certain operations on the underlying quiver affect the spectrum of the Laplacian, we provide the following collection of propositions.

Theorem. *Let $\mathbf{A}_\bullet$ be a representation of Q . The following inequalities hold for appropriate integers i . The constants k, r, w_1, w_2 which appear below depend on $\mathbf{A}_\bullet$ as well as the modification performed to create a new representation $\mathbf{A}'_\bullet$ of a new quiver Q' .*

1. (Proposition [A.1.1\(a\)](#)) *If Q' is obtained by removing edges from Q and $\mathbf{A}'_\bullet$ is the restriction of $\mathbf{A}_\bullet$ to Q' , then*

$$\lambda_{k+i}(\mathbf{A}_\bullet) \leq \lambda_{k+i}(\mathbf{A}'_\bullet) \leq \lambda_{k+r+i}(\mathbf{A}_\bullet).$$

2. (Proposition [A.1.1\(b\)](#)) *If Q' is obtained by removing vertices (and incident edges) from Q and $\mathbf{A}'_\bullet$ is the restriction of $\mathbf{A}_\bullet$ to Q' then*

$$\lambda_i(\mathbf{A}_\bullet) - w_1 \leq \lambda_i(\mathbf{A}'_\bullet) \leq \lambda_{k+i}(\mathbf{A}_\bullet) - w_2.$$

3. (Proposition [3.5.5](#)) *If Q' and $\mathbf{A}'_\bullet$ are the result of performing an admissible homotopy of Q , then*

$$\varphi^{-2}\lambda_i(\mathbf{A}_\bullet) \leq \lambda_i(\mathbf{A}'_\bullet) \leq \varphi^2\lambda_i(\mathbf{A}_\bullet)$$

where φ is the golden ratio.

4. (Proposition [3.5.9](#)) *If Q' and $\mathbf{A}'_\bullet$ are the result of performing an admissible Kron reduction of $\mathbf{A}_\bullet$, then*

$$\lambda_i(\mathbf{A}_\bullet) \leq \lambda_i(\mathbf{A}'_\bullet) \leq \lambda_{i+r}(\mathbf{A}_\bullet).$$

As well as giving a form of stability of the spectrum with respect to the underlying quiver, these operations are necessary in Chapter [4](#) in order to simplify

a quiver representation to a less computationally taxing representation with isomorphic global sections while giving bounds on the change of the approximate sections.

Chapter 4: Feature Selection

In this chapter we formulate the problem of recombining the results of feature selection on subpopulations in terms of quiver representations and finding their approximate sections. Figure 1 gives a graphical overview of the pipeline for compatible feature selection using the quiver Laplacian.

We assume that we are given a dataset X of data points and a vector space of features on these points. A ‘black box’ feature selection algorithm essentially takes in a set of features defined on a subset of X and outputs a (vector) subspace of the features on X . An example of such a feature selection algorithm is Principal Component Analysis (PCA).

Importantly, the space of features on X should come equipped with a notion of inner product. This does not need to have been derived from X itself. In the chromatin accessibility example which we explore in Chapter 5, the inner product between features will be based on genomic distance.

We will subsequently make use of a cover $\mathcal{U} = \{U_1, U_2, \dots, U_n\}$ of X into subsets representing subpopulations. This cover gives rise to a topology on X . The subsets could represent a clustering of X , but allowing overlaps of clusters gives access to more ‘continuous’ topologies. From this cover we define a quiver. Applying the feature selection algorithm on a set of features restricted to each element of the cover of X endows them with a vector space of selected features. Comparison maps between the vector spaces are given by orthogonal projection using the inner product structure, and hence we obtain the structure of a quiver representation.

We now aim to understand how features that are relevant for subsets of X (ele-

ments of the cover) relate to each other. For a collection of features, indexed by the vertices of the quiver, we describe two notions of compatibility – local and global. Briefly, a collection of such features is locally compatible if a feature that has been selected with respect to some subset $\bigcap_{i \in \sigma} U_i$ of X continues to be selected on larger subsets $\bigcap_{j \in \tau} U_j$ where $\tau \subset \sigma$. This indicates that this feature is in a sense robust to the different scales in the cover $\mathcal{U}$. Restricting to the locally compatible features of the quiver representation gives a nicer structure – that of a cellular sheaf. On the other hand, the globally compatible features are exactly global sections of the representation: collections of features indexed by the vertices of the quiver who project onto each other in a consistent fashion along the quiver. Such a collection of features is desirable as it indicates a coherent group of features that are relevant across the dataset and at multiple scales.

We show how the quiver representation can be modified such that its space of global sections correspond to the features that are both locally and globally compatible. However, generically the space of such features will be trivial – hence we will aim to find features that are approximately local and global. There is another motive for finding approximate sections: we may have selected relevant features on different subsets that are highly similar (under the provided inner product), and finding approximate global sections allow us to ‘glue’ together these similar features, as exhibited in Chapter 5.

We prove in Theorems 4.2.14 and 4.2.15 that the representations used to compute approximately compatible features can be reduced to much simpler representations, while providing bounds on how the eigenvalues of the Laplacian changes, allowing us to perform these computations more efficiently.

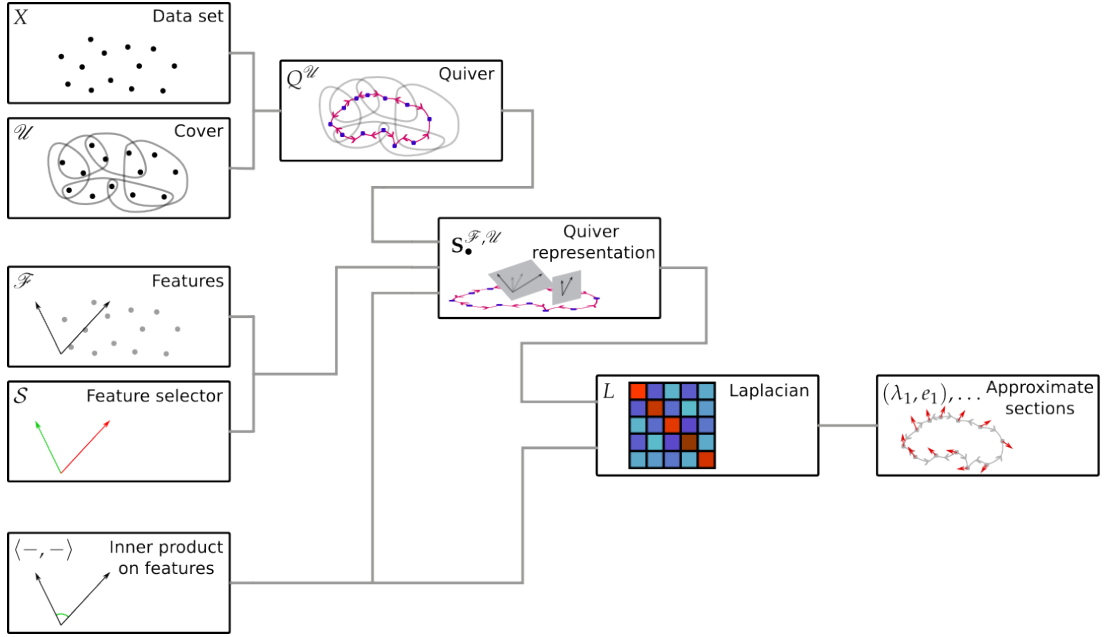


Figure 1: Pipeline for finding consistent features from features selection. The inputs consist of a dataset X , a cover $\mathcal{U}$ of X , a set of features $\mathcal{F}$ on X , a feature selection algorithm $\mathcal{S}$, and an inner product $\langle -, - \rangle$ on features of X . The cover $\mathcal{U}$ is used to create a quiver Q with vertices corresponding to subsets of $\mathcal{U}$ and their non-trivial intersections. The feature selection algorithm $\mathcal{S}$ is applied to the features $\mathcal{F}$ restricted to each element of $\mathcal{U}$, to construct a representation $\mathbf{S}_{\bullet}^{\mathcal{F}, \mathcal{U}}$ of Q . A quiver Laplacian L is then constructed from $\mathbf{S}_{\bullet}^{\mathcal{F}, \mathcal{U}}$, from which approximate sections are computed as the eigenvectors of L with eigenvalue no greater than some desired value ε .

Chapter 5: Chromatin Accessibility

This chapter now applies the methodology set out in Chapter 4 to single-cell chromatin accessibility data.

Within the cell nucleus, DNA is condensed in the form of chromatin, formed primarily of the DNA itself together with histone proteins that act as spools for the DNA to wind around. The regions of this structure where the DNA is less condensed, referred to as open chromatin, can be accessed by proteins that affect transcription, such as transcription factors, which control genes being ‘on’ or ‘off’. The accessibility of the different chromatin regions forms part of the epigenome,

which explains why cells with the same DNA can have different phenotypes, for example expressing themselves as different cell types.

Chromatin accessibility can be measured by various technologies such as ATAC-seq. These technologies use certain enzymes to split the DNA at open sites, giving fragments of DNA for each cell, which are then sequenced. The end points of these fragments indicate where the chromatin is open. This data is difficult to work with: it is highly sparse and its dimensionality is theoretically 3 billion, the number of base-pairs in the human genome. It is therefore necessary to perform feature selection.

Peak callers are algorithms that use a population of cells to infer sites or regions that are significantly accessible within the population. This presents a method of feature selection for ATAC-seq data. In order to combine the outputs of peak calling on subpopulations of a single-cell dataset, we can apply the methodology outlined in Chapter 4.

Figure 2 gives an overview of how the feature selection and integration pipeline presented in Figure 1 is applied in practice. The resulting approximate sections of the quiver representation now correspond to accessible regions in the genome that are compatible across cell types. We apply this pipeline to a large dataset of tumour-infiltrating T cells as shown in Figures 5.1 and 5.2 on pages 128 and 129, and find biologically relevant signals.

We observe that the approximate sections closest to *bone fide* global sections each exhibit a collection of relevant genomic positions, one for every vertex of the quiver or subpopulation or cell type, that lies entirely within a 100 base-pair region in the genome. Essentially this illustrates that the algorithm has managed to glue together a genomic region relevant across all cell types that is peak-called at a slightly different positions in the different subpopulations. This indicates that each of these approximate sections correspond to a genomic region that is significantly

accessible for all sequenced T cell subtypes. For example, we find a position close to a gene that is known to be involved in the cell cycle, hence it makes sense that this region is accessible for all cell types.

At higher eigenvalues, i.e. approximate sections that are further away from being global sections, we also find collections of relevant genomic regions that are only supported on a topologically consistent subset of the population. Hence we see that the feature selection integration pipeline is able to pick up features that are coherently found to be relevant for a subpopulation. In particular, we find a region close to a gene that is known to be an important regulator for differentiation and proliferation in the cell types that the subpopulation in question is enriched in.

Chapter 6: Quiver Laplacians and Persistence Modules

In the final chapter, we present work in progress that applies the quiver Laplacian defined in Chapter 3 to a major tool in the field of topological data analysis, namely persistence modules. Persistence modules are¹ representations of A_n quivers, derived from applying homology to a filtered topological space (persistent homology).

Endowing persistence modules with an inner product can add geometric information that normal persistent homology cannot pick up. For example, they are able to detect the rate at which topological features change. Persistence modules that are endowed with an inner product are called harmonic persistence modules.

However, unlike ordinary persistence, harmonic persistence modules do not admit a decomposition into a nice set of indecomposables, even in the case of one-parameter persistence modules. Even in the one-parameter case we therefore face the problems that plague multiparameter persistence, where such decompositions are also not feasible.

¹when discretised and with one parameter

Rather than aiming for finding a decomposition, we can therefore aim to find an invariant that will give us information about the harmonic persistence module. Examples of invariants used in multiparameter persistence are the Hilbert function and the rank invariant.

A harmonic persistence module is an example of a quiver representation, hence we can define the associated quiver Laplacian. The spectrum of the quiver Laplacian is an invariant of the harmonic persistence module up to isometric isomorphism.

We show in a toy example case study that the spectrum of the quiver Laplacian is able to pick up geometric information that the underlying normal persistence module cannot pick up. We moreover show that it can differentiate multiparameter persistence modules that the rank invariant cannot tell apart. Hence studying the spectrum of the quiver Laplacian for harmonic persistence modules presents an interesting new tool for topological data analysis.

Future directions

One of the general themes of this thesis is to build mathematical frameworks for how to deal with substructures and multiple scales in single-cell bioinformatics.

The structure of a quiver representation, introduced in Chapter 3 to describe feature selection in the presence of subpopulations, has the potential to be exploited in a much more general setting in bioinformatics. For example, it could potentially be applied to the problem of integrating datasets – either those of the same modality, where we have several datasets from different runs of the same experiment and we want to integrate these together while accounting for batch effects, or where the different datasets have different modalities, such as simultaneous measurements of transcriptomic and epigenetic information.

The reason why quiver representations are an interesting data structure for integrating datasets is that the category of quiver representations valued in finite-dimensional Hilbert spaces has a rich structure. In particular it is both finitely complete co-complete. This means that if we are given two datasets in the form of quiver representations, we can use a pushout or pullback to join the two together.

Besides bioinformatics, another interesting avenue to explore would be the application of quiver representations and their spectra to network theory more generally. Networks and graph theory are ubiquitous in their application, from graph neural nets to neuroscience. Quivers and quiver representations are generalisations of graphs or networks. One can use a quiver representation for example to join two graphs together along certain collections of vertices. There could be fruitful connections for applying quiver representations in network theory, and quiver representations could provide a meaningful data structure within network theory more generally.

Another interesting direction to explore is the connection of the quiver Laplacian to spectral graph theory in the field of combinatorics. Eigenvalue bounds for the graph Laplacian are studied in spectral graph theory, and our results for the quiver Laplacian Propositions [A.1.1\(a\)](#), [A.1.1\(b\)](#), [3.5.5](#) and [3.5.9](#) in particular extend and improve these results on the level of graphs. It could be the case that studying the quiver Laplacian and its spectrum further could have application within spectral graph theory.

A famous question by Lipman Birs, popularised by Marc Kac in 1966 [\[45\]](#), asked whether we can hear the shape of a drum, in other words, whether there exist surfaces on which the Laplacian operator has the same spectrum. The answer turns out to be yes, i.e. isospectral surfaces exist and we cannot hear the shape of a drum. Similarly, isospectral graphs are known to exist [\[94\]](#). An interesting follow-up question is to ask: can we hear the shape of a quiver representation?

In particular, one would wonder whether there exist isospectral non-isomorphic quiver representations if we fix the underlying quiver and dimensions, and if so, how many?

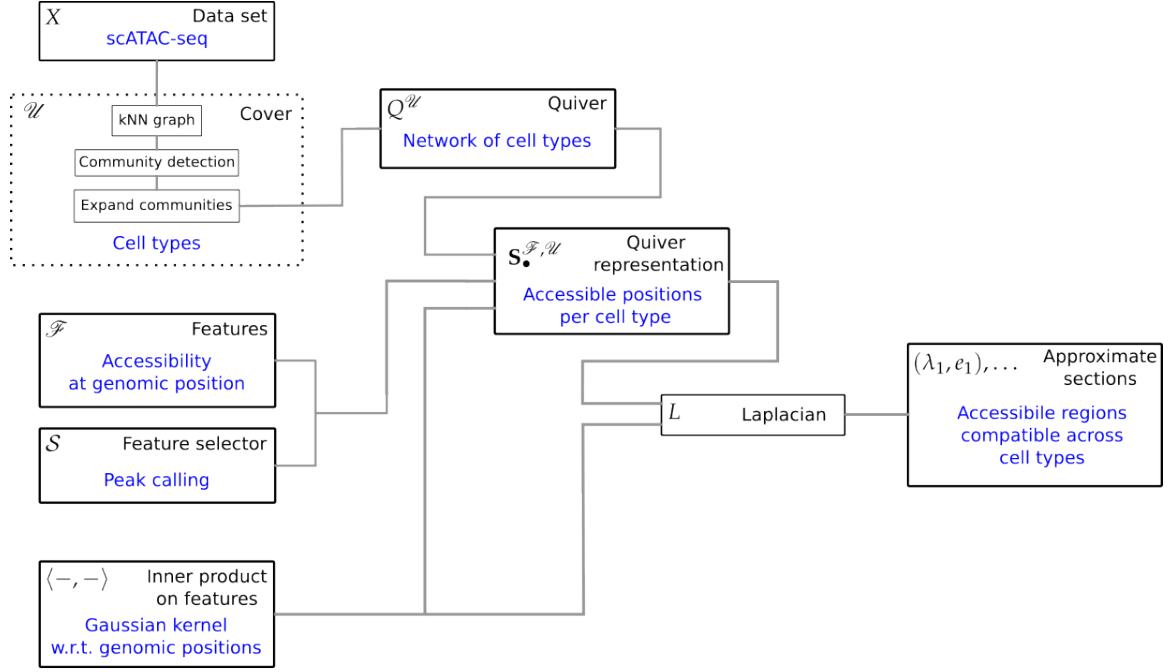


Figure 2: Application of feature selection pipeline to single-cell chromatin accessibility data. The cover $\mathcal{U}$ is constructed by the following sub-pipeline: First, a kNN graph is constructed from the scATAC-seq data using a count matrix constructed from 2.5kbp windows. Then community detection is performed on the graph, and these communities are expanded by including all direct neighbours, resulting in a cover, with each covering element corresponding to a cell type. The set of features are the accessibility status of each genomic position for each cell. The peak calling algorithm is MACS2. The inner product is a Gaussian kernel between genomic positions with a bandwidth of 1kbp. The quiver now represents a network of cell types, and the quiver representation associates to each cell type the significantly accessible positions. From this information we construct our quiver Laplacian which allows us to compute the approximate sections, corresponding to compatible accessible regions.

Chapter 1

The Rayleigh Quotient for Feature Selection

The following chapter expands on the work in [\[40\]](#).

1.1 Background on Graph Laplacians

Graphs

A (finite) *simple graph* $G = (V, E)$ consists of a (finite) set of vertices V and a set of edges E , where each edge $e \in E$ is a set $\{u, v\}$ of two distinct vertices in V . If $\{u, v\} \in E$, we will denote this relation by $u \sim v$, and call u and v neighbours. For this chapter, we will only consider simple graphs, and simply refer to them as graphs.

We may also want weight the strength of the relation given by an edge; for a graph $G = (V, E)$, an edge weighting is a function $w : E \rightarrow \mathbb{R}_{>0}$. Given such a w , we will call G a weighted graph.

Graph Laplacians

For a vertex $v \in V$, its degree d_v is the number of vertices $u \in V \setminus \{v\}$ such that $v \sim u$. Suppose G has n vertices. Then the degree matrix D associated to a graph is the $n \times n$ diagonal matrix indexed by its vertices such that the (u, u) entry is d_u . The adjacency matrix A is the $n \times n$ matrix indexed by vertices whose (u, v) entry is 1 if $u \sim v$, else 0. If G has an edge weighting, we instead define the degree of each vertex v as $\sum_{u \sim v} w(\{u, v\})$ and give the (u, v) -entry of the adjacency matrix $w(\{u, v\})$ if $u \sim v$ and 0 otherwise.

The *graph Laplacian matrix*, sometimes known as the Kirchoff Laplacian, L associated to the graph G is defined as the difference

$$L = D - A. \quad (1.1)$$

An equivalent, and perhaps more enlightening definition, is via a difference operator. First let us define two vector spaces associated to a graph: the space of real vertex functions

$$C_0(G) := \{f : V \rightarrow \mathbb{R}\} \cong \bigoplus_{v \in V} \mathbb{R}$$

and the space of real edge functions

$$C_1(G) := \{f : E \rightarrow \mathbb{R}\} \cong \bigoplus_{e \in E} \mathbb{R}.$$

We will also call a real vertex function a *graph signal*. We give the space $C_0(G)$ a standard inner product structure by

$$\langle f, g \rangle = \sum_{v \in V} f(v)g(v).$$

Now in order to define a difference operator, we need a choice of direction

along the edges of G . An orientation on a graph is defined by two function $s, t : E \rightarrow V$ defining the *source* and *target* of each edge. Then for such a graph, define its *boundary matrix*¹ as an operator

$$B : C_0(G) \rightarrow C_1(G)$$

given in matrix form with rows indexed by the edges E , and columns indexed by vertices V where

$$B_{e,v} = \begin{cases} 1 & \text{if } s(e) = v \\ -1 & \text{if } t(e) = v \\ 0 & \text{otherwise.} \end{cases}$$

We can apply the boundary matrix to a graph signal, giving

$$(Bf)_e = f(s(e)) - f(t(e))$$

for each edge $e \in E$, hence B can be thought of as a discrete analogue of the gradient ∇ .

Now, the graph Laplacian as defined in Equation 1.1 can be written as

$$L = B^T B$$

thus the Laplacian is symmetric and positive semi-definite. Note that the Laplacian is invariant under the choice of orientation: changing orientation amounts to pre-multiplying B with a diagonal matrix O whose diagonal entries are ± 1 , hence $O^T O = \text{Id}$.

If the graph is weighted, we can view the edge weights as giving an inner

¹There are various names for this matrix, including a the signed incidence matrix or the difference operator. Perhaps we should call this the coboundary operator, inline with algebraic topology.

product on $C_1(G)$, where the norm of an edge is its weight. If W is the matrix of weights, i.e. the adjacency matrix of the weighted graph, then the dual of B with respect to this new inner product is the matrix $B^T W$, hence the Laplacian of the weighted graph can be defined as $B^T W B$.

Applying L to a graph signal $f : V \rightarrow \mathbb{R}$ gives us

$$\begin{aligned} (Lf)_v &= \sum_{u \sim v} (f(v) - f(u)) \\ &= d_v f(v) - \sum_{u \sim v} f(u) \\ &= d_v \left[f(v) - \frac{1}{d_v} \sum_{u \sim v} f(u) \right], \end{aligned}$$

hence $(Lf)_v$ is proportional to the average value of f on the neighbours of v . This can be viewed as a discretisation of the Laplace operator Δ .

If the graph is non-regular, then vertices with large degrees will have a large influence on $\|Lf\|$, so we may wish to normalise the Laplacian to account for this. The *symmetrically normalised Laplacian* [14] $\mathcal{L}$ is defined as

$$\mathcal{L} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} A D^{-1/2}$$

where D is the degree matrix of G .

The Rayleigh Quotient and Laplacian Spectrum

For a graph G , the boundary matrix B applied to a graph signal f measures the difference of the signal across edges of G . If we compute the squared norm of Bf we have

$$\|Bf\|^2 = \langle Bf, Bf \rangle = \langle f, B^T Bf \rangle = \langle f, Lf \rangle.$$

The value $\langle f, Lf \rangle$ can be rearranged to

$$\mathcal{E}_L(f) := \langle f, Lf \rangle = \sum_{u \sim v} (f(u) - f(v))^2. \quad (1.2)$$

We will call this value the *Dirichlet energy* of f . The Dirichlet energy measures the total variation of f with respect to the underlying graph. From Equation 1.2, it is clear that the Dirichlet energy of f is zero if, and only if, f is constant on each connected component of G . Normalising the Dirichlet energy by the norm of f gives us the *Rayleigh quotient* of a graph signal:

$$R_L(f) = \frac{\mathcal{E}_L(f)}{\|f\|^2}.$$

As L is symmetric and positive semi-definite all its eigenvectors are real and non-negative. Label the eigenvalues (with multiplicity) as

$$0 \leq \lambda_0(L) \leq \dots \leq \lambda_{n-1}(L)$$

where n is the number of nodes of G . As L is symmetric, $C_0(G)$ can be given an orthonormal basis of eigenvectors, $e_0, \dots, e_{n-1}$ where e_i is the eigenvector corresponding to λ_i .

The Rayleigh quotient has a strong connection to the eigenvalues of L :

$$R_L(e_i) = \frac{\langle e_i, Le_i \rangle}{\|e_i\|^2} = \lambda_i.$$

Let $\mathbf{1}$ be the constant graph signal with value 1. Then, as discussed earlier, $\mathcal{E}_L(\mathbf{1}) = 0$ hence $\mathbf{1} \in \ker L$, therefore the first eigenvalue $\lambda_0(L)$ must be zero, and the multiplicity of the zero eigenvalue gives the number of connected components of the graph.

A useful property of the symmetrically normalised Laplacian $\mathcal{L}$, is that its spectrum is bounded – all its eigenvalues lie in the interval $[0, 2]$ (see [14]).

1.2 The Laplacian Score

1.2.1 Graph Structure from Data

When analysing a finite dataset $X \subset \mathbb{R}^n$, a common assumption is that X is sampled from a low dimensional manifold embedding in $\mathbb{R}^n$. The manifold hypothesis may not always be accurate, e.g. data may exhibit tree-like structure, however building a topological model for X based on local information has several advantages. Firstly, restricting to a nice class of objects allows the use of ideas from the established fields of topology, algebraic topology, differential geometry, and network theory. Further, we may assume that very close data points represent meaningfully similar states or objects, while data points further away may represent very different states or objects. From a purely computational view, using only local information is a form of sparsification, reducing the complexity of analysis.

Graphs are perhaps the simplest topological models. there are various methods to construct a graph G from a dataset $X \subset \mathbb{R}^n$. In the following examples, let d be a metric on X .

Proximity graph The proximity graph² of X with parameter ε is given by a graph where the set of vertices V is X and there exists an edge between x_i and x_j if $d(x_i, x_j) \leq \varepsilon$.

k -nearest neighbour graph (k NN) For an integer $k > 0$, the k NN graph of X has vertices given by X and where there is an edge between x_i and x_j if either x_i one of the nearest k neighbours of x_j or vice versa.

²This is the 1-skeleton of the Vietoris-Rips complex of X .

Nerve Suppose $\mathcal{U}$ is a cover of X , a collection of subsets of X such that their union is X . The *nerve graph* of $\mathcal{U}$ is the graph with vertices given by $\mathcal{U}$ and where there is an edge between U_i and U_j in $\mathcal{U}$ if $U_i \cap U_j$ is non-empty.

Mapper Mapper is an algorithm introduced in [77] that constructs a graph, an approximation of the Reeb graph, from a dataset X , a function $h : X \rightarrow \mathbb{R}^m$, and a cover $\mathcal{U}$ of $\mathbb{R}^m$: each node of the graph G is a cluster (under some clustering algorithm) within the preimage $h^{-1}(U_i)$ of some $U_i \in \mathcal{U}$. Each edge of the graph is given by non-trivial intersections between these clusters.

Each of these methods can be adapted to output a weighted graph. For example, the much used *UMAP* [54] dimensionality reduction algorithm builds a weighted k NN graph.

1.2.2 Feature Selection

Given a dataset $X \subset \mathbb{R}^n$, suppose we construct a graph G with vertices being the points of X . Now, the original coordinates of X are graph signals $f_i : V \rightarrow \mathbb{R}$ for $i = 1, \dots, n$. We wish to understand which features f_i are most coherent with the graph, and thus more likely explain the structure of G rather than being due to noise. As we have seen, the Rayleigh quotient of a graph signal f measures the variation of f with respect to the graph. The Laplacian score, as introduced in the paper [39], is based off this principle, with further normalisation.

Substituting the normalised Laplacian into the Rayleigh quotient definition we have we have the following equation:

$$R_{\mathcal{L}}(f) = \frac{\langle f, \mathcal{L}f \rangle}{\langle f, f \rangle} = \frac{\sum_{u \sim v} \left(\frac{1}{\sqrt{d_u}} f(u) - \frac{1}{\sqrt{d_v}} f(v) \right)^2}{\sum_u f(u)^2}.$$

However, as the constant function $\mathbf{1}$ will not necessarily be in the kernel of $\mathcal{L}$,

we will want to first normalise [14] signals by $f \mapsto D^{1/2}f$. Then

$$R_{\mathcal{L}}(D^{1/2}f) = \frac{\langle D^{1/2}f, \mathcal{L}D^{1/2}f \rangle}{\langle D^{1/2}f, D^{1/2}f \rangle} = \frac{\langle f, Lf \rangle}{\langle f, Df \rangle} = \frac{\sum_{u \sim v} (f(u) - f(v))^2}{\sum_u f(u)^2 d_u}.$$

Then the graph mean $\mu_G(f)$ of a signal f is defined as

$$\mu_G(f) = \frac{1}{\sum_{v \in V} d_v} \sum_{v \in V} d_v f(v). \quad (1.3)$$

Definition 1.2.1. If we re-centre the graph signal f by setting $\tilde{f} = f - \mu_G(f)$, then the *Laplacian score* of f , as defined in [39], is defined as

$$LS(g) = R_{\mathcal{L}}(D^{1/2}\tilde{f}) = \frac{\sum_{u \sim v} (f(u) - f(v))^2}{\sum_{v \in V} (d_v (f(v) - \mu_G(f))^2)}. \quad (1.4)$$

The Laplacian score has been further generalised to functions on simplicial complexes in [32].

Both the Rayleigh quotient and the Laplacian score measure consistency of the graph signal with the underlying graph structure: for a graph signal f , if the value of f varies wildly across the graph, thus more likely corresponding to a nuisance or noise signal, its Laplacian score will be higher – and vice versa for a lower score.

However the Rayleigh quotient and Laplacian score give one-dimensional rankings of features. Consider the two graph signals in Figure 1.1, with the node colour representing the value of the signal. The signal on the left slowly changes with respect to the graph, so we would expect it to have a low Laplacian score. On the other hand, the right signal is highly locally consistent, except at the boundary of its support – we would expect its Laplacian score to be higher than the left signal. However, when it comes to the question ‘which signal is relevant, given the graph’ arguments could be made for both signals being interesting. The problem is that they both exhibit consistency on the graph, but at different scales.

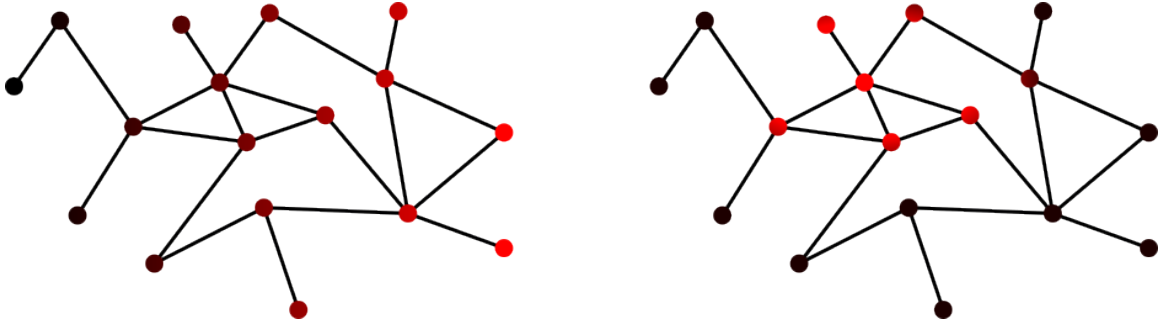


Figure 1.1: Two graph signals

In this chapter, we present three extensions of the Rayleigh quotient and Laplacian score that provide a *multiranking* of graph signals, giving different rankings based on different notions of scale within the graph. Briefly, these methods are as follows:

Eigenscores The eigenscore method decompose the Laplacian score in terms of eigenvectors of the Laplacian, giving a ranking of graph signals for each i th eigenvector.

Multiscale Laplacian Score (MLS) The multiscale Laplacian score uses the continuous-time solution of a random walk on a graph to give a ranking of graph signals for each time $t \geq 0$.

Persistent Rayleigh Quotient (PRQ) The persistent Rayleigh quotient uses a different notion of scale – along a *filtration* of the graph, a partial ordering of vertices, by considering the coherence of a graph signal with respect to the change in topology of the graph along the filtration.

1.3 Eigenscores

The Graph Fourier Transform

The Fourier transform is a fundamental operation in signal processing. One can decompose a time-dependent function into its frequencies, then perform operations in the frequency domain, such as function multiplication or ‘low pass filtering’, which become convolution and ‘denoising’ respectively, under the inverse Fourier transform. The Fourier transform works by taking the inner product with the function and an orthonormal basis of periodic functions. This basis consists of the eigenfunctions (with positive eigenvalue) of the Laplace operator $-\Delta$ on $\mathbb{R}$ – the Fourier transform can thus be extended to non-time domains by using an eigenbasis of the corresponding Laplace operator. These ideas have been long since been extended to discrete domains – for example, the compression method *JPEG* for images uses the discrete cosine transform, a variant of the discrete Fourier transform, on a pixel grid.

Graphs, however, are more irregular than grids. It is not obvious how to define a periodic function for graphs. Thus, we should define the graph Fourier transform directly via the eigenvectors of the graph Laplacian.

Definition 1.3.1 (See e.g. [65, 76]). Suppose $G = (V, E)$ is a graph, $f : V \rightarrow \mathbb{R}$ a graph signal. Then the *graph Fourier transform* of f at λ_i is

$$\hat{f}(\lambda_i) = \langle f, e_i \rangle$$

where λ_i is an eigenvalue with eigenvector e_i of the graph Laplacian L for $i = 0, \dots, n - 1$. If we use the eigendecomposition of the normalised Laplacian $\mathcal{L}$, we will call $\hat{f}$ the normalised graph Fourier transform. Given a real number $\hat{f}(\lambda_i)$ for each $i = 0, \dots, n - 1$, we can define the *inverse graph Fourier transform* of $\hat{f}$ as the

graph signal

$$f := \sum_{i=0}^{n-1} \hat{f}(\lambda_i) e_i.$$

Suppose we have a graph signal $f : V \rightarrow \mathbb{R}$, and let $g = D^{1/2}f$ be its degree-normalisation. We can write g as $g = \sum_{i=0}^{n-1} \hat{g}(\lambda_i) e_i$ and applying the Rayleigh quotient to g gives us

$$\begin{aligned} R_{\mathcal{L}}(g) &= \frac{\langle \sum_i \hat{g}(\lambda_i) e_i, \sum_j \lambda_j \hat{g}(\lambda_j) e_j \rangle}{\langle \sum_i \hat{g}(\lambda_i) e_i, \sum_j \hat{g}(\lambda_j) e_j \rangle} \\ &= \frac{\sum_i \lambda_i \hat{g}(\lambda_i)^2}{\sum_i \hat{g}(\lambda_i)^2} = \sum_{i=0}^{n-1} \lambda_i \left(\frac{\hat{g}(\lambda_i)}{\|g\|} \right)^2. \end{aligned}$$

This motivates the definition of the eigenscore:

Definition 1.3.2. Given a graph signal $f : G \rightarrow \mathbb{R}$ we define the i th eigenscore eig_i by

$$\text{eig}_i(f) = \frac{\langle D^{1/2}f, e_i \rangle}{\|D^{1/2}f\|}.$$

where e_i is the i th eigenvector of the Laplacian $\mathcal{L}$ for $i = 0, \dots, n-1$.

Now

$$R_{\mathcal{L}}(D^{1/2}f) = \sum_{i=0}^{n-1} \lambda_i \text{eig}_i(f)^2$$

We can now consider individual contributions to the Rayleigh quotient separately. Generically for a connected graph, the Laplacian will have non-degenerate eigenspaces, thus each λ_i will be distinct. Note however, that while the Rayleigh quotient is invariant to the choice of either $+e_i$ or $-e_i$, the i th eigenscore will not. We shall therefore consider both $+e_i$ and $-e_i$.

The ordering of the eigenvalues by magnitude explains the multiscale nature of the eigenscore. Expressing a graph signal in terms of Laplacian eigenvector contributions can be viewed as expanding in a frequency basis. Here, ordering the eigenvectors according to increasing eigenvalue corresponds to considering waves

of increasing frequency.

In general, computing the full eigendecomposition is expensive; however, algorithms exist for computing the first few dominant eigenvectors of a symmetric sparse matrix [11].

If G is connected, the signal $D^{1/2}\mathbf{1}$ is the eigenvector with eigenvalue 0 of the normalised Laplacian. Then

$$\text{eig}_0(g) = \frac{\langle D^{1/2}g, D^{1/2}\mathbf{1} \rangle}{\|D^{1/2}g\| \|D^{1/2}\mathbf{1}\|} = \|D^{1/2}\mathbf{1}\| \mu_G \left(\frac{g}{\|D^{1/2}g\|} \right),$$

where μ_G is the graph mean, as defined in Equation (1.3).

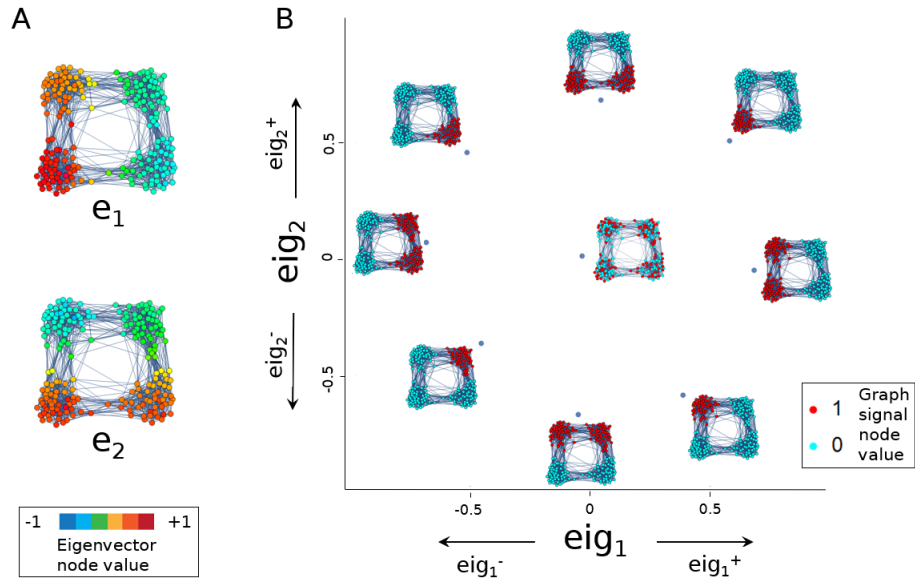


Figure 1.2: The eigenscore method demonstrated on a graph. The dataset X consists of 100 points drawn randomly from four tangential balls in $\mathbb{R}^{30}$. The graph G is constructed as the 15NN graph of X , visualised in 2D using a force-directed layout. (A) Laplacian eigenvectors e_1 and e_2 distinguish the left and right two clusters and the top and bottom two clusters, respectively. (B) Different graph signals align or anti-align differently with the two eigenvectors, resulting in a plot of eigenscore $(\text{eig}_1, \text{eig}_2)$ -space that differentiates the various signals. A random signal plots near the origin. Figure from [40]

Eigenscores to Visualise Graph Signals

Projecting graph signals onto the eigenspace spanned by low-frequency eigenscores allows us to visualise the featurespace and identify meaningful signals (see Figure 1.2). Noisy signals are mapped close to zero, and interesting signals lie on the periphery in such an eigenspace plot. Constructing such an embedding using Laplacian eigenvectors is reminiscent of Laplacian eigenmaps [6]. However, in [6], Laplacian eigenvectors are used to construct an embedding of the nodes of the graph, whereas we embed signals on the graph.

1.4 Multiscale Laplacian Score

Let $G = (V, E)$ be a graph. A *random walk* on a connected graph G consists of a walker starting at a vertex $v \in V$ and randomly moving to neighbouring vertex $u \sim v$ of v with probability d_v^{-1} . If the graph is weighted, the probability of transitioning to u will be proportional to $w(\{u, v\})$. For $t \geq 0$, we will define a function $p_t : V \rightarrow [0, 1]$ such that $\sum_{v \in V} p_t(v) = 1$ where $p_t(v)$ represents the probability of finding the walker at vertex v at time t . This is described³ by

$$p_{t+1} = D^{-1} A p_t.$$

This can be rewritten

$$p_{t+1} - p_t = (D^{-1} A - \text{Id}) p_t = -D^{-1} L p_t \quad (1.5)$$

The matrix $D^{-1} L$ is known as the *random walk Laplacian* L_{rw} . If we want to allow continuous time t , the following differential equation can be derived from Equa-

³In the random walk literature it is more common to consider the action of $A D^{-1}$ acting on p_t on the right.

tion 1.5:

$$\frac{dp}{dt} = -L_{\text{rw}}p \quad (1.6)$$

whose solution is $e^{-tL_{\text{rw}}}p_0$. As $\mathbf{1}$ is an eigenvector with eigenvalue 0, if p_0 is a probability distribution, then $e^{-tL_{\text{rw}}}p_0$ will also be a probability distribution for all $t \geq 0$. We then define the continuous-time random walk matrix as $P(t) = \exp(-tL_{\text{rw}})$. In fact, if we replace L_{rw} with the Laplacian L in Equation 1.6, the term $\exp(-tL)$ is known as the *heat kernel*.

As the random walk Laplacian L_{rw} is similar to the symmetrically normalised Laplacian $\mathcal{L}$, L_{rw} also has real non-negative eigenvalues. From this one can see that in the limit $t \rightarrow \infty$, $e^{-tL_{\text{rw}}}p_0$ is the projection of p_0 onto the kernel of L_{rw} . For a connected graph, the kernel of L_{rw} is spanned by the vector $\mathbf{1}$ of ones, thus the limit distribution of the random walk is the uniform distribution.

Plugging in the continuous-time random walk matrix gives us the following definition.

Definition 1.4.1. Let $G = (V, E)$ be a graph, and $f : V \rightarrow \mathbb{R}$ a graph signal. For $t \geq 0$, The *multiscale Laplacian score* of f at resolution t is defined as

$$MLS(f, t) = \frac{\langle D^{1/2}\tilde{f}, D^{1/2}(I - P(t))\tilde{f} \rangle}{\langle D^{1/2}\tilde{f}, D^{1/2}\tilde{f} \rangle} = \frac{\sum_{u \sim v} d_u \exp(-tL_{\text{rw}})_{uv} (f(u) - f(v))^2}{\sum_{v \in V} d_v \tilde{f}^2},$$

where we use the identity $d_u P(t)_{uv} = d_v P(t)_{vu}$ for all $t \in \mathbb{R}_{\geq 0}$ and all $u, v \in V$.

As we have a score of a signal for each time t , we need to decide which values of t to use in the analysis. Continuous time random walks with the transition matrix $P(t)$, are used in community detection to find natural partitions of graphs (see [40] for further details). For each time t , many different reasonable partitions may be found. If for a specific t_0 these partitions are very similar, this indicates that the graph has a stable community structure at time t_0 . An appropriate notion of similarity between partitions [55, 72] is given by *variation of information*:

Definition 1.4.2. Let $\{C_i\}_{1 \leq i \leq k}$ and $\{C'_j\}_{1 \leq j \leq k'}$ be two partitions of a graph with N nodes. Then, their variation of information (VI) is defined to be

$$\text{VI}(\{C_i\}, \{C'_j\}) = 2H(\{C_i\}, \{C'_j\}) - H(\{C_i\}) - H(\{C'_j\}),$$

where

$$H(\{C_i\}, \{C'_j\}) = - \sum_{i,j} \frac{|C_i \cap C'_j|}{N} \log_2 \left(\frac{|C_i \cap C'_j|}{N} \right)$$

is the joint entropy of the two partions and

$$H(\{C_i\}) = - \sum_{i=1}^k \frac{|C_i|}{N} \log_2 \left(\frac{|C_i|}{N} \right).$$

is the entropy of each partion.

To find key times t where the graph has a stable community structure, we look for t that are local minima of the VI across t , as illustrated in Figure 1.3.

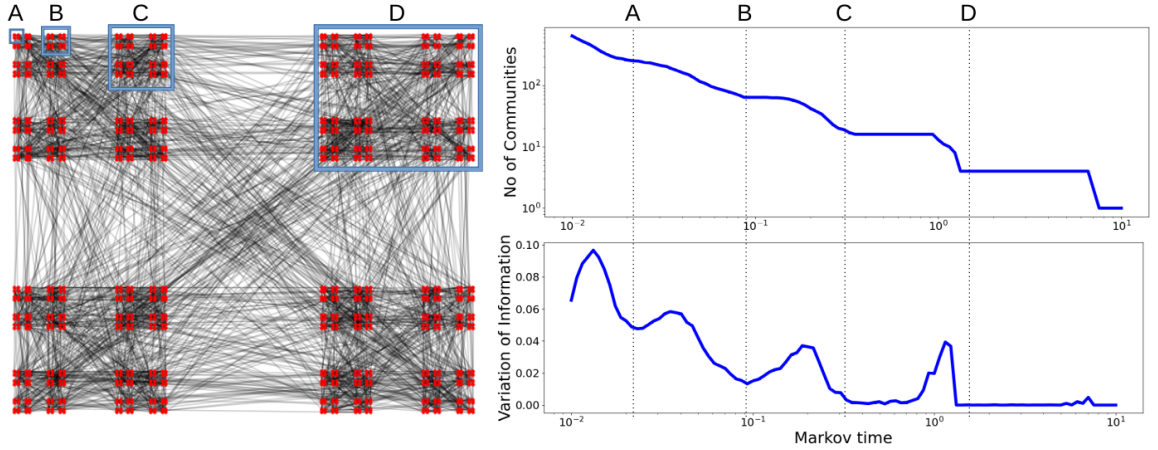
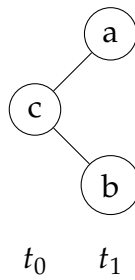


Figure 1.3: The graph on the **left** displays community structures at four different scales, exemplified by the groups A, B, C and D. When computing the mean pairwise variation of information (**right**) as a function of scale (Markov time), we find local minima corresponding to resolutions A (256 communities), B (64 communities), C (16 communities) and D (4 communities). Figure inspired by [4].

1.5 Persistent Rayleigh Quotient

For a graph G we may have additional information associated to each node of G that we would like to use to further inform our analysis. In particular, we are interested in cases where the data comes from a process or experiment where each vertex appears at a certain time, and we associate this time data to the vertex. The toy model we consider in this section is the following graph:



Here, we have a state c that appears at time t_0 and splits into two states a and b at time t_1 . The biological context, as will be described in Chapter 2, is that c represents a parent stem-cell like cell type which later bifurcates into two possible daughter cell types.

While the Rayleigh quotient can rank signals by their coherence with the underlying graph, it is ignorant to any time data we attach to the vertices. Consider the two graph signals below:

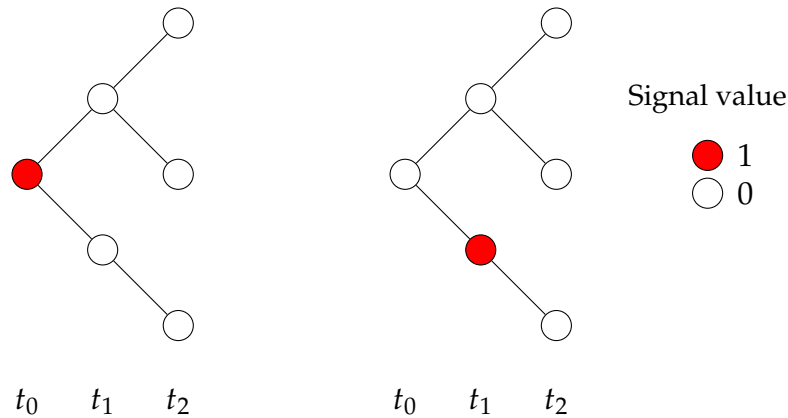


Figure 1.4

The Dirichlet energy of both graph signals is 2, hence their Rayleigh quotients are the same. However, from the point of view of the time data, the signals play a different role, thus we somehow need to introduce the time information into our analysis. Let us first introduce some notation:

For a graph G , for a subset $\alpha \subset V$ define $G[\alpha]$ to be the subgraph of G with vertices α and all edges in E that have both endpoints in α , let $L|_\alpha$ be the Laplacian for the graph $G[\alpha]$. For a function $\mathbf{t} : V \rightarrow \mathbb{R}$ and a real number $t_i \in \mathbb{R}$ define

$$\{\mathbf{t} \leq t_i\} = \{v \in V : \mathbf{t}(v) \leq t_i\}.$$

and similarly for $\{\mathbf{t} = t_i\}$ and $\{\mathbf{t} \geq t_i\}$.

We might attempt to solve the problem by studying how the signals change when we restrict the graph to only the vertices at time t . If we apply this to Figure 1.4 we would look at the graphs $G[\{\mathbf{t} = t_i\}]$ for $i = 0, 1, 2$. However, this does not give any information: Both $G[\{\mathbf{t} = 0\}]$ and $G[\{\mathbf{t} = 1\}]$ have no edges, thus the Rayleigh quotient is zero for both signals. Instead, we should look at a *filtration* of subgraphs of G : either the sublevel sets

$$G[\{\mathbf{t} \leq t_0\}] \leq \cdots \leq G[\{\mathbf{t} \leq t_k\}] \leq G$$

or the superlevel sets

$$G[\{\mathbf{t} \geq t_k\}] \leq \cdots \leq G[\{\mathbf{t} \geq t_0\}]$$

for a sequence $(t_i)_{i=0}^k$. Now considering the Rayleigh quotient across the sub/super-level set filtration will distinguish the signals in Figure 1.4. Consider the next example in Figure 1.5.

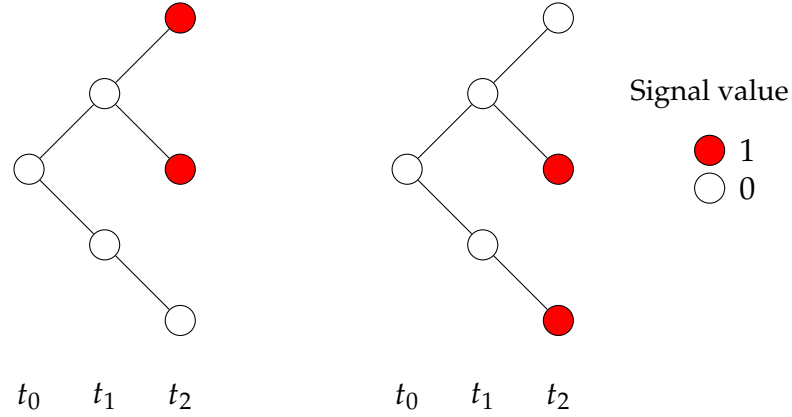


Figure 1.5

Using the Rayleigh quotient on the superlevel set filtration (or sublevel) will not be able to discern these signals, but the two red vertices on the left signal have a common ancestor vertex one timestep back whilst on the right signal only have a common ancestor two timesteps back, therefore we should construct a score that considers the left graph signal more coherent with the graph than the right signal.

1.5.1 Kron Reduction

Dirichlet and Neumann Boundary problems

The graph Laplacian allows us to define an analogue of the Laplace equation

$$\Delta f = 0. \quad (1.7)$$

A function f satisfying Equation 1.7 is known as *harmonic*. In Section 1.4, we supplied an initial condition solve the equation for graphs. The Laplace equation is usually solved on a bounded domain $D \subset \mathbb{R}^n$ by supplying a boundary condition on the domain, e.g. a Dirichlet boundary condition, where we fix $f|_{\partial D}$, or a Neumann boundary condition, where we fix $\langle \mathbf{n}(x), \nabla f(x) \rangle$ for the unit normal $\mathbf{n}(x)$ for $x \in \partial D$.

For $\emptyset \neq \alpha \subset V$, define α^c as $V \setminus \alpha$. We shall model D as the set α^c and its boundary⁴ as the graph $G[\alpha]$. The Dirichlet problem is then solving $(Lf)_v = 0$ for $v \in \alpha^c$ with the boundary condition $f|_\alpha = g$ for some fixed $g : \alpha \rightarrow \mathbb{R}$. If such an f exists and is unique, it is termed the *harmonic extension* of g , – the graph signal that extends g and is harmonic at vertices in α^c . The solution f can be thought of as the stationary state of fixing values g on α and allowing these to diffuse throughout α^c .

The solution to the Dirichlet problem for graphs exists, and is unique [96]. First, for subsets $\alpha, \beta \subseteq V$ define $L[\alpha, \beta]$ as the submatrix of L with rows indexed by α and columns indexed by β . Under an appropriate reordering of the vertex labels, the Laplacian L has block form

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

Now the Dirichlet problem can be written as

$$\begin{bmatrix} L[\alpha, \alpha] & L[\alpha, \alpha^c] \\ L[\alpha^c, \alpha] & L[\alpha^c, \alpha^c] \end{bmatrix} \begin{bmatrix} g \\ f|_{\alpha^c} \end{bmatrix} = \begin{bmatrix} \star \\ \mathbf{0} \end{bmatrix} \quad (1.8)$$

hence we must have

$$L[\alpha^c, \alpha]g + L[\alpha^c, \alpha^c]f|_{\alpha^c} = \mathbf{0}$$

If no vertices in α^c are disconnected from α , then the matrix $L[\alpha^c, \alpha^c]$ is weak-chained diagonally dominant (see [75]), and thus is non-singular. Hence

$$\begin{aligned} f|_\alpha &= g, \\ f|_{\alpha^c} &= -L[\alpha^c, \alpha^c]^{-1}L[\alpha^c, \alpha]g, \end{aligned}$$

and thus the solution f exists and is unique. What about the value $\star$ in Equ-

⁴Usually in graph theory, the boundary of α^c is the set of vertices not in α^c but neighbouring α^c . Here, the reader should think of $G[\alpha]$ representing a ‘thickened’ boundary of α^c .

tion 1.8? This is given by

$$\star = L[\alpha, \alpha]g + L[\alpha, \alpha^c]f|_{\alpha^c}. \quad (1.9)$$

Substituting in the solution to $f|_{\alpha^c}$ we now have

$$\star = \left(L[\alpha, \alpha] - L[\alpha, \alpha^c]L[\alpha^c, \alpha^c]^{-1}L[\alpha^c, \alpha] \right) g$$

The matrix $L[\alpha, \alpha] - L[\alpha, \alpha^c]L[\alpha^c, \alpha^c]^{-1}L[\alpha^c, \alpha]$ is known as the *Schur complement* $L/L[\alpha^c, \alpha^c]$.

The Laplacian equation can be also solved by providing a Neumann boundary condition. However, there is not an immediate notion for the normal vector $\mathbf{n}(v)$ at a vertex v in a boundary of a graph. Arguably, the analogous object for a set $\alpha \subset V$ is the submatrix B_α of the boundary matrix B , consisting of the α columns of B . Recall that the Neumann condition is given by fixing the value of $\langle \mathbf{n}(x), \nabla f(x) \rangle$ for $x \in \partial D$. Then in the case of graphs, we must fix the value $B_\alpha^T (Bf)$. As it turns out, in the Dirichlet problem above, this is exactly given in Equation 1.9:

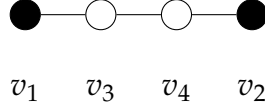
$$L[\alpha, \alpha]g + L[\alpha, \alpha^c]f|_{\alpha^c} = B_\alpha^T (Bf).$$

Therefore, the solution of the Dirichlet problem defined by a signal $g : \alpha \rightarrow \mathbb{R}$ on the boundary $G[\alpha]$ is also the solution (up to a constant) of the Neumann problem defined by $(L/L[\alpha^c, \alpha^c])g$. The Schur complement $(L/L[\alpha^c, \alpha^c])$ for graphs is sometimes referred to as the *Dirichlet-to-Neumann operator* [19].

Kron Reduction

Returning to the problem at the beginning of Section 1.5, the Dirichlet-to-Neumann operator will allow us to measure how coherent a signal defined on a subgraph

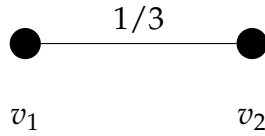
is with respect to the entire graph; we fix the signal on the subgraph, allow this to diffuse through the remainder of the graph, and then measure the flux at the subgraph needed to maintain this signal. Consider the following example:



where α consists of vertices v_1 and v_2 . If we set $g(v_1) = 1$ and $g(v_2) = 0$ then $f|_{\alpha^c}(v_3) = 2/3$ and $f|_{\alpha^c}(v_4) = 1/3$. The matrix $L/L[\alpha^c, \alpha^c]$ is given by

$$L/L[\alpha^c, \alpha^c] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \left(-\frac{1}{3} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \right) \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 1/3 & -1/3 \\ -1/3 & 1/3 \end{bmatrix}$$

hence the Neumann condition is $1/3$ at v_1 and $-1/3$ at v_2 . Notice, however, that $L/L[\alpha^c, \alpha^c]$ is also the Laplacian of the weighted graph



where the weight between v_1 and v_2 is the reciprocal of the distance between v_1 and v_2 in the original graph.

This phenomenon is not isolated – in fact, the class of Laplacians of (weighted) graphs is closed under the Schur complement:

Theorem 1.5.1 ([19, 3]). *Suppose $G = (V, E)$ is a graph, and $\emptyset \neq \alpha \subsetneq V$ a subset of nodes. If α^c is not disjoint from G , in the sense that there exists a path from each $v \in \alpha^c$ to a vertex in α , then there exists a weighted graph G_α with vertices α whose Laplacian equals $L[\alpha, \alpha]/L[\alpha^c, \alpha^c]$.*

Given such a G and α as in Theorem 1.5.1, we will call the graph G_α the *Kron reduction of G to α* . The structure of G_α can be deduced from G and α without constructing the Laplacian. For any two vertices $u, v \in \alpha$, there exists an edge

between u and v in G_α if and only if there exists a path from u to v in G that only passes through vertices in $\alpha^c \cup \{u, v\}$. To find the weight of this edge, let us first introduce effective resistance: for two vertices u, v of a weighted graph G , the effective resistance $\mathcal{R}_{u,v}^G$ is given by

$$\mathcal{R}_{u,v}^G = (\mathbb{1}_u - \mathbb{1}_v)L^+(\mathbb{1}_u - \mathbb{1}_v)$$

where $\mathbb{1}_u$ and $\mathbb{1}_v$ are the functions 1 on u and v respectively, and zero elsewhere, and where L^+ is the Moore-Penrose inverse of the Laplacian L . Physically, the weighted graph G represents a circuit where each edge $\{u, v\}$ corresponds to a resistor with resistance $1/w(u, v)$. Then the effective resistance is the resistance found when putting a current with intensity 1 entering u and leaving v with intensity 1. The Kron reduction of a graph preserves the effective resistance [19]:

$$R_{u,v}^{G_\alpha} = (\mathbb{1}_u - \mathbb{1}_v)(L[\alpha, \alpha]/L[\alpha^c, \alpha^c])^+(\mathbb{1}_u - \mathbb{1}_v) = (\mathbb{1}_u - \mathbb{1}_v)L^+(\mathbb{1}_u - \mathbb{1}_v) = R_{u,v}^G.$$

The Persistent Laplacian

Topological data analysis studies the evolution of topological invariants⁵ across filtered topological spaces, such as filtered graphs

$$G[\alpha_1] \leq \dots \leq G[\alpha_k]$$

for an increasing sequence $\alpha_1 \subset \dots \subset \alpha_k \subset V$. The most common tool is persistent homology [28], which computes how invariants, such as connected components in $G[\alpha_i]$, persist in the larger graph $G[\alpha_{i+1}]$. Suppose H is a subgraph of G with vertices α . Then each connected component H_i of H can be identified by some vertex $v_i \in H_i$ for $i = 1, \dots, \beta_0(H)$ where $\beta_0(H)$ is the number of connected components

⁵This will be discussed further in Chapter 6.

of H . Fix some ordering of the connected components H_i . Then for $i \geq 1$, we shall say the connected component H_i *persists* in G if, considered as a subgraph of G , H_i is disconnected from H_1 . The number of connected components of H that persist in G is termed the (degree 0) *persistent Betti number* $\beta_0(H \leq G)$. If G has no components disconnected from H , we can compute the Kron reduced graph G_α from Theorem 1.5.1. Recall that two vertices in G_α are connected if they are connected in G , thus the number of connected components of G_α , is $\beta_0(H \leq G)$.

The notion of Kron reduction of a graph has been extended to simplicial complexes in [56, 95] where it is known as the persistent Laplacian. The persistent Laplacian extends Kron reduction in the sense that nullity of the degree k persistent Laplacian of simplicial complexes $K \leq L$ is the persistent Betti number of K with respect to L in degree k .

1.5.2 Persistent Rayleigh Quotient

As noted earlier, the Schur complement $L/L[\alpha^c, \alpha^c]$ requires that $L[\alpha^c, \alpha^c]$ is invertible. If this is not the case, i.e. when there is a vertex in α^c disconnected from α , we will instead use the *generalised Schur complement*

$$L//L[\alpha^c, \alpha^c] := L[\alpha, \alpha] - L[\alpha, \alpha^c]L[\alpha^c, \alpha^c]^+L[\alpha^c, \alpha]$$

where $+$ denotes the Moore-Penrose inverse.

Let G be a graph, and $\mathfrak{t} : V \rightarrow \mathbb{R}$ a function on the nodes of the graph representing time. For $t_1 \leq t_2$, define L_{t_1, t_2} as the generalised Schur complement $L|_{\{\mathfrak{t} \leq t_2\}}$ to $G_{\{\mathfrak{t} \leq t_1\}}$:

$$L_{t_1, t_2} := L//[\{\mathfrak{t} \leq t_1\}^c, \{\mathfrak{t} \leq t_1\}^c].$$

Definition 1.5.2. For $t_1 \leq t_2$, define the *persistent Rayleigh quotient*

$$\text{PRQ}(t_1, t_2)(f) = \frac{\langle f|_{\{t \leq t_1\}}, (L_{t_1, t_2}) f|_{\{t \leq t_1\}} \rangle}{\langle f|_{\{t \leq t_1\}}, f|_{\{t \leq t_1\}} \rangle}$$

and the *normalised persistent Rayleigh quotient* as (cf. Equation 1.2.2)

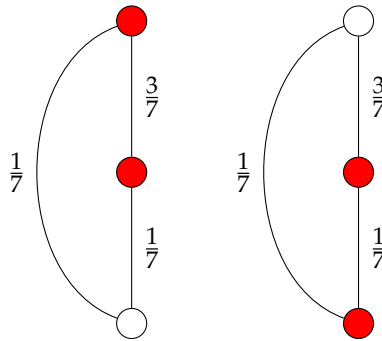
$$\widehat{\text{PRQ}}(t_1, t_2)(f) = \frac{\langle f|_{\{t \leq t_1\}}, (L_{t_1, t_2}) f|_{\{t \leq t_1\}} \rangle}{\langle f|_{\{t \leq t_1\}}, D_{t_1, t_2} f|_{\{t \leq t_1\}} \rangle}$$

where D_{t_1, t_2} is the diagonal matrix corresponding to the Laplacian L_{t_1, t_2} .

Let us return to Figure 1.5. We will study the superlevel set filtration

$$G[\{t \geq t_2\}] \leq G[\{t \geq t_1\}] \leq G[\{t \geq t_0\}].$$

In the definition of the PRQ, we use the sublevel set filtration for t . To use the superlevel set filtration, we must reverse the time direction: we instead consider $t' := t_2 - t$ and $t'_i = t_2 - t_i$ for $i = 0, 1, 2$. Consider the full graph G and Kron reduce the graph to $\alpha = \{t' \leq 0\}$. The two signals from Figure 1.5 on this reduced graph are shown below:



The Dirichlet energy of the left signal is $2/7$, and $4/7$ for the right, thus the left signal has a lower PRQ and so is more consistent than the left, as desired.

Code Availability

Code for computing the multiscale Laplacian score and the persistent Rayleigh quotient can be found at:

<https://github.com/osumray/multiscale-signal-selection-single-cell>

Chapter 2

Gene Selection

2.1 Single-Cell Transcriptomics

Proteins make up most of the dry mass of a biological cell, and perform most of the functions within the cell: from internal behaviour, such as catalysing reactions, to external behaviour, such as inter-cell signaling. The information to make a specific protein is in a particular region of DNA, a *gene*. The DNA does not produce the protein directly, but must be first be copied into messenger RNA (mRNA), a process known as *transcription*. This mRNA strand (after processing) is then *translated* into a chain of amino acids, which then folds into the resultant functional protein. This process, from DNA to RNA to protein, is known as the ‘central dogma of molecular biology’. While a simplified model, this idea gives a general principle for the flow of information in a cell.

In order to understand how cells function, biologists need to perform an *assay*, a measurement. For example, one might be interested in understanding what differentiates cell type A from cell type B e.g. which proteins exist in large quantities in cells of type A and not in type B. However, directly measuring many different proteins is technically very difficult. It is much easier to instead measure the

amount of mRNA for each gene, the *expression* of the gene.

The technology to sequence RNA (RNA-seq) with high-throughput began in the 2000s, with the introduction of *next-generation sequencing*. The RNAs are isolated from the sample, amplified, and then sequenced and aligned to a reference genome. For each gene g in the genome, we now have a count $c(g) \geq 0$ of the number of mRNA molecules measured in the sample that map to the gene g – its gene expression. There are more than 20,000 coding genes in the human genome. Performing RNA-seq on a whole sample in this way is known as *bulk sequencing*.

Although virtually all cells in the human body contain the same copies of DNA, their phenotypes vary drastically. Biologists classify cells into broad categories known as *cell types*, such as neuron cells, epithelial cells or blood cells. Within the blood cell type, there are several subtypes; T cell, B cell, macrophage, monocyte, natural killer cells and so on. Even within these subtypes, there exist further subtypes – there are helper (CD4+) T cells, cytotoxic (CD8+) T cells, T regulatory cells, and even these have subtypes. Eventually, we begin talking about specific transcriptomic *cell states*, the exact level and combination of gene expressions.

Bulk RNA sequencing mixes together all the mRNA found from a sample. A classic analogy is a fruit smoothie – the result is a complete mixture of all the constituent fruits. Though we may be able to tell the relative abundance of broad cell types, through genes whose expression is specific to a cell type, a *marker gene*, for many experiments this is not granular enough.

The ability to sequence individual cells simultaneously is a relatively recent advancement in biology and bioinformatics. There are many platforms and technologies that provide single-cell RNA-seq (scRNA-seq), one of the most popular high-throughput methods is with microfluidics. The 10X platform separates dissociated single¹ cells into droplets; these cells are then broken down (lysed) and

¹Mostly single, however some droplets contain multiple cells, known as *doublers*, and some contain no cells

each mRNA in the droplet is combined with a unique barcode, identifying that the mRNA read originated from that particular cell. These are then amplified and sequenced in a similar manner to bulk RNA-seq. The result is a scRNA-seq dataset, given as a matrix X . Each row of X corresponds to an individual cell, and each column corresponds to the expression level of a particular gene. Single cell transcriptomic technologies can sequence many thousands of cells from one sample, and by aggregating multiple samples, scRNA-seq datasets can have hundreds of thousands of rows – very much pushing bioinformatics into the age of ‘big data’.

2.2 Standard Analysis of scRNA-seq

The typical aims in scRNA-seq analysis are to:

1. classify the cells into cell types, either known cell types or *de novo* cell types, and
2. find *differentially expressed genes*: genes that have significantly different expression between these cell types.

2.2.1 Cell classification

The output of scRNA-seq first needs to be processed in various ways, e.g. filtering low-quality cells and removing doublets. Subsequently, the data needs to be normalised, for which we highlight two standard methods:

Log-normalisation Log normalisation is the simplest method based on a common principle in bioinformatics, that fold-change is more meaningful than difference when comparing, e.g. number of mRNA reads. A common normalisation is then

$$X_{i,j} \mapsto \log\left(1 + s \frac{X_{i,j}}{\sum_k X_{i,k}}\right)$$

for some chosen scaling factor s , usually set to 10^5 . The value $\sum_k X_{i,k}$ is the *sequencing depth* for the i th cell, the number of unique mRNA transcripts found for that cell. The sequencing depth is expected to be on the order of 10^3 – 10^5 , hence the term $s \frac{X_{i,j}}{\sum_k X_{i,k}}$ normalised the number of reads for each cell and gene to an expected sequencing depth.

Variance-Stabilising Transform A more sophisticated method, introduced to single-cell analysis in the paper [36], that accounts various technical confounders in scRNA-seq, such as the linear relationship between sequencing depth of a cell and its genes' expressions, by transforming the data to residuals of a negative binomial regression model.

In order to find cell types, the data is then clustered. There are many algorithms for clustering, however, in the most popular packages for single-cell analysis (Seurat [38] and Scanpy [98], clustering is not performed on the data directly. Instead, an intermediary representation of the data is built, a network, which is then clustered. This is usually done with a kNN graph and then applying the Louvain or Leiden community detection methods [87].

2.2.2 Differential Gene Expression

Given a clustering of the data X into cell types, the next task is to discover which genes have significantly different expression between cell types. Suppose we have two clusters (cell types) C_i and C_j . Now we perform a statistical test for the gene expressions between C_i and C_j – common methods are the Wilcoxon signed rank test (used by Seurat [38]), or the Wald test under a negative binomial distribution, as used by DESeq2 [52]. The p -values, after correction to account for multiple testing, rank the genes by relevance, with lower being more significant.

2.2.3 Limitations of Current Analyses

2.2.3.1 Models of Cell Type

While all² cells in an organism have the same genome, they can exhibit wildly different morphologies and phenotypes. A cell type is a group of cells (or rather cell states) that have a similar phenotype. This definition is not exact – while it is straightforward to distinguish a neuron from a red blood cell, it is more difficult to distinguish between e.g. helper T cells and cytotoxic T cells or T regulatory cells which are usually defined in terms of specific surface markers (CD4 and CD8 respectively). Furthermore, the type of a particular cell is not necessarily static; while some cells are fully differentiated, i.e. they will not change into a more specific cell type, others cells are multipotent – they may differentiate further. For example, the common myeloid progenitor cell differentiate into various cell types, including myeloblasts which further differentiate into neutrophils etc. Some cells also exhibit *plasticity* such as cancer cells or T cells which can go from a naïve to eventually exhausted cell state.

Cell types are usually modelled as a discrete set, rather than a continuous spectrum, as exemplified by the use of clustering algorithms. Any clustering algorithm requires some choice of scale, different choices of which will affect downstream analysis. Moreover, using clusters will bias analysis to the average of those clusters and could miss ‘outlier’ cells that are transitioning from one cell type to another. Consider performing differential gene expression on the following toy example:

²Barring gamete cells.

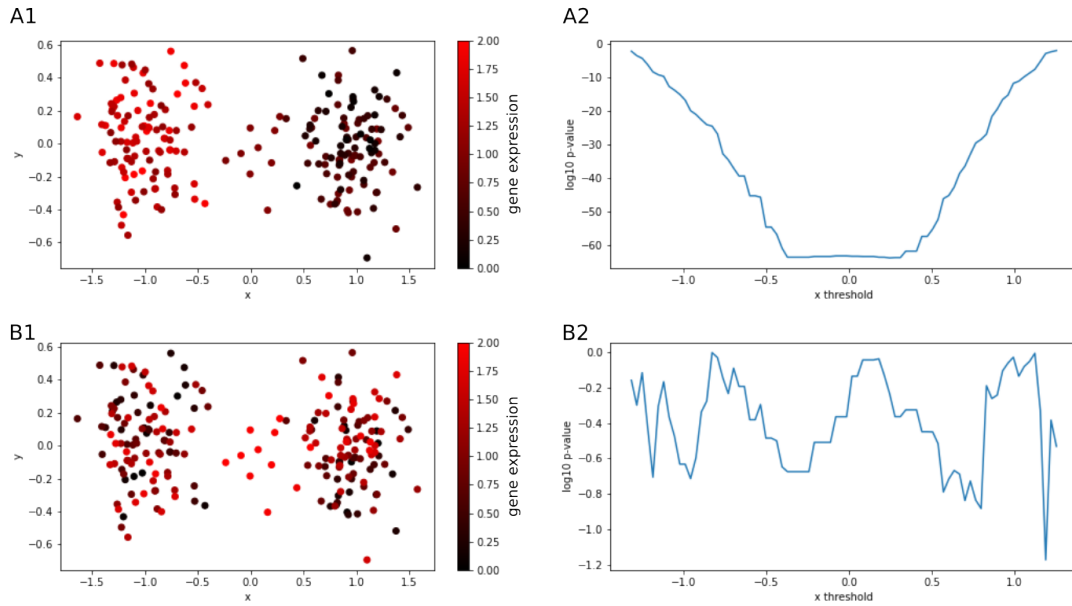


Figure 2.1: Differential gene expression depends on how the data is clustered. Subfigures **A1** and **B1** plot two gene expression patterns on a toy dataset consisting of two large clusters joined by a smaller cluster. In Subfigures **A2** and **B2** display the p -value associated to performing a t -test on between gene expressions on the left and right sides of the x -threshold. For each value of x , the data is split into two populations, left and right, based on the x coordinate. For the gene expression pattern in **A1**, the p -values vary gradually however the exact ranking of the gene will depend on the x chosen. On the other hand, the gene **B1** is only highly and consistently expressed in the smaller transitionary cluster, and randomly expressed in the left and right clusters. Dividing the data into two populations does not give a significant p -value for any choice of x . Furthermore, the p -value, and hence the ranking of the gene, changes dramatically across different x -thresholds.

Figures 2.1A1 and B1 show two different gene expression patterns on a toy dataset consisting of two large groups of cells, joined by a smaller group of transitionary cells. The expression in plot A1 has a gradual change in expression from the left cluster to the right. Plots A2 and B2 display the p -value resulting from performing a t -test on the signals A1 and B1 respectively, where the two groups are decided by splitting on some x value. For the signal in A1, we see the p -value is highly significant for most values of x . Naturally, the p -value changes for different x , and so the final ranking of that gene will depend on the choice of clustering. The signal in A2 is only constant for the transitionary cells, with the expression values

random by construction for the remaining cells. No value of x gives a significant p -value for this signal, showing that underdetermining the number of clusters will mean that certain interesting genes, such as those driving the transition between different groups of cells, will not be ranked as interesting by differential gene expression algorithms.

A solution we propose in this chapter is to build a model of the space of cell states with a graph (network) rather than disjoint clusters, and to rank genes by using the spectral graph signal analysis techniques we introduced in Chapter 1. For the Laplacian score, such analysis has been applied in [32].

2.3 Eigenscores and MLS Applied to PBMC and T Cell Data

2.3.1 Datasets and Preprocessing

We will apply the eigenscores and MLS methods described in Chapter 1 to two scRNA-seq datasets. The first consists of 2700 human peripheral blood mononuclear cells (PBMC) [27], a standard dataset used in e.g. [70, 35]. The broad cell types found in the PBMC data set are lymphocytes (T cells, NK cells, B cells), monocytes, and dendritic cells. This data set also contains a small group of *PPBP*+ cells, which have been variously identified as platelets or their progenitors, megakaryocytes [70, 35, 97]. For simplicity, we will refer to them as platelets in our discussion, following [70].

The second dataset is consist of 24,911 human T cells infiltrating lung tumours and adjacent normal tissue [46] (previously analysed with the Laplacian score [32]). While the majority of the cells are CD4+ and CD8+ T cells, some NK cell clusters are also present in this data set. The aim of the T cell experimental data set was to

determine how the cellular composition of the stromal region surrounding lung tumours differs from the stromal region of healthy lung tissue. The authors analysed scRNA-seq data to identify different stromal cell compositions (or signatures) that correlate with survival.

Preprocessing for PBMC and T Cell Datasets

The PBMC and T cell scRNA data sets were normalised using the variance stabilizing transform (VST) [36]. Before constructing the graph, the dimensionality of the datasets was reduced using the standard practice³ of first restricting to the most variable genes (3000 in this case) and then using the first 30 principal components. Due to the large size of the T cell dataset, 3000 cells were sampled at random to form the final preprocessed dataset, (following [32]).

We then constructed a 15NN graph from the dimensionality-reduced datasets, using two notions of distance: cosine-dissimilarity for the PBMC data (following the Seurat tutorials) and Pearson correlation-distance for the T cell data (following [32]). Both graphs were weighted according to the weights given by the dimension reduction algorithm UMAP [54].

Previous Results on PBMC Data

In the UMAP plot generated from the variance stabilised PBMC data (see Figure 2.8), five large populations (clusters) and two smaller ones are visible. By colouring the UMAP plot by marker gene expression, the VST vignette in [35] identifies that the large populations roughly correspond to CD4 T cells, CD8 T cells, NK cells, B cells and monocytes, while the small components contain platelets and dendritic cells. The clustering algorithm applied to the data further subdivides these populations. By performing differential gene expression (DGE) analysis by

³As in e.g. Seurat.

use of a non-parametric Wilcoxon rank sum test [69], the VST vignette suggests that the CD4 and CD8 T cells can be split into three subpopulations. The B cells and NK cells are decomposed into two subpopulations, each closely linked to a high expression of known marker genes. The top ten differentially expressed genes in the twelve resulting clusters are given in Table 2.1.

2.3.1.1 Previous Results on T Cell Data

Lambrechts et al. [46] identify several subclusters of T and natural killer (NK) cells based on clustering in t-SNE (nine clusters) and marker genes (six subgroups, see Figure 2.2). There is only one visibly connected component in the original t-SNE plot, while our UMAP plot shows more structure. Govek et al. [32] applied the combinatorial Laplacian score to identify a variety of genes, including *HAVCR2*, *RSAD2* and *GZMK*, that are consistent with the topological structure of the data set but do not correspond to the clusters defined in [46]. Govek et al. also demonstrated on the T cell data that the discriminating power of the combinatorial Laplacian score (measured by the area under the receiver-operating characteristic curve) is comparable to that of conventional DGE methods and superior to feature variance without taking topology into account [32].

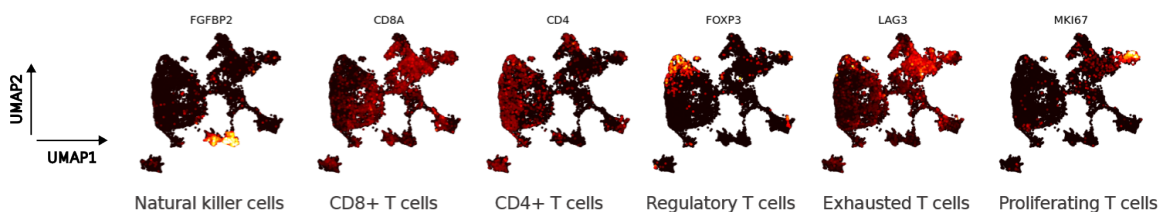


Figure 2.2: Lambrechts et al. [46] classified T cells into six sub-cell types based on marker genes. T cells plotted using 2D UMAP. To reduce overplotting and assist visualisation, points with non-zero expression were plotted on top for this figure and Figure 2.10.

2.3.2 Eigenscores

Eigenscores test whether features such as genes, viewed as functions on nodes representing cells of a graph, align or anti-align with the Laplacian eigenvectors on the graph. They can be used to rank genes similarly to DGE but on different scales in the data, according to their coherence with the topology of the cell similarity graph. They can also be applied to explore gene expression by visualising genes in a gene space. By scoring genes for alignment with individual eigenvectors, as well as selecting relevant genes, we can often shed light on the biological processes in which these genes are involved. Eigenscores are meaningful in data sets with clear community structures, e.g., the PBMC data set. Eigenscores are also meaningful in data sets that cannot be decomposed into distinct clusters but rather have a continuous structure, e.g., the T cell data set. In either case, the eigenscores do not rely on predefined clusters; they scan through the data in an unsupervised way.

2.3.2.1 The Geometry of PBMC Genes via Eigenscores

We compute eigenscores of the top genes in the PBMC data set [27] (see Figure 2.9). While many selected genes overlap with cluster marker genes identified by differential gene expression (DGE) via Seurat’s default implementation of the non-parametric Wilcoxon rank sum test [69], validating eigenscores, this method identifies 26 additional genes (see Table 2.1). To interpret the genes, we compare to 12 finer cell subtypes or seven previously determined broader cell subtypes [70] (see Figures 2.3A and 2.8). Eigenscores select genes that are enriched in broader cell types (e.g., *MALAT1* for all lymphocytes or *FTL* and *FTH1* for all monocytes). Due to their expression in multiple cell clusters, differential gene expression (DGE) does not identify these as highly ranked significant genes for one cell cluster. *MALAT1* is a highly-expressed non-coding RNA, whose enrichment may be a feature of damaged or low-quality cells in some scRNA-seq data sets [2, 66]. However, the

biological differential expression of *MALAT1* between immune cell types has also been previously reported, including increased expression in lymphocytes compared with monocytes [79]. In any case, the increased expression of *MALAT1* is certainly a feature shared by multiple lymphocyte clusters in the PBMC data set. Genes *FTL* and *FTH1* encode, respectively, the light and heavy chain of ferritin, which is a major protein involved in iron storage and homeostasis. It is highly expressed within myeloid (monocyte) cells, which can be further modulated by inflammatory stimuli [15, 85, 105].

While traditional DGE identifies markers unique to a given cluster or grouping of cells, eigenscores can also identify genes that exhibit biological variation within multiple disparate cell types. For example, the gene *FCGR3A* can distinguish between major subpopulations of monocyte cells [62, 26, 17], major subpopulations of NK cells [93, 18] and major subpopulations of T cells [62]. Eigenscore analysis ranks *FCGR3A* highly (Figure 2.9, e_4 and e_5), whereas with traditional DGE, it is not ranked in the top 10 genes for any cluster, which is perhaps due to its high level of expression across clusters in subtypes from disparate cell lineages. Moreover, *PPBP*, a highly differentially expressed marker gene on platelets, is ranked highly by eigenscores and not DGE. For completeness, we include a qualitative and quantitative comparison between eigenscore and DGE ranking by non-parametric Wilcoxon rank sum test (Figure 2.4) and present the set complement of the top scores.

We can further explore the relationships between genes by projecting the gene space of eigenscores via UMAP (Figure 2.3B). The visualisation of 16 dimensional low-frequency eigenscores (eigenscore 1–16 corresponding to $0 < \lambda_i < 0.1$) emphasises gene signals that are most coherent with the cell similarity graph structure. We interpret this visualisation of gene space in Figure 2.3B. Genes plotting in the centre (blue) have low eigenscores, where the gene expression is incoher-

ent with the graph topology (see, for example, gene *BAG4*). Genes with similar expression patterns in the data set plot together, and a sequence of continuous transitions of gene signals plots continuously in eigenscore space (as illustrated in Figure 1.2). The flares in the gene space with high eigenscores correspond to groups of genes strongly expressed on clusters corresponding to broad cell types or the cell cycle. We explore and interpret the continuous signal of gene space in Figure 2.3B with flares interpreted clockwise as groups of genes expressed on: (I) platelets, (II) different subtypes of monocytes with continuous transition of genes, (III) B cells and dendritic cells, (IV) cytotoxic lymphocytes (NK cells and CD8 T cells), (V) all lymphocytes but not myeloid cells (monocytes) (these are two separate developmental lineages), and (VI) a previously unidentified subpopulation of cells within a larger cluster expressing a high level of cell cycle markers.

Figure 2.4C gives a quantitative comparison of the amount of overlap in the eigenscore and DGE rankings, showing that there is consistent overlap while we also find additional genes. We highlight that two flares of the gene space geometry are missed by DGE via non-parametric Wilcoxon rank sum test (see Figure 2.4B), particularly region VI. This region corresponds to genes involved with cell cycle progression, which are highest in a highly proliferative subpopulation of B cells in the PBMC data set. The genes in this region show strong overlap with gene signatures used by Seurat to identify the S phase (*PCNA*, *TYMS*, *RRM2*) and G2/M phase (*TOP2A*, *BIRC5*, *UBE2C*, *HMGB2*, *SMC4*, *CKS1B*). Notably, our eigenscore method is able to identify the cell cycle as an important source of biological variation in this data set that is missed by traditional cluster-based DGE without requiring the need to run an additional signature scoring method relying on the expression of a predefined set of markers for this biological phenomenon.

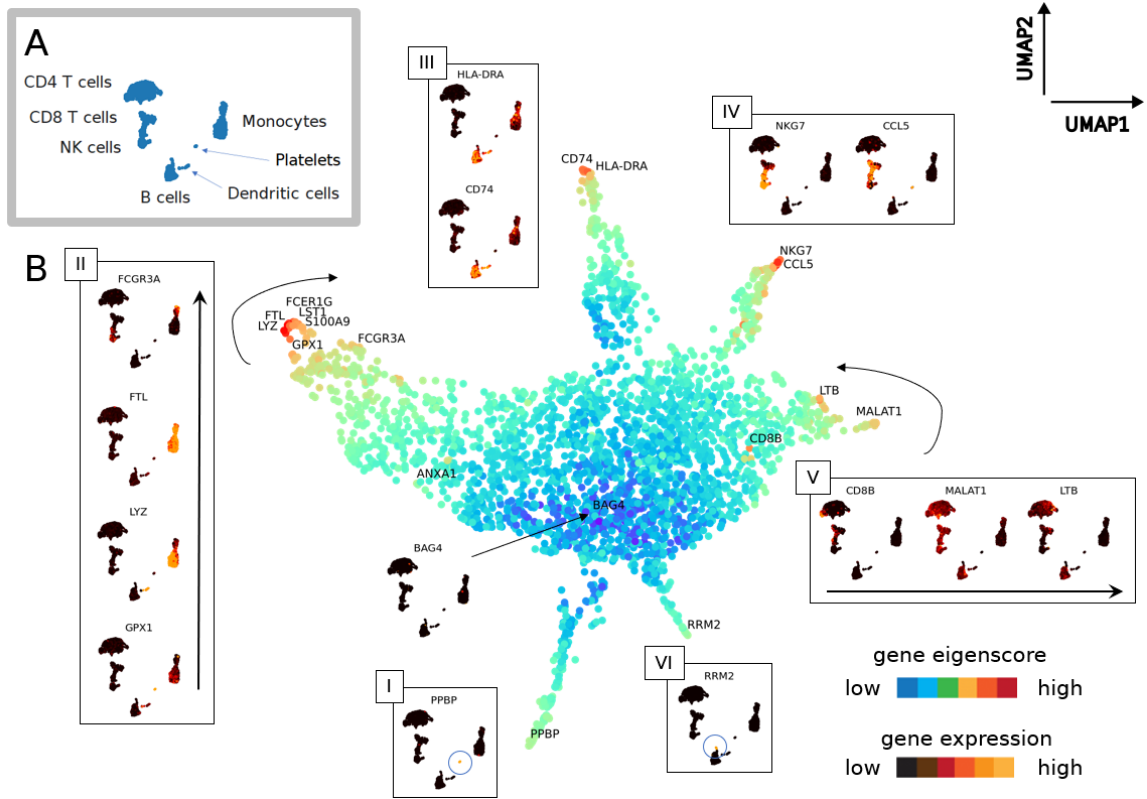


Figure 2.3: Geometry of cell space and gene space. (A) Cell types in PBMC data [27]. (B) UMAP of genes set in eigenscore space for eigenvectors 1–16. Genes (dots) are colour-coded for the logarithm of the norm of the vector in 16-dimensional eigenscore space. Genes with similar expression patterns in the PBMC single-cell data [27] plot close together in eigenscore space, and expression patterns vary continuously as we move through this space. The outward branches I–VI correspond to genes that are expressed highly on specific groups of cells.

2.3.2.2 Eigenscores for Analysing Data with Continuous Structure: T Cells

We next compute eigenscores of the T cell data [46] (corresponding to $0 < \lambda < 0.1$). As before, we can visualise the gene space in Figure 2.5A where genes are near to another gene if they have similar expression patterns. For example, a large number of mitochondrial genes, such as *MT-CO3*, that are lowly expressed in cells across the data set are grouped together as a distinct gene cluster. The detection of cells expressing high levels of such mitochondrial genes can be an important step in filtering poor quality cells from scRNA-seq data sets [81].

The continuous nature of the T cell data set is reflected in the many intermedi-

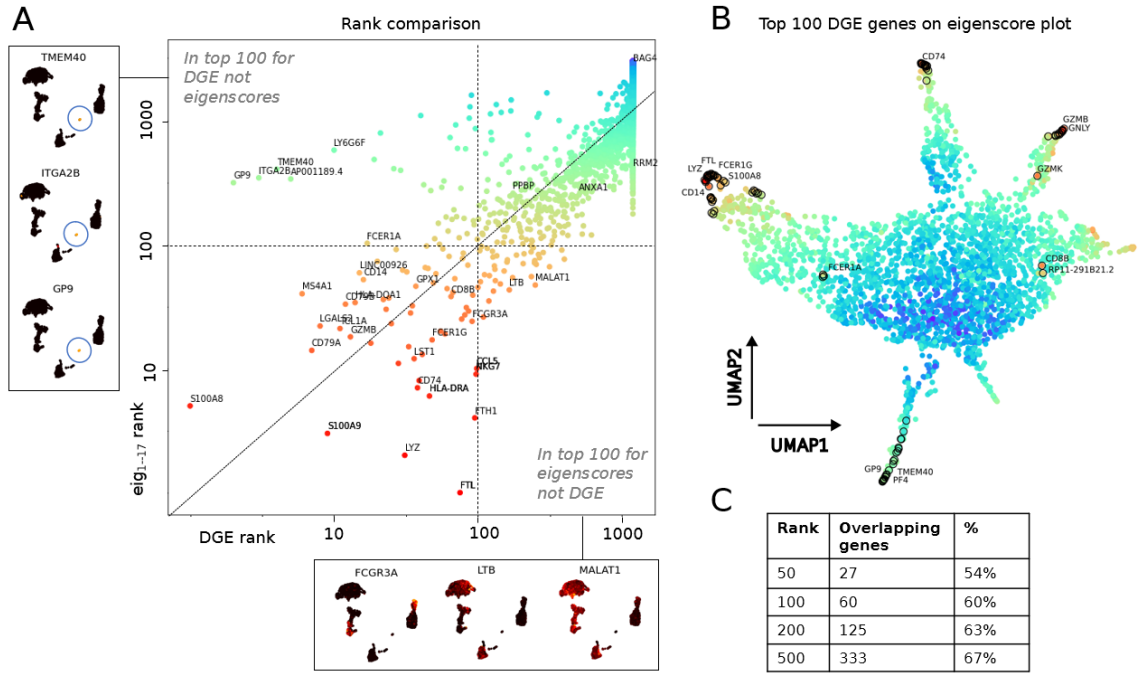


Figure 2.4: Eigenscores compared to differential gene expression (DGE) on PBMC data set [27]. (A) Comparative study of DGE ranking using Seurat clustering and a non-parametric Wilcoxon rank sum test (log of rank computed from adjusted p-value on x-axis) versus ranking by norm in eigenscore space (log of eigenscore rank of 16 lowest frequencies on y-axis). Example genes in top 100 for one ranking but not the other shown on the sides. (B) Top 100 genes ranked by adjusted p-value in DGE marked on the eigenscore UMAP plot of genes from Figure 2.3. Two regions in the UMAP not found in the top of DGE are branch V from Figure 2.3B (T cell and lymphocyte genes that are expressed in larger groups of cells); branch VI (genes expressed in RRM2+ cluster that is not found by DGE). (C) Quantitative comparison of gene ranks given by adjusted p-value in DGE versus norm in 16-dimensional eigenscore space.

ate to high eigenscore genes that show coherent regions in the data set on multiple scales. While we can identify groups of cells that have coherent gene expression behaviour, such as the clusters formed by *EEF1A1*+ cells, *HBB*+ cells, *ANXA1*+ cells and *HSPA1A*+ cells, we also find genes that have unique expressions that are unlike any other gene signals (e.g., *GNLY* and *GZMB*, which are both known secreted cytotoxic effector genes found across T and NK cell subsets). To explore the geometry of genes further, we analyse the top 20 genes ranked by eigenscore norm in 1–19 dimensional eigenscore space. We find *AREG* (9th in eigenscore rank) does

not show up on the DGE ranking. As shown in the UMAP cell subfigure, *AREG* is expressed *in between* clusters of cell types, connecting NK cells and a cluster of CD8 T cells (Figure 2.2, see Section 2.3.3.2 for biological implications). In this way, eigenscores provide insight into both the continuous and discrete nature of T cell behaviour.

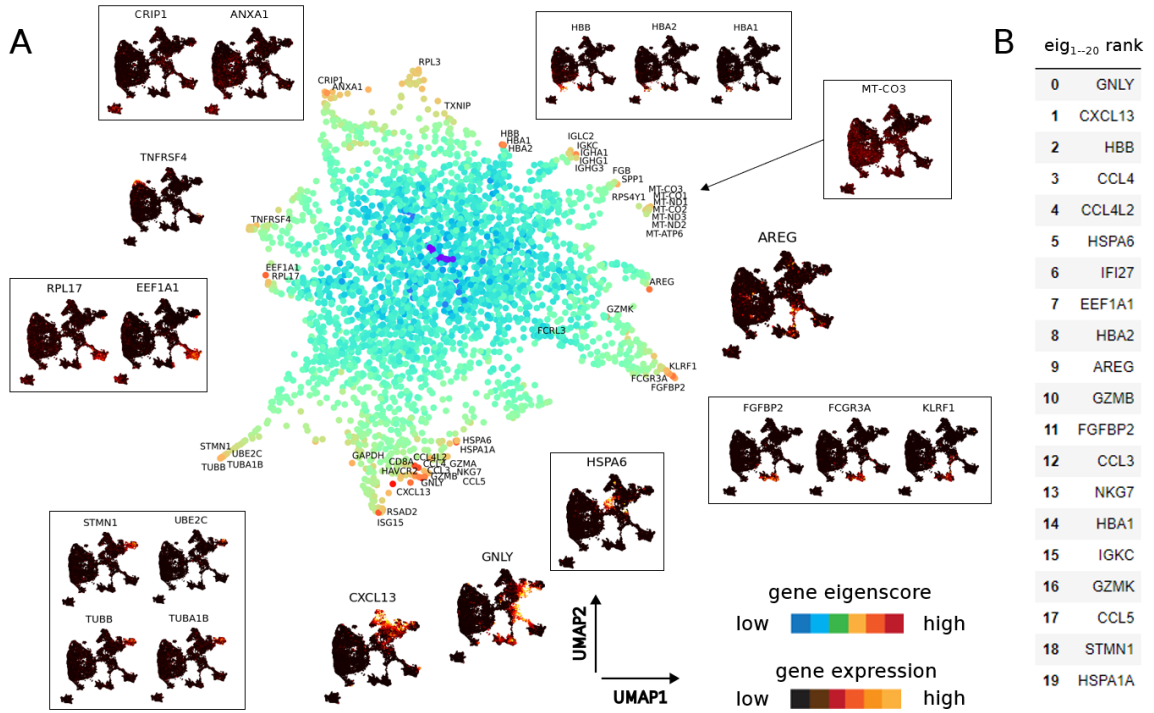


Figure 2.5: (A) UMAP of genes from T cell data set [46] in eigenscore space for eigenvectors 1–19, colour-coded for the logarithm of norm of the vector in 19-dimensional eigenscore space. Genes with similar expression group together and reveal substructure in the data set. Some genes have unique expression patterns not matched by other genes. Boxed genes represent a group of genes with similar expression whereas unboxed genes represent isolated gene behaviour. (B) Top 20 genes ranked by norm in 1–19 dimensional eigenscore space.

2.3.3 Multiscale Laplacian Score

The multiscale Laplacian score (MLS), similar to the 0-dimensional combinatorial Laplacian score [32] and gene connectivity score [67], extends DGE to settings in which a stable partition of cells into groups is not feasible; therefore, no assignment

of cells into groups is required. The MLS ranks genes by their consistency with the topological structure of the data set and performs such topological consistency analyses at multiple resolutions.

The resolutions are determined by finding scales in the data that provide stable community structures. We reiterate that the MLS calculation does not use the obtained communities; rather, we use the resolution that provides a stable communities via local minima in variation of information (VI). We highlight that the communities we find in both the PBMC and T cell data increase in size with increasing Markov time (see Figures 2.6 and 2.7, panels A). This size increase illustrates that the MLS at the selected Markov times tests for consistency at different scales in gene space.

2.3.3.1 Multiscale Laplacian Score of PBMC Data

In Figure 2.6, we apply MLS to the PBMC data set [27]. These data permit a stable clustering into five larger groups of cells. However, these clusters contain non-stable substructures (see Figure 2.6A). The substructures largely align with the clusters found in the Seurat VST vignette [35] (Figure 2.8). We find that the genes *GZMK* and *CD8B* exhibit a higher consistency with the structures at the first resolution (t_1) than at later resolutions (Figure 2.6B). The gene *GZMK* is highly expressed on the intersection of two communities at t_1 which correspond to naive and memory CD8 T cells, but it is not highly expressed on the union of these two communities (the two communities merge at resolution t_2). *GZMK* seems to mark a transition between these two clusters. Similarly, *CD8B* expression is detected within the left-most community assigned to the broader CD4 T cell cluster of the UMAP plots (Figure 2.8), which is merged into a larger community at t_2 . The genes *GZMB* and *XCL2* are examples of features with low MLS at t_2 but relatively high MLS at t_3 . The former is highly expressed on a community corresponding to

effector CD8 T cells (cluster 6 in Figure 2.8), the latter on the intersection of the effector and the naive/memory T cell communities (clusters 6 and 5 and 7 in Figure 2.8). At resolution t_3 , clusters 5 and 7 are merged with cluster 6. The communities at resolution t_3 correspond to the seven populations describing the different cell types with the NK and CD8 T cells merged.

Examples of genes with low MLS at t_3 , relative to the standard LS as in [32] and MLS at other resolutions, include *AIF1* and *CTSS*. Both genes are highly and consistently expressed on the community consisting of *CD14+* and *FCGR3A+* monocytes (Figure 2.8) [70, 35]. Resolution t_3 is the first resolution at which *CD14+* and *FCGR3A+* monocytes form a single community.

2.3.3.2 Multiscale Laplacian Score of T Cell Data

We compute the MLS to a human T cell data set from Labrechts et al. [46] (Figure 2.7). As remarked by [32], these data do not allow for any partitioning into stable clusters (see high VI values in Figure 2.7A). Next, we determine three resolutions of interest based on the VI. Genes with a relatively low MLS at the finest resolution, t_1 , include *IGKC* and *IFI27*. Both are highly expressed on a small group of cells (in the center of left hand side and top right of UMAP plot respectively; see Figure 2.7B).

The gene *IGKC* is an immunoglobulin gene, an antibody component found in B cell subsets, particularly plasma cells [48]. Cells expressing *IGKC* are also *JCHAIN+* and positive for antibody subtypes suggestive of class switching (e.g., *IGHG1* and *IGHA1*). Since this is a T cell data set, this almost certainly indicates that the cells in question are doublets (two cells in the same experimental droplet), specifically T cells binding B cells. While not representing single cell states, it is important that these readings are picked up in the analysis.

The gene *IFI27* is part of an antiviral/interferon-induced (IFI) response signa-

ture. It is particularly interesting that MLS can detect a specific transcriptional programme shared across multiple cell types (CD4+ and CD8+ T cells).

This could represent a shared T cell programme directed against viruses or induced during stress responses (e.g., for scRNA-seq processing) [89]. As a particular example of informative gene prioritisation, we highlight *AREG* at resolution t_2 , where it is expressed highly on a group of cells bridging nearby clusters identified as NK cells and CD8 T cells based on overall marker expression of each cluster (Figure 2.2). Within the immune system, *AREG* is expressed by subsets of NK cells and other types of innate lymphoid cells (ILCs) [18, 58], where it plays an important role in mediating type 2 immunity [58, 104]. Despite bridging different clusters in our global clustering and that of the original authors [46], this population likely represents cells in various states of transition between two previously described *AREG*+ NK cell phenotypes: one with high levels of secreted molecules associated with effector functions (*CCL3*, *CCL4*) and the other expressing homing receptors associated with a more circulatory phenotype (*CD44*, *SELL*) [18] (see Figure 2.10). Therefore, while the original authors identified these two subsets as discrete NK and type 1 ILC-like cell types, respectively ([46] Figure S13), both our eigenscore and MLS-informed approaches highlight *AREG* as a shared feature, supporting the notion that these two populations may be consistent with a more continuous transition between *CCL3*+ and *SELL*+ states within the NK cell population. This interpretation is further supported by the preserved expression of NK cell markers (e.g., *CD94/KLRD1*, *NKG2A/KLRC1*) in the *SELL*+*AREG*+ cells, which is often considered to be a feature of NK cells that is not shared by otherwise closely related type 1 ILCs [7, 8].

Similarly, *GZMB* is highly expressed on the intersection of exhausted and proliferating T cells (see Figure 2.2), two clusters of which are visible in the community structure at t_2 but merge at t_3 . Finally, at Markov time t_3 , *FGFBP2* and *NKG7* are

examples of genes with relatively low expression that are highly and consistently expressed on the cluster of NK cells.

Table 2.1: Differentially expressed genes for PBMC data found by Seurat.

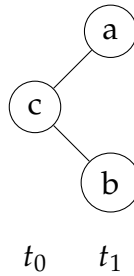
0	1	2	3	4	5	6	7	8	9	10	11
RPS27	LTB	S100A8	CD79A	IFITM3	GZMK	GZMB	GZMH	CD8B	FCER1A	IFIT1	GP9
RPL32	IL32	LGALS2	MS4A1	RP11-290F20.3	CCL5	FGFBP2	CST7	RP11-291B21.2	ENHO	IFIT3	ITGA2B
RPS6	IL7R	S100A9	TCL1A	LST1	NKG7	SPON2	NKG7	CD8A	CLEC10A	RTP4	TMEM40
RPS12	CD3D	CD14	CD79B	AIF1	LYAR	GNLY	CCL5	S100B	SERPINF1	SPATS2L	AP001189.4
RPL31	AQP3	FCN1	HLA-DQA1	MS4A7	GZMA	PRF1	GZMA	CARS	CD1C	DDX58	LY6G6F
RPS14	LDHB	TYROBP	LINC00926	IFI30	IL32	XCL2	FGFBP2	RPS12	CACNA2D3	RSAD2	sep-05
RPS25	CD2	MS4A6A	VPREB3	CD68	CD8A	AKR1C3	CD8A	RPL13	HLA-DQB2	MX1	HGD
LDHB	CD40LG	LYZ	HLA-DQB1	FCER1G	CTSW	CLIC3	GZMB	RPS6	HLA-DQA2	ISG15	PTCRA
RPS3A	TPT1	GPX1	CD74	CFD	CST7	KLRD1	CTSW	CCR7	HLA-DQA1	IFI6	TREML1
RPL30	CD3E	CST3	HLA-DRA	SERPINA1	HOPX	CST7	CCL4	RPL32	NDRG2	HERC5	ITGB3

2.4 Persistent Rayleigh Quotient

2.4.1 Application to Cell Bifurcation

As introduced in Section 1.5, we have designed the persistent Rayleigh quotient to analysis signals on graphs with time information. Such an example occurs in single-cell RNA-seq within the process of cell differentiation.

We consider a simple bifurcation model, where a multipotent parent cell type c that exists at developmental time t_0 can differentiate into two different daughter cell types a and b that exist at developmental time t_1 , as in the following graph:



Assigning time data to a such a bifurcation may come from original experimental design, as we will see in Section 2.4.2, or by inference from already collected data by using trajectory inference algorithms, seek to assign a pseudotime to each cell by fitting a tree onto the data set [68, 92] or fit a statistical model to each

gene and then test against a null model [90, 88, 63, 51, 42]. The Rayleigh quotient and Laplacian score have proven useful in selecting genes [32] but are agnostic to any prior cell knowledge or meta data. Here, we use additional time information to filter the graph and seek to separate genes that have different roles in this differentiation process.

Now consider the following graph signals, representing expressions of genes:

$$g_1 : a, b, c \mapsto 1, 1, 1,$$

$$g_2 : a, b, c \mapsto 1, 0, 1,$$

$$g_3 : a, b, c \mapsto 1, 1, 0,$$

$$g_4 : a, b, c \mapsto 0, 1, 0.$$

Figure 2.11 displays the results of comparing the usual normalised Rayleigh quotient, corresponding to $\widehat{\text{PRQ}}(t_1 - t_0, t_1 - t_0)$, to the normalised persistent Rayleigh quotient $\widehat{\text{PRQ}}(0, t_1 - t_0)$, where we are using reversed time. As shown in Figure 2.11C, these measurements separates the binary graph functions g_i on G .

Interpreting these signals as gene expressions, gene g_2 lies above the diagonal in Figure 2.11C and is highly expressed in the parent cell type and only one of the daughter cell types. Such behaviour indicates that a gene, or its transcriptional regulators, may be involved in cell differentiation towards a particular lineage or fate. Similarly, gene g_3 , which lies below the diagonal, is expressed in both daughter cell types but not the parent cell type, and, as such, it may correspond to a class of genes which represents markers of differentiation shared across cell fates. Genes corresponding to g_1 are expressed at a constant rate, representing possible ‘house-keeper’ genes and, thus, have zero (or close to zero) persistent Rayleigh quotients. Finally, gene g_4 is only expressed in a single daughter cell type and, as such, represents markers of differentiation to a particular cell fate. Gene g_4 lies along the

diagonal.

2.4.2 PRQ Applied to Mouse Foetal Liver Cell Data Set

In this section, we apply the PRQ to a scRNA-seq dataset of 447 mouse foetal liver cells at different stages of development [101]. The aim of the mouse liver experimental data was to identify divergence in gene expression (and associated time-course) as progenitor cells in the liver (i.e., hepatoblasts) specialise or differentiate into two different types of mature cells, hepatocytes and cholangiocytes. Cells in the mouse foetal liver data set were sampled on embryonic days 10, 11, 12, 13, 14, 15, and 17.

Preprocessing

As the mouse data set is substantially smaller than the other two datasets considered so far, we used a simpler preprocessing strategy. The public data from [101] were provided in transcripts-per-million (TPM), and we further applied a $\log_e(x + 1)$ transform. The 10,000 most highly varying genes were retained and the UMAP-weighted k -nn graph was built using $k = 3$ with the Euclidean metric. Figure 2.12 displays the resulting graph, plotted on the first two principal components of the data.

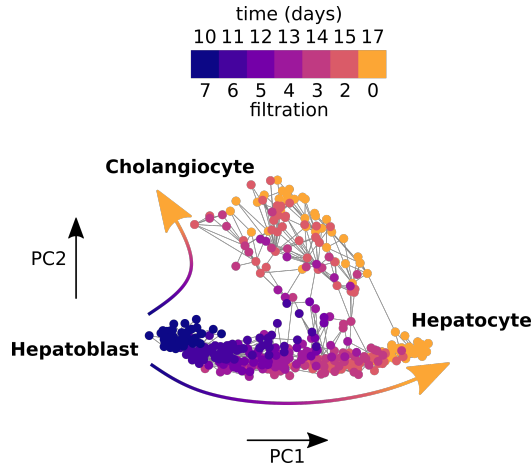


Figure 2.12: A weighted graph constructed from mouse foetal liver cells sampled from days 10–17 during development. Parent cell type hepatoblasts differentiate into two daughter cell types, cholangiocytes and hepatocytes.

Previous Results on Mouse Foetal Liver Cell Data Set

Yang et al. [101] selected 1761 heterogeneously expressed genes that correlate with the first two principal components of the data which were then clustered. In a separate study on different data, Mu et al. [59] ranked genes that were differentially expressed in hepatocyte development.

PRQ Analysis

We apply the PRQ to a bifurcation describing the differentiation of mouse hepatic cells by Yang et al. [101] (see Figure 2.13). Hepatoblasts, hepatocytes, and cholangiocytes were sampled from mouse embryos at seven time points, from embryonic day 10 to day 17. Hepatoblasts are a parental cell type whose daughter cells differentiate into hepatocyte and cholangiocyte cell types. As the topologically interesting direction is in ‘reverse time’, we assign a filtration to the graph by assigning a node from day t the filtration value $17 - t$. For each pair of steps i and j with $i \leq j$ in the filtration of the graph, we apply the normalised persistent

Rayleigh quotient to a particular gene. We represent this as a two-dimensional⁴ score for each gene in Figure 2.13. In order to show the largest differences between parts of the PRQ score for a gene, we compare the normalised persistent Rayleigh quotients $\widehat{\text{PRQ}}(2, 7)$ with $\widehat{\text{PRQ}}(7, 7)$ in Figure 2.13C.

We have highlighted genes that were found to be differentially expressed during hepatoblast differentiation in [59] (Figure 2.13A,B,D,E). As expected, genes such as *Tubb5*, *Mdk*, and *Igfbp1*, which are expressed in hepatoblasts and only one of hepatocytes or cholangiocytes, have a higher value for the persistent Rayleigh quotient than the full Rayleigh quotient (Figure 2.13A,B). The gene *Mdk* is known to show a decrease during hepatoblast maturation towards hepatocytes in utero, corresponding with the upregulation of genes involved in hepatocyte function including *Aldob* [59, 82]; however, this gene is preserved in cholangiocyte populations [84], agreeing with our observations using the PRQ. *Igfbp1* is not known to play a role in differentiation into hepatocytes; however, the expression of *Igfbp1* has been previously shown in hepatocytes, where it has a prosurvival role that can be enhanced by p53 activity [50]. Genes *Aldob* and *Mt2* are expressed in cholangiocytes and hepatocytes but not hepatoblasts. Hence, their persistent Rayleigh quotient has a lower value than the full Rayleigh quotient (Figure 2.13D). Finally, the persistent Rayleigh quotient and full Rayleigh quotient for *Fabp1* and *Ahsg* are almost the same. This corresponds to the fact that *Fabp1* and *Ahsg* are highly expressed in only one of the daughter cell types (Figure 2.13E). The full Rayleigh quotient $\widehat{\text{PRQ}}(7, 7)$ can sort genes based on how coherently expressed they are with respect to the underlying graph, but it does not distinguish between different expression patterns relevant to development. In contrast, the persistent Rayleigh quotient can differentiate genes whose expression pattern is relevant to bifurca-

⁴This is reminiscent of the birth–death persistence diagram in topological data analysis where the x-axis records the filtration step where certain topological features first appear, known as the birth time, while the y-axis records where these features finally vanish, known as the death time.

tion.

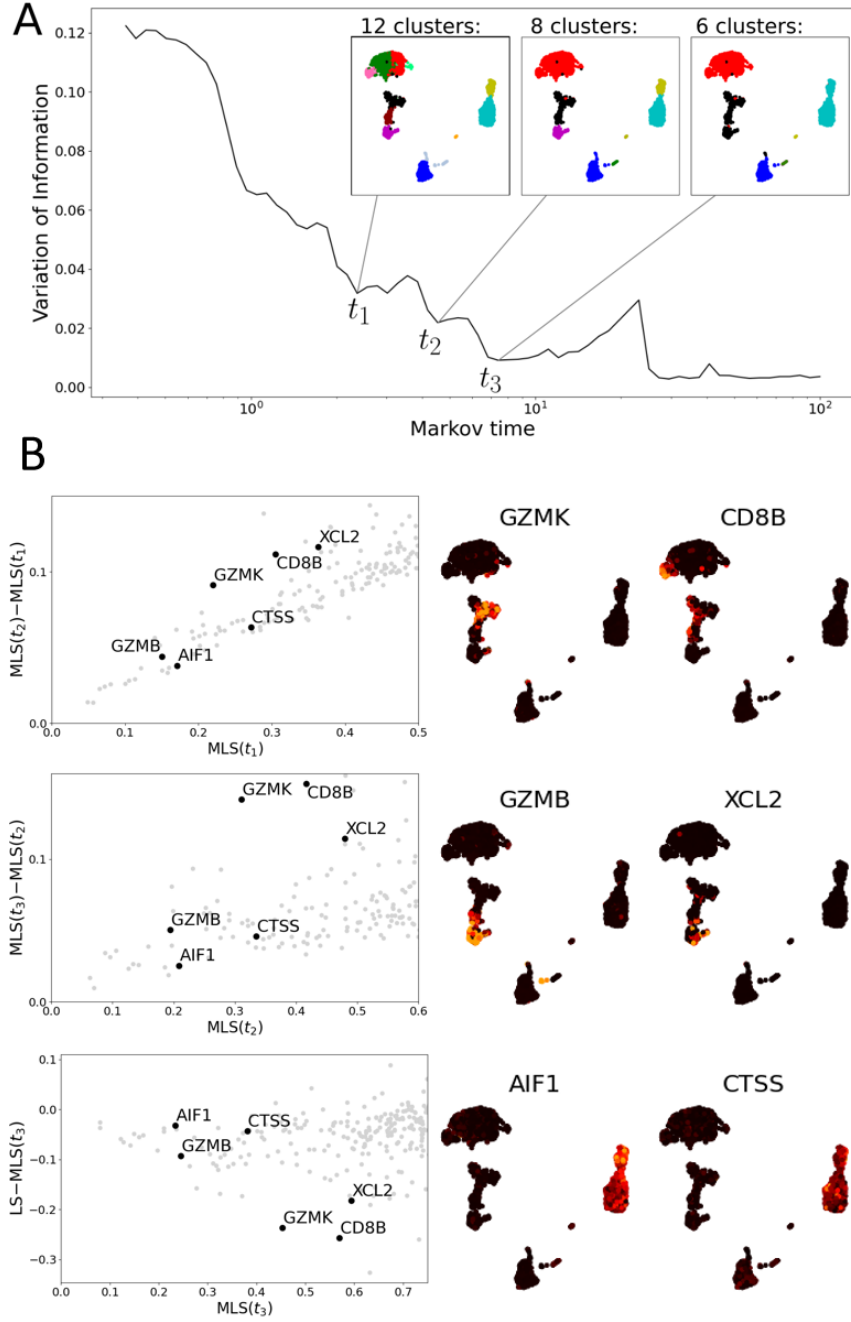


Figure 2.6: Multiscale Laplacian scores of PBMC data set [27]. (A) The graph of variation of information of community structures returned by 100 iterations of the Louvain algorithm at each Markov time. Local minima indicate stable community structures and, hence, scales of interest. The community structures at three such minima are shown by colourings of UMAP plots. (B) Left: three scatter plots comparing the multiscale Laplacian scores of genes (grey dots) at successive times to one another (upper two) and of t_3 to the combinatorial Laplacian score (in all plots, axes are truncated). We highlight 6 genes of interest (annotated). Middle and Right: UMAP plots visualising the gene expression of six genes selected based on their MLS.

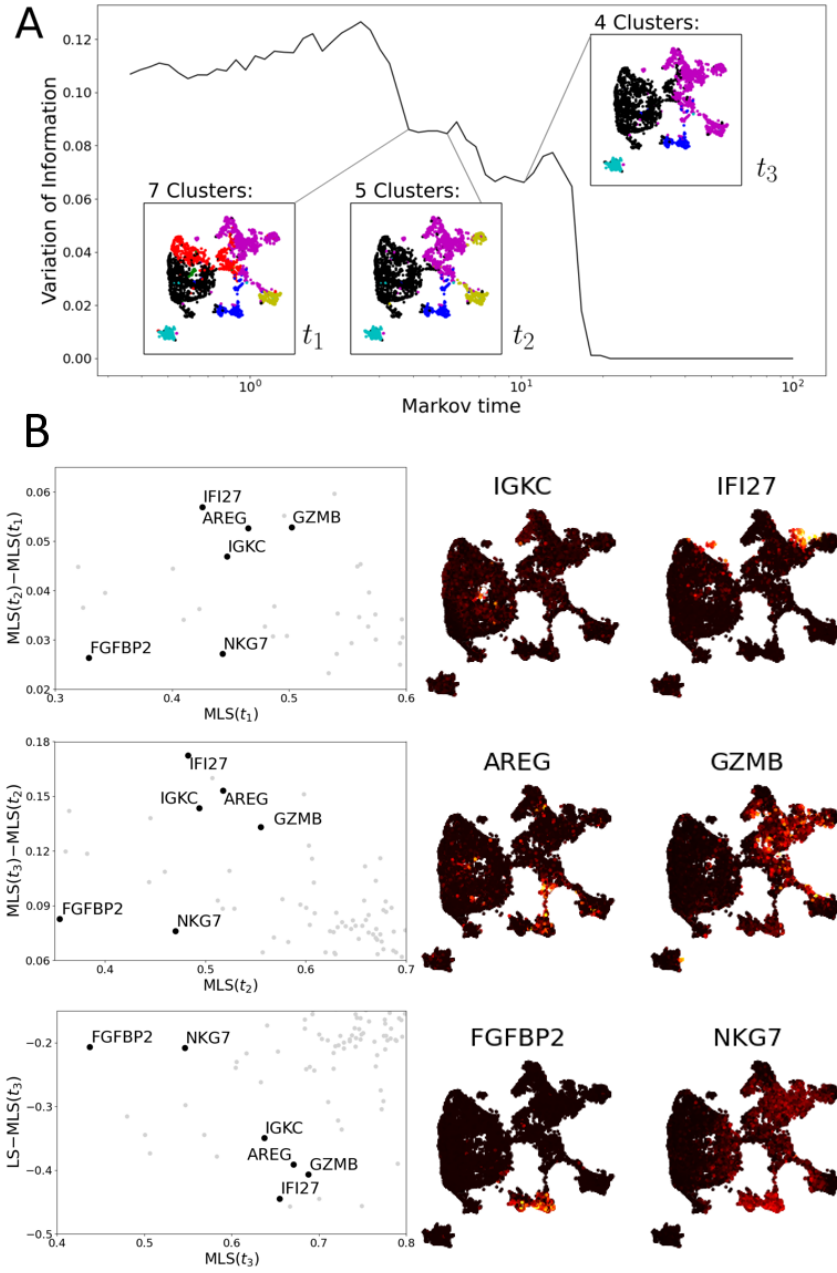


Figure 2.7: Multiscale Laplacian score of human T cell data set [46]. (A) The graph of variation of information of community structures. Again, local minima (or plateaus) indicate scales of interest. Community structures at three scales are picked out. (B) (Left): three scatter plots comparing the multiscale Laplacian scores of genes (grey dots) at successive times to one another (left and middle plot) and of t_3 to the combinatorial Laplacian score (in all plots, axes are truncated). We highlight 6 genes of interest (black dots; annotated). (Middle and Right): UMAP plots visualising the gene expression of six genes selected based on their MLS.

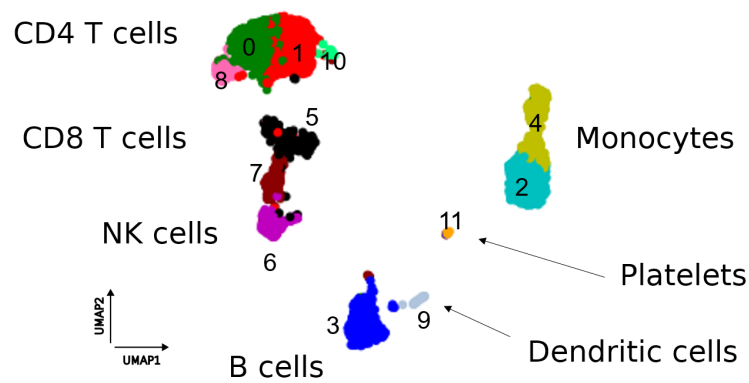


Figure 2.8: Seurat clusters on PBMC data [27] from Seurat VST Vignette [35] numbered according to the vignette and interpretations of overarching cell types inferred from previous results. Plot uses a 2D UMAP projection. The cells in this data set divide into broad clusters corresponding to the cell types found in peripheral blood mononuclear cells: lymphocytes (T cells, NK cells, B cells), monocytes, and dendritic cells, as also platelets which are not mononuclear but are found in this specific data set. The DGE analysis from Seurat (non-parametric Wilcoxon rank sum test [69]) defines twelve smaller clusters, in particular subclustering T cells, NK cells and monocytes, and searches only for differentially expressed genes on these subclusters.

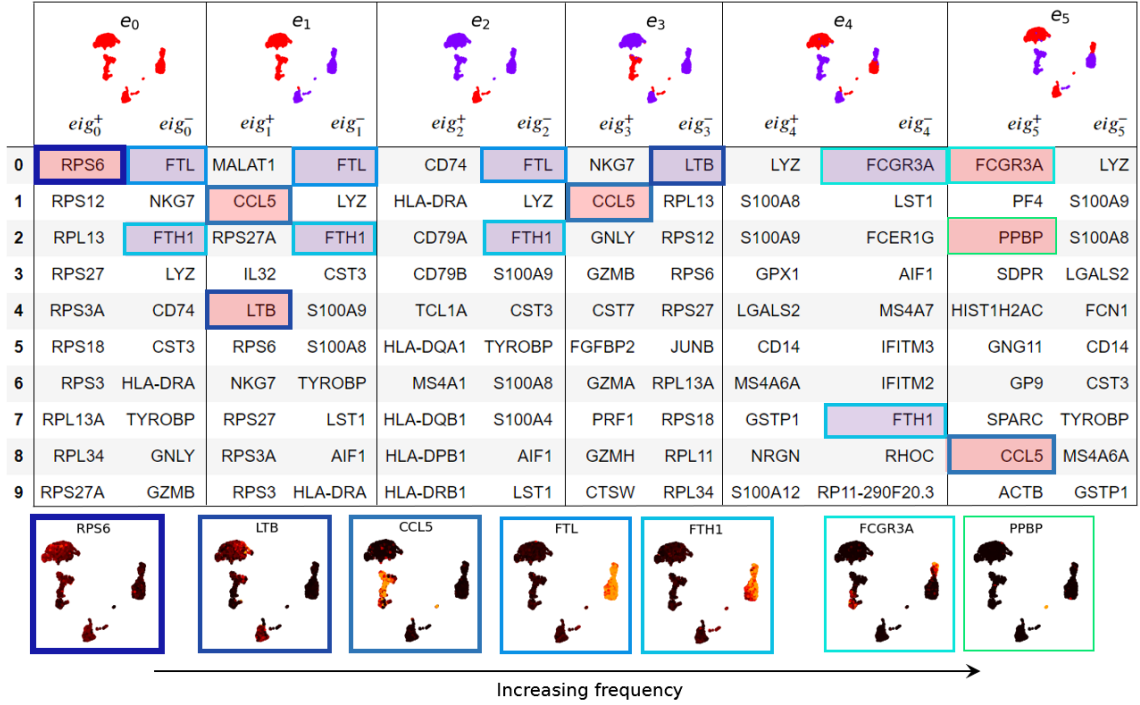


Figure 2.9: Eigenscore ranks for PBMC data [27]. On the top row are plots of the Laplacian eigenvectors, coloured by sign (red positive, purple negative). For each eigenvector e_i , genes are listed with the highest alignment (eig_i^+) and highest anti-alignment (eig_i^-) with e_i . Below the table are a selection of genes ranked highly by eigenscores. For example gene *FTL* shown below the table is strongly expressed on the monocyte cluster on the right, which is purple (negative) for both e_1 and e_2 , hence *FTL* has high scores on eig_1^- and eig_2^- .



Figure 2.10: Expression of relevant marker genes in the T cell data set.

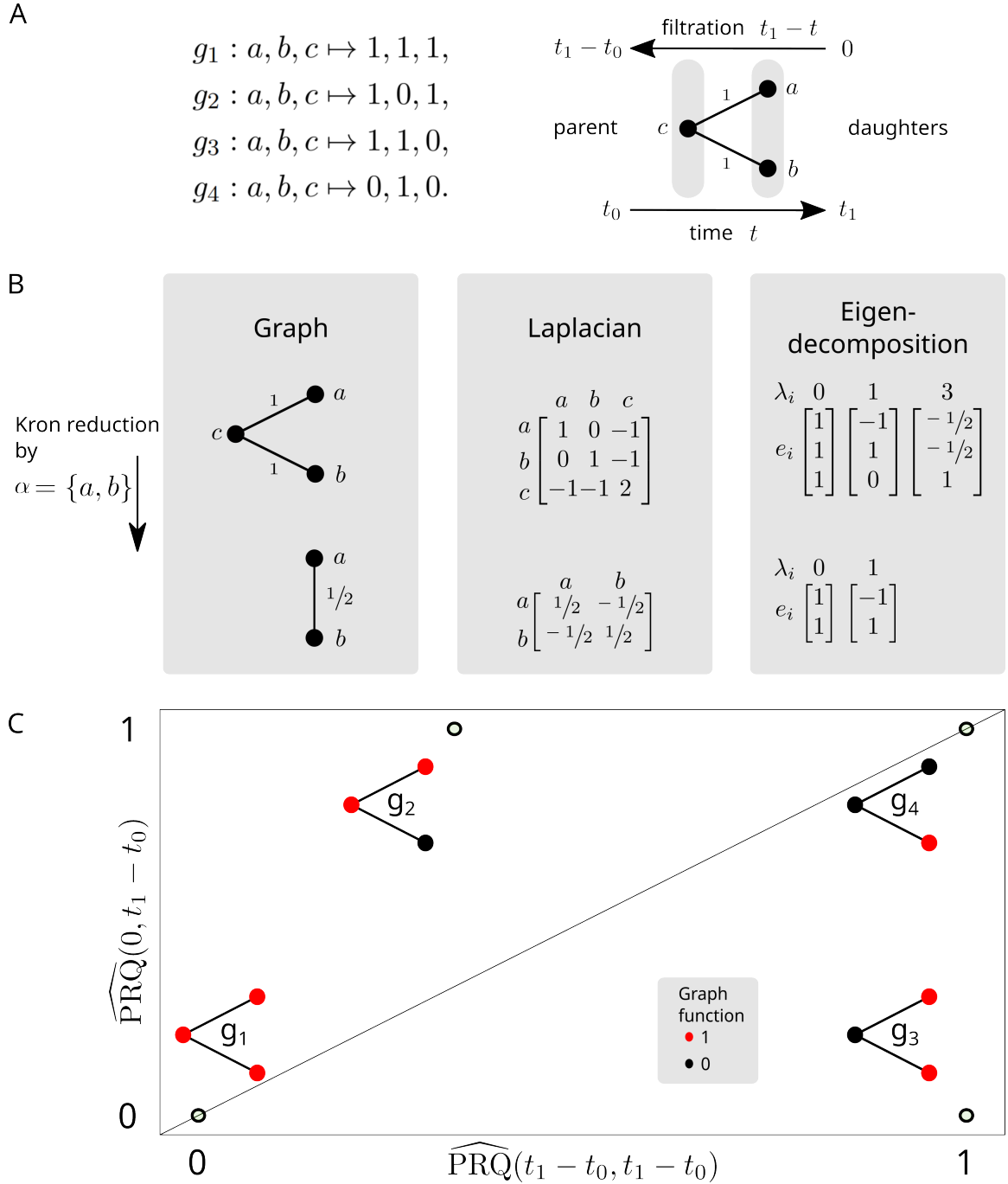


Figure 2.11: The persistent Rayleigh quotient for cell differentiation. **(A) (left)** Signals (genes) on the graph that we aim to differentiate. **(right)** The model for the bifurcating differentiation process. **(B)** The effects on the graph and graph Laplacian after applying the Kron reduction process 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 high and low gene expression of a particular gene. The persistent Rayleigh quotient separates these genes based on relevance to the bifurcation: g_1 is expressed in all cell types, g_2 is expressed in the parent and one daughter cell type, g_3 is expressed only in both daughter cell types, g_4 is expressed only in one daughter cell type.

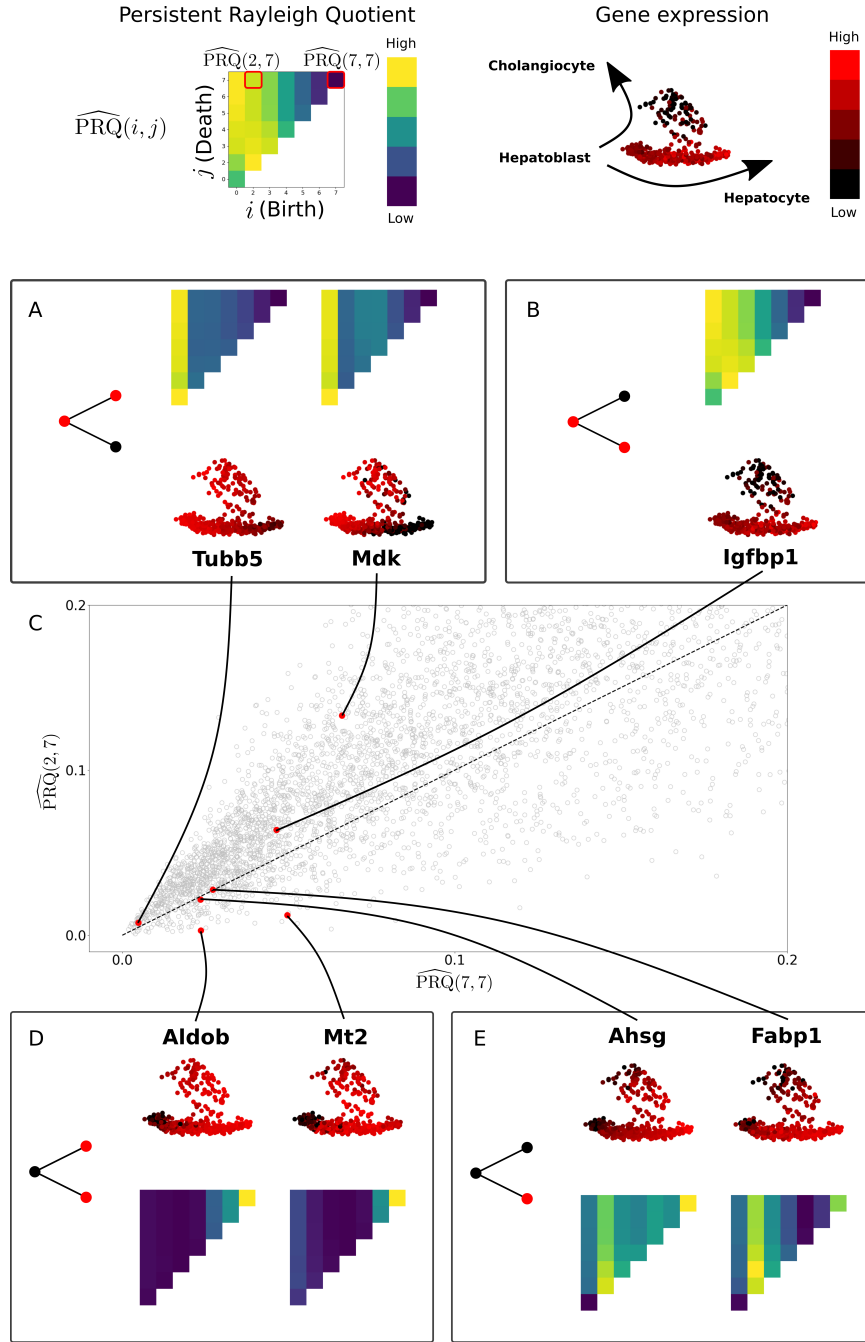


Figure 2.13: The persistent Rayleigh quotient separates genes by their role in a cell differentiation process. The PRQ is parameterised by birth (i) and death (j), each pair (i, j) assigning a non-negative number to every gene. We plot these values for each gene for $(i = 7, j = 7)$ on the x-axis and $(i = 2, j = 7)$ on the y-axis on subfigure (C). Selected for display (A,B,D,E) are top differentially expressed genes from [59] on the data from [101] (see Figure 2.12). 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 they lie below the diagonal. Genes *Ahsg* and *Fabp1* are only expressed in one daughter cell type (E) and lie on the diagonal (compare with Figure 2.11).

Chapter 3

Quiver Laplacians

So far, we have explored how the graph Laplacian can be used to analyse signals on graphs, such as gene expression signals on cell-cell networks. A graph signal is scalar valued: $f : V \rightarrow \mathbb{R}$. However, there are some data that are best represented as a vector-valued function on a graph. For instance, a colour image can be thought as a 3-dimensional vector on a grid, with each component is the red, green, and blue values respectively. Another example is multi-omic single-cell data: there now exist technologies that can sequence not only RNA expression but other modalities such as chromatin accessibility (discussed in Chapter 5) or spatial positioning (spatial transcriptomics) simultaneously for each cell. The apparatus of Chapter 1 can be adapted to a function $f : V \rightarrow \mathbb{R}^n$ by simply applying the graph Laplacian to each component f_i , however, this has several deficiencies. Firstly, this does not allow comparisons between f_i and f_j for different i and j . In the multi-omics example, we are interested in how the different modalities relate, to answer questions about how e.g. DNA modifications affect gene transcription, or how a gene's transcription varies spatially across cells. Secondly, our vector data may change dimension across our graph: in the multi-omics case, we may have only measured one modality on a subset of the data.

In this chapter, we shall model such data as a *quiver representation*, a graph where each vertex and edge is assigned a vector space and linear map respectively. We shall then introduce a Hermitian positive semi-definite operator that generalises the graph Laplacian to quiver representations. We shall see that the equivalent of the harmonic graph function (i.e. in the kernel of the Laplacian) is the notion of a *global section* of a quiver representation. The quiver Laplacian allows us to approximate such global sections, which will need in Chapter 4 when we apply the quiver Laplacian to feature selection in epigenetic data. The rest of the chapter is then devoted to proving various spectral stability results of our Laplacian under morphisms of representations and natural graph operations.

The quiver Laplacians is closely related to other flavours of Laplacian, in particular the cellular sheaf Laplacian [37] (but see also [29, 60]) and the ideas behind principal bundles on graphs [25].

3.1 Quiver Representations

A *quiver* Q is a very general notion of a graph: it consists of a set of vertices Q_0 , a set of edges Q_1 and two maps $s, t : Q_1 \rightarrow Q_0$, denoting the source and target of each edge. More verbosely, a quiver is simply an unweighted loopy multi-edged directed graph. As usual, in this chapter we will only consider finite quivers, where Q_0 and Q_1 are finite.

Given some category $\mathbb{D}$, a *representation* of $\mathbf{A}_\bullet$ of Q valued in $\mathbb{D}$ consists of the following data: for each vertex $v \in Q_0$, an object $\mathbf{A}_v \in \mathbb{D}$, and for each edge $e \in Q_1$, a morphism $\mathbf{A}_e : \mathbf{A}_{s(e)} \rightarrow \mathbf{A}_{t(e)}$ in $\mathbb{D}$. In this thesis, we will only consider $\mathbb{D}$ to be the category $\mathbf{FdHilb}(\mathbb{k})$ of finite-dimensional Hilbert spaces over the field $\mathbb{k} = \mathbb{R}$ or $\mathbb{C}$, i.e. finite-dimensional vector spaces with inner products. Crucially

for our purposes, every morphism $A : U \rightarrow V$ in this category¹ admits an *adjoint* morphism $A^* : V \rightarrow U$ characterised by $\langle Au, v \rangle = \langle u, A^*v \rangle$ for all (u, v) in $U \times V$.

For a quiver representation $\mathbf{A}_\bullet$, the *dimension vector* $\underline{\dim} \mathbf{A}_\bullet$ is defined as the integer vector $(\dim \mathbf{A}_v)_{v \in Q_0}$ while the *total dimension* of $\mathbf{A}_\bullet$ is defined as

$$\dim_{\text{Tot}} \mathbf{A}_\bullet = \sum_{v \in Q_0} \dim \mathbf{A}_v.$$

We will now give some prototypical examples of quiver representations.

Example 3.1.1. In the following, let Q denote a quiver.

- 3.1.1(a) The *trivial* representation $\mathbf{0}_\bullet$ of Q is the unique representation where $\mathbf{0}_v = \mathbf{0}$ for all $v \in Q_0$.
- 3.1.1(b) Suppose C is a finite-dimensional vector space with an inner-product. The *constant* representation with value C is the representation $\mathbf{C}_\bullet$ where $\mathbf{C}_v = C$ and $\mathbf{C}_e = \text{Id}_C$ for all $v \in Q_0$ and $e \in Q_1$. If $C = \mathbb{k}$ for the underlying field $\mathbb{k}$, we call this constant representation the *unital* representation² $\mathbf{1}_\bullet$.

Suppose G is a graph with an orientation on its edges, making it a quiver. Then the unital representation of G is essentially a copy of G , while a weighting of the edges of G corresponds to a representation whose dimension vector is all 1s, and whose maps all have positive sign.

The following example will be an important source of representations from data that we will study in Chapters 4 and 6.

Definition 3.1.2. Let Q be a quiver and suppose H is an inner product space.

A *subspace arrangement* of Q in H is a finite set $\mathbf{S} = \{\mathbf{S}_v\}_{v \in Q_0}$ of finite-dimensional subspaces of H . For each $v \in Q_0$ denote the inclusion of S_v in H by $\iota_v^{\mathbf{S}}$ and the or-

¹**FdHilb** is an example of a *Dagger category*: a category $\mathbf{D}$ with a functor $\dagger : \mathbf{D}^{\text{op}} \rightarrow \mathbf{D}$ such that $\dagger$ fixes objects and is involutive on morphisms.

²This name is not standard.

thogonal projection of H onto $\mathbf{S}_v$ as $\pi_v^{\mathbf{S}}$. Then $\mathbf{S}$ induces a representation of Q , which we will write as $\mathbf{S}_\bullet$, where the map associated to an edge $e \in Q_1$ is

$$\mathbf{S}_e = \pi_{t(e)}^{\mathbf{S}} \iota_{s(e)}^{\mathbf{S}}.$$

A *morphism* of quiver representations $\tau : \mathbf{A}_\bullet \rightarrow \hat{\mathbf{A}}_\bullet$ is a collection of linear maps

$$\tau = \left\{ \tau_v : \mathbf{A}_v \rightarrow \hat{\mathbf{A}}_v \right\}$$

such that the diagram

$$\begin{array}{ccc} \mathbf{A}_{s(e)} & \xrightarrow{\mathbf{A}_e} & \mathbf{A}_{t(e)} \\ \downarrow \tau_{s(e)} & & \downarrow \tau_{t(e)} \\ \hat{\mathbf{A}}_{s(e)} & \xrightarrow{\hat{\mathbf{A}}_e} & \hat{\mathbf{A}}_{t(e)} \end{array}$$

commutes for all $e \in Q_1$. We will also require a looser notion of map between quiver representations: If we have a collection of linear maps $\tau = \left\{ \tau_v : \mathbf{A}_v \rightarrow \hat{\mathbf{A}}_v \right\}$ without the commuting condition, we will simply call this a *transformation* of quiver representations. We will consider two $\mathbf{FdHilb}$ -valued representations of Q to be equivalent if they are *isometrically isomorphic* i.e. there exists a morphism τ such that each component τ_v is unitary (if $\mathbb{k} = \mathbb{C}$) or orthogonal (if $\mathbb{k} = \mathbb{R}$).

3.1.1 Global Sections

Given a representation $\mathbf{A}_\bullet$ of a quiver Q , its *total space* is defined as

$$\text{Tot}(\mathbf{A}_\bullet) := \bigoplus_{v \in Q_0} \mathbf{A}_v,$$

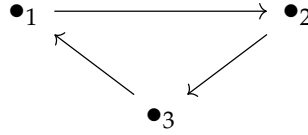
where we give $\text{Tot}(\mathbf{A}_\bullet)$ the induced inner product. One perspective of a quiver representation is that the edges encode constraints on the total space.

Definition 3.1.3 ([73]). A *global section* of $\mathbf{A}_\bullet$ is a vector $x \in \text{Tot}(\mathbf{A}_\bullet)$ that is consistent with $\mathbf{A}_\bullet$: for every edge $e \in Q_1$ we have

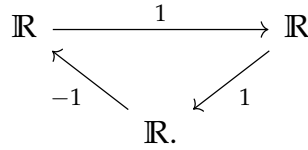
$$\mathbf{A}_e x_{s(e)} = x_{t(e)}$$

where x_v denotes the component of x in $\mathbf{A}_v$. The subspace of all global sections is denoted $\Gamma(\mathbf{A}_\bullet)$.

The space of global sections can be thought of as a generalisation of the connected components of a graph. Given a graph G , consider the unital representation $\mathbf{1}_\bullet$ of G , where we give the edges of G an orientation. Then the space of sections $\Gamma(\mathbf{1}_\bullet)$ is isomorphic to $\mathbb{k}^{\beta_0}$ where β_0 is the number of (weakly³) connected components of G . However, unlike the connected components of a graph, the dimension of the space of global sections may be zero, as can be seen in the quiver



and the following representation



which admits no global sections.

In Chapter 4, we will want to compute the space of global sections. There are different methods to do so, e.g. [73], however, we are particularly interested in quiver representations arising in data. Generically, the space of sections will be trivial, and as data is often noisy, we should not expect to be able to recover the

³The connected components of the underlying undirected graph.

space of sections accurately. Instead, we should aim to find *approximate global sections*. First, let us introduce the quiver Laplacian.

3.2 The Quiver Laplacian

Here we fix a finite quiver $Q = (s, t : Q_1 \rightarrow Q_0)$ and consider a representation $\mathbf{A}_\bullet$ of Q valued in $\mathbf{FdHilb}(\mathbb{k})$. First we define *target space* of $\mathbf{A}_\bullet$ as the direct sum

$$\mathrm{Tar}(\mathbf{A}_\bullet) := \bigoplus_{e \in Q_1} \mathbf{A}_{t(e)}.$$

Since we have assumed that Q is finite, both the total and target spaces inherit a Hilbert space structure from the individual factors of the form $\mathbf{A}_v$. We will denote the identity map on $\mathbf{A}_v$ by Id_v rather than $\mathrm{Id}_{\mathbf{A}_v}$.

Definition 3.2.1. Given an $\mathbf{FdHilb}(\mathbb{k})$ -valued quiver representation $\mathbf{A}_\bullet$ of Q , its **boundary operator**

$$B_{\mathbf{A}} : \mathrm{Tot}(\mathbf{A}_\bullet) \rightarrow \mathrm{Tar}(\mathbf{A}_\bullet)$$

is the linear map given in component form by

$$(B_{\mathbf{A}})_{e,v} = \begin{cases} \mathbf{A}_e - \mathrm{Id}_v & \text{if } s(e) = t(e) = v, \\ \mathbf{A}_e & \text{if } s(e) = v \text{ and } t(e) \neq v, \\ -\mathrm{Id}_v & \text{if } s(e) \neq v \text{ and } t(e) = v, \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

If the quiver Q is finite, then it follows from a simple calculation that $\ker B_{\mathbf{A}}$ is isomorphic (as a Hilbert space) to the space $\Gamma(Q; \mathbf{A}_\bullet)$ of $\mathbf{A}_\bullet$'s sections. From the boundary operator of a quiver representation, we can construct a new associated operator, the Laplacian, which will allow us to compute approximate sections.

Definition 3.2.2. The **Laplacian** of $\mathbf{A}_\bullet$ is the endomorphism $L_{\mathbf{A}} : \text{Tot}(\mathbf{A}_\bullet) \rightarrow \text{Tot}(\mathbf{A}_\bullet)$ given by

$$L_{\mathbf{A}} = B_{\mathbf{A}}^* B_{\mathbf{A}}.$$

In component form, we have

$$(L_{\mathbf{A}})_{v,v} = \sum_{\substack{e \in Q_1 \\ s(e)=v}} \mathbf{A}_e^* \mathbf{A}_e + \sum_{\substack{e \in Q_1 \\ t(e)=v}} \text{Id}_v - \sum_{\substack{e \in Q_1 \\ s(e)=t(e)=v}} [\mathbf{A}_e^* + \mathbf{A}_e]$$

and for $u \neq v$

$$(L_{\mathbf{A}})_{u,v} = - \sum_{\substack{e \in Q_1 \\ s(e)=u \\ t(e)=v}} \mathbf{A}_e^* - \sum_{\substack{e \in Q_1 \\ s(e)=v \\ t(e)=u}} \mathbf{A}_e.$$

Since $\ker L_{\mathbf{A}} = \ker B_{\mathbf{A}}$, the kernel of $L_{\mathbf{A}}$ also computes the sections of $\mathbf{A}_\bullet$. As with the graph Laplacian, the quiver Laplacian is a Hermitian and positive semi-definite matrix. As such, it enjoys favourable spectral properties; in particular, all of its eigenvalues are all real and non-negative. We shall order these eigenvalues in increasing fashion, with multiplicity:

$$\lambda_1(L_{\mathbf{A}}) \leq \dots \leq \lambda_n(L_{\mathbf{A}})$$

where $n := \dim \text{Tot}(\mathbf{A}_\bullet)$. Let $\tau : \mathbf{A}_\bullet \rightarrow \widehat{\mathbf{A}}_\bullet$ be an isometric isomorphism of quiver representations. Define $\tau_{\text{Tot}} = \bigoplus_{v \in Q_0} \tau_v$ and $\tau_{\text{Tar}} = \bigoplus_{e \in Q_1} \tau_{t(e)}$. Then

$$L_{\mathbf{A}} = B_{\mathbf{A}}^* B_{\mathbf{A}} = \tau_{\text{Tot}}^* B_{\widehat{\mathbf{A}}}^* \tau_{\text{Tar}} \tau_{\text{Tar}}^* B_{\widehat{\mathbf{A}}} \tau_{\text{Tot}} = \tau_{\text{Tot}}^* L_{\widehat{\mathbf{A}}} \tau_{\text{Tot}}$$

where τ_{Tot} is orthogonal (resp. unitary), thus we may speak unambiguously of the eigenvalues $\lambda_i(\mathbf{A}_\bullet) := \lambda_i(L_{\mathbf{A}})$ of a quiver representation $\mathbf{A}_\bullet$.

3.3 Approximate Sections

To measure the distance in $\text{Tot}(\mathbf{A}_\bullet)$ between an arbitrary vector x and the space $\Gamma(\mathbf{A}_\bullet)$ of sections, we consider the *Dirichlet energy*:

$$\mathcal{E}_\mathbf{A}(x) := \langle x, L_\mathbf{A} x \rangle = \sum_{e \in Q_1} \|\mathbf{A}_e(x_{s(e)}) - x_{t(e)}\|_{\mathbf{A}_{t(e)}}^2 \in \mathbb{R}_{\geq 0}$$

Here $\langle -, - \rangle$ is the induced inner product on $\text{Tot}(\mathbf{A}_\bullet)$, while x_v denotes the component of x in $\mathbf{A}_v$, and $\| - \|$ denotes the norm induced by the inner product on $\text{Tot}(\mathbf{A}_\bullet)$. As $L_\mathbf{A}$ is Hermitian, we can form an orthonormal eigenbasis $e_1, \dots, e_n$ of $\text{Tot}(\mathbf{A}_\bullet)$ where e_i corresponds to $\lambda_i(\mathbf{A}_\bullet)$. Writing x in this basis, we have

$$\mathcal{E}_\mathbf{A}(x) = \mathcal{E}_\mathbf{A} \left(\sum_i x_i e_i \right) = \sum_i \lambda_i(\mathbf{A}_\bullet) x_i^2.$$

Thus, $\sqrt{\mathcal{E}_\mathbf{A}(x)}$ is the λ_i -weighted distance from x to $\Gamma(\mathbf{A}_\bullet)$; and moreover,

$$\sum_{i: \lambda_i(\mathbf{A}_\bullet) = 0} x_i e_i$$

is the closest vector to x in $\Gamma(\mathbf{A}_\bullet)$. This motivates the following definition. (A similar notion for cellular sheaves was introduced in [44]).

Definition 3.3.1. For a quiver Q and an $\mathbf{FdHilb}(\mathbb{k})$ -valued representation $\mathbf{A}_\bullet$ of Q , an ε -approximate section of $\mathbf{A}_\bullet$ is an $x \in \text{Tot}(\mathbf{A}_\bullet)$ such that $\|x\| = 1$ and $\mathcal{E}_\mathbf{A}(x) \leq \varepsilon$.

By Theorem A.3.1, if $x \in \text{span}\{e_1, \dots, e_k\}$ and $\|x\| = 1$, then x is a λ_k -approximate section. Note this is not exhaustive: there can exist λ_k -approximate sections of $\mathbf{A}_\bullet$ that are not in the span of the first k eigenvectors, e.g. $\lambda_1 x_1 e_1 + \lambda_{k+1} x_{k+1} e_{k+1}$ for small enough x_{k+1} . Regardless, to compute approximate sections of a quiver representation one can compute eigenvectors of $L_\mathbf{A}$ with small eigenvalues. As we have connected approximate sections with the spectrum of the quiver represent-

ation, we now want to understand how the spectrum changes if we perturb the quiver representation (§3.4), or the underlying quiver (§3.5).

3.4 Spectral Stability of Quiver Representations

Let Q be a quiver and suppose $\mathbf{A}_\bullet$ and $\widehat{\mathbf{A}}_\bullet$ are two representations of Q . Suppose τ is a transformation $\{(\tau_v : \mathbf{A}_v \rightarrow \widehat{\mathbf{A}}_v) : v \in Q_0\}$ between $\mathbf{A}_\bullet$ and $\widehat{\mathbf{A}}_\bullet$. Define the *defect* of τ at an edge $e \in Q_1$ to be

$$\partial(\tau, e) = \left\| \widehat{\mathbf{A}}_e \tau_{s(e)} - \tau_{t(e)} \mathbf{A}_e \right\|,$$

Here, and henceforth, the expression $\| - \|$ when applied to a linear map between Hilbert spaces will always indicate the operator norm — for $M : V \rightarrow W$, the norm $\|M\|$ equals the supremum of $\|Mx\|_W$ over unit vectors $x \in V$. Now the *total defect* of τ is:

$$\partial(\tau) = \left(\sum_{e \in Q_1} \partial(\tau, e)^2 \right)^{\frac{1}{2}}.$$

The defect of τ is zero if and only if when τ prescribes a morphism of quiver representations. Given bases of $\text{Tot}(\mathbf{A}_\bullet)$ and $\text{Tot}(\widehat{\mathbf{A}}_\bullet)$, we view τ as a block diagonal matrix. Let τ^+ denote its Moore-Penrose pseudoinverse. The values $\|\tau\|$ and $\|\tau^+\|$ are unique up to unitary transforms of $\text{Tot}(\mathbf{A}_\bullet)$ and $\text{Tot}(\widehat{\mathbf{A}}_\bullet)$. Finally, we let

$$\kappa(\tau) := \|\tau\| \|\tau^+\|.$$

be the *generalised condition number* of τ .

We seek to describe how the existence of τ forces a relationship between the Laplacian spectra of $\mathbf{A}_\bullet$ and $\widehat{\mathbf{A}}_\bullet$. The first step in this direction is the following lemma.

Lemma 3.4.1. *For a matrix τ , if $x \in (\ker \tau)^\perp$ then*

$$\|\tau x\| \geq \|\tau^+\|^{-1} \|x\|.$$

Proof. Let $u_1, \dots, u_m$ and $v_1, \dots, v_n$ be left and right singular vectors of τ respectively. Let $x = \sum_{i=1}^n x_i v_i$. Then $\tau x = \sum_{i=1}^{\min\{n,m\}} \sigma_i x_i u_i$, but as $x \in \ker \tau^\perp$ we have that $x_j = 0$ when $\sigma_j = 0$ or $j > \min\{m, n\}$. Thus

$$\|\tau x\| \geq \left\| \sum_{i=1}^{\min\{n,m\}} \sigma_i x_i u_i \right\| = \sigma_+ \|x\| = \|\tau^+\|^{-1} \|x\|$$

where σ_+ is the smallest non-zero singular value of τ . □

The next theorem describes how the existence of a transformation between quiver representations constrains their eigenvalues. In the statement below (and henceforth), we use $\text{null}(f)$ as a shorthand for the dimension of the kernel of a linear map $f : U \rightarrow V$ of finite-dimensional Hilbert spaces.

Theorem 3.4.2. *Suppose $\mathbf{A}_\bullet$ and $\widehat{\mathbf{A}}_\bullet$ are two representations of a quiver Q whose total spaces have dimensions n and m respectively. Let $\tau = \{\tau_v : \mathbf{A}_v \rightarrow \widehat{\mathbf{A}}_v \mid v \in Q_0\}$ be a collection of linear maps, viewed as a single map $\text{Tot}(\mathbf{A}_\bullet) \rightarrow \text{Tot}(\widehat{\mathbf{A}}_\bullet)$. Set $q := \text{null}(\tau)$; for every integer k such that $1 \leq k \leq n - q$, we have*

$$\lambda_k(\widehat{\mathbf{A}}) \leq \kappa(\tau)^2 \lambda_{k+q}(\mathbf{A}) + \partial(\tau) \|\tau^+\| \left[2\kappa(\tau) \lambda_{k+q}(\mathbf{A})^{\frac{1}{2}} + \partial(\tau) \|\tau^+\| \right].$$

In particular, if τ is a morphism of quiver representations, then $\partial(\tau)$ vanishes and hence

$$\lambda_k(\widehat{\mathbf{A}}) \leq \kappa(\tau)^2 \lambda_{k+q}(\mathbf{A}).$$

Proof. Let $x_1, \dots, x_n$ and $y_1, \dots, y_m$ be eigenvectors of $\mathbf{A}_\bullet$ and $\widehat{\mathbf{A}}_\bullet$ respectively, ordered by increasing eigenvalue. Suppose k is an integer such that $1 \leq k \leq n - q$. Let

$l = k + q$ and define $X_l = \text{span}\{x_1, \dots, x_l\}$ and $Y^k = \text{span}\{y_k, \dots, y_m\}$. As $l = k + q$ we have that $\dim \tau(X_l) \geq k$ thus $\dim \tau(X_l) + \dim Y^k$ is at least $m + 1$ hence $\tau(X_l)$ and Y^k intersect non-trivially. Thus there exists some non-zero $y \in \tau(X_l) \cap Y^k$, and as $y \in Y^k$, by Theorem A.3.1 it holds that

$$\lambda_k(\hat{\mathbf{A}}) \leq \frac{1}{\|y\|^2} \sum_{e \in Q_1} \left\| \hat{\mathbf{A}}_e y_{s(e)} - y_{t(e)} \right\|^2.$$

Now we shall bound the right-hand side of the above. Since $y \in \tau(X_l)$ there exists some $x \in X_l \cap (\ker \tau)^\perp$ such that $\tau x = y$. Using Lemma 3.4.1 have that

$$\frac{1}{\|\tau x\|^2} \sum_{e \in Q_1} \left\| \hat{\mathbf{A}}_e \tau_{s(e)} x_{s(e)} - \tau_{t(e)} x_{t(e)} \right\|^2 \leq \frac{\|\tau^+\|^2}{\|x\|^2} \sum_{e \in Q_1} \left\| \hat{\mathbf{A}}_e \tau_{s(e)} x_{s(e)} - \tau_{t(e)} x_{t(e)} \right\|^2.$$

Then applying the triangle inequality gives us

$$\begin{aligned} & \frac{1}{\|\tau x\|^2} \sum_{e \in Q_1} \left\| \hat{\mathbf{A}}_e \tau_{s(e)} x_{s(e)} - \tau_{t(e)} x_{t(e)} \right\|^2 \\ & \leq \frac{\|\tau^+\|^2}{\|x\|^2} \sum_{e \in Q_1} \left\| \tau_{t(e)} \mathbf{A}_e x_{s(e)} - \tau_{t(e)} x_{t(e)} \right\|^2 \\ & \quad + 2 \frac{\|\tau^+\|^2}{\|x\|^2} \sum_{e \in Q_1} \left\| \tau_{t(e)} \mathbf{A}_e x_{s(e)} - \tau_{t(e)} x_{t(e)} \right\| \left\| \hat{\mathbf{A}}_e \tau_{s(e)} x_{s(e)} - \tau_{t(e)} \mathbf{A}_e x_{s(e)} \right\| \\ & \quad + \frac{\|\tau^+\|^2}{\|x\|^2} \sum_{e \in Q_1} \left\| \hat{\mathbf{A}}_e \tau_{s(e)} x_{s(e)} - \tau_{t(e)} \mathbf{A}_e x_{s(e)} \right\|^2 \end{aligned}$$

and then with the Cauchy-Schwarz inequality we have

$$\begin{aligned} & \frac{1}{\|\tau x\|^2} \sum_{e \in Q_1} \left\| \hat{\mathbf{A}}_e \tau_{s(e)} x_{s(e)} - \tau_{t(e)} x_{t(e)} \right\|^2 \\ & \leq \kappa(\tau)^2 \frac{1}{\|x\|^2} \sum_{e \in Q_1} \left\| \mathbf{A}_e x_{s(e)} - \tau_{t(e)} x_{t(e)} \right\|^2 \\ & \quad + \kappa(\tau) \|\tau^+\| \left[\frac{1}{\|x\|^2} \sum_{e \in Q_1} \left\| \mathbf{A}_e x_{s(e)} - \tau_{t(e)} x_{t(e)} \right\|^2 \right]^{\frac{1}{2}} \partial(\tau) + \|\tau^+\|^2 \partial(\tau)^2 \end{aligned}$$

and finally since $x \in X_l$, we may apply Theorem A.3.1 to obtain

$$\begin{aligned} & \frac{1}{\|\tau x\|^2} \sum_{e \in Q_1} \left\| \hat{\mathbf{A}}_e \tau_{s(e)} x_{s(e)} - \tau_{t(e)} x_{t(e)} \right\|^2 \\ & \leq \kappa(\tau)^2 \lambda_l(\mathbf{A}) + 2\kappa(\tau) \|\tau^+\| \lambda_l(\mathbf{A})^{\frac{1}{2}} \partial(\tau) + \|\tau^+\|^2 \partial(\tau)^2 \end{aligned}$$

completing the proof. $\square$

We shall now describe how Theorem 3.4.2 gives us a bound on the spectral pseudo-metric between quiver representations. If the total spaces of $\mathbf{A}$ and $\hat{\mathbf{A}}_\bullet$ have the same dimension, then we can directly compare the associated Laplacian spectra. Following [33], we use the Wasserstein metric to perform such comparisons in order to account for the case where total spaces have different dimensions.

Definition 3.4.3. For each real number $r \geq 0$, let $B(Q, r)$ be the set of representations $\mathbf{A}_\bullet$ of Q which satisfy $\max_{e \in Q_1} \|A_e\| \leq r$.

By design, if $\mathbf{A}_\bullet \in B(Q, r)$, then its eigenvalues are bounded by $|Q_1|(r+1)^2$.

Given a representation $\mathbf{A}_\bullet$ of Q with total dimension n , its *spectral measure* is

$$\mu_{\mathbf{A}} = \sum_{i=1}^n \delta_{\lambda_i(\mathbf{A})}$$

where $\delta_{\lambda_i(\mathbf{A})}$ is the Dirac measure concentrated at $\lambda_i(\mathbf{A})$. Given two probability measures μ_1, μ_2 on $\mathbb{R}_{\geq 0}$, a *coupling* γ of μ_1 and μ_2 is any probability measure on $\mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$ whose marginals on each factor are μ_1 and μ_2 respectively. Denote by $\Gamma(\mu_1, \mu_2)$ the set of all such couplings. Then the Wasserstein p -metric between μ_1 and μ_2 is defined as

$$W_p(\mu_1, \mu_2) = \inf_{\gamma \in \Gamma(\mu_1, \mu_2)} \left(\int_{\mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}} \|x - y\|^p d\gamma(x, y) \right)^{\frac{1}{p}}$$

Adapting the proof of Theorem 6.3 in [33] we have the following:

Theorem 3.4.4. Suppose $\mathbf{A}_\bullet, \hat{\mathbf{A}}_\bullet \in B(Q, r)$ with total dimensions n and m respectively, where $n \geq m$. Suppose τ is a non-zero transformation from $\mathbf{A}_\bullet$ to $\hat{\mathbf{A}}_\bullet$. If $\text{rank}(\eta) \geq \text{null}(\eta^*)$ then

$$W_1(\mu_{\mathbf{A}}, \mu_{\hat{\mathbf{A}}}) \leq C$$

where C depends on τ, n and r . Moreover, if $\|\tau\|, \|\tau^+\| \leq 1 + \varepsilon$ and $\partial(\tau), \partial(\tau^*) \leq \varepsilon$ for some $\varepsilon > 0$, then we have

$$\lim_{\varepsilon \rightarrow 0} C = \frac{2 \text{null}(\tau) + 2 \text{null}(\tau^*)}{n} |Q_1| (r+1)^2,$$

which, if $\text{null}(\tau) + \text{null}(\tau^*) < n/2$, will be less than naïve bound of $|Q_1|(r+1)^2$.

Proof. Define $c_\tau = \text{null}(\tau) + \text{null}(\tau^*)$. We will now define a coupling γ between $\mu_{\mathbf{A}}$ and $\mu_{\hat{\mathbf{A}}}$. For each positive integer p , let $[p]$ denote the set $\{1, 2, \dots, p-1, p\}$. Setting $\ell = \text{null}(\tau^*)$ and $u = n - \text{null}(\tau)$, define the following partition of $[n] \times [m]$:

$$\begin{aligned} P_1 &= \{(i, i) : i \in [u] \setminus [\ell]\} \\ P_2 &= \{(i, j) : i \in [\ell] \text{ and } j \in [\ell]\} \\ &\quad \cup \{(i, j) : i \in [\ell] \text{ and } j \in [m] \setminus [u]\} \\ &\quad \cup \{(i, j) : i \in [n] \setminus [u] \text{ and } j \in [\ell]\} \\ &\quad \cup \{(i, j) : i \in [n] \setminus [u] \text{ and } j \in [m] \setminus [u]\} \\ P_3 &= \{(i, j) : i \in [\ell] \text{ and } j \in [u] \setminus [\ell]\} \\ &\quad \cup \{(i, j) : i \in [n] \setminus [u] \text{ and } j \in [u] \setminus [\ell]\} \\ P_4 &= ([n] \times [m]) \setminus (P_1 \cup P_2 \cup P_3). \end{aligned}$$

Now define the desired coupling γ as

$$\gamma = \sum_{\substack{1 \leq i \leq n \\ 1 \leq j \leq m}} w_{ij} \delta_{(\lambda_i(\mathbf{A}), \lambda_j(\hat{\mathbf{A}}))} \quad \text{where} \quad w_{ij} = \begin{cases} \frac{1}{n} & (i, j) \in P_1, \\ \frac{1}{m(n-c_\tau)} & (i, j) \in P_2, \\ \left(\frac{1}{m} - \frac{1}{n}\right) \frac{1}{n-c_\tau} & (i, j) \in P_3, \\ 0 & (i, j) \in P_4. \end{cases}$$

Set $R = |Q_1|(r+1)^2$, which is an upper bound for the eigenvalues of $\mathbf{A}_\bullet$. Then using Theorem 3.4.2 for τ and τ^* and substituting R gives us

$$\kappa(\tau)^{-2} \lambda_{i-\text{null}(\tau^*)}(\mathbf{A}_\bullet) - a_{\tau^*, R} \leq \lambda_i(\hat{\mathbf{A}}_\bullet) \leq \kappa(\tau)^2 \lambda_{i+\text{null}(\tau)}(\mathbf{A}_\bullet) + b_{\tau, R}$$

with the constants

$$a_{\tau^*, R} = 2\kappa(\tau)^{-1} \|\tau^+\| \partial(\tau^*) \sqrt{R} + \kappa(\tau)^{-2} \|\tau^+\|^2 \partial(\tau^*)^2, \text{ and} \\ b_{\tau, R} = 2\kappa(\tau) \|\tau^+\| \partial(\tau) \sqrt{R} + \|\tau^+\|^2 \partial(\tau)^2.$$

Computing cost of γ gives us

$$W_1(\mu_{\mathbf{A}}, \mu_{\hat{\mathbf{A}}}) \leq \frac{1}{n} \left[\sum_{k=\text{null}(\tau^*)+1}^{n-\text{null}(\tau)} |\lambda_k(\mathbf{A}_\bullet) - \lambda_k(\hat{\mathbf{A}}_\bullet)| + c_\tau R \right] \\ \leq \frac{1}{n} \left[\sum_{k=\text{null}(\tau^*)+1}^{n-\text{null}(\tau)} (\kappa(\tau)^2 \lambda_{k+\text{null}(\tau)}(\mathbf{A}_\bullet) - \kappa(\tau)^{-2} \lambda_{k-\text{null}(\tau^*)}(\mathbf{A}_\bullet)) \right. \\ \left. + (n - c_\tau)(a_{\tau^*, R} + b_{\tau, R}) + c_\tau R \right].$$

Rearranging the sum on the right side yields

$$\frac{1}{n} \left[c_\tau \kappa(\tau)^2 R + (n - 2c_\tau)(\kappa(\tau)^2 - \kappa(\tau)^{-2})R + (n - c_\tau)(a_{\tau^*, R} + b_{\tau, R}) + c_\tau R \right],$$

which is the desired C . □

3.4.1 Stability of Subspace Arrangements

Suppose A, B are a pair of orthogonal matrices with dimensions $n \times k$ and $n \times l$ respectively. Then the matrix A^*B is an orthogonal projection $\text{im } B \rightarrow \text{im } A$, where $\text{im } A$ denotes the image of A . Let the singular values of A^*B be $\sigma_1, \dots, \sigma_r$ where $r = \min\{k, l\}$. Then the principal angles between $\text{im } A$ and $\text{im } B$ are defined by $\cos \theta_i = \sigma_i$ and the Grassmannian metric between $\text{im } A$ and $\text{im } B$ as defined by [102] is

$$d_{\text{Gr}_\infty}(\text{im } A, \text{im } B) := \left(\sum_{i=1}^r \theta_i^2 \right)^{\frac{1}{2}}.$$

Corollary 3.4.5. *Let Q be a quiver and H an inner product space. Suppose $\mathbf{S}, \mathbf{T}$ are two subspace arrangements of Q in H . Suppose that*

$$\max_{v \in Q_0} d_{\text{Gr}_\infty}(\mathbf{S}_v, \mathbf{T}_v) \leq \varepsilon < \frac{\pi}{2}$$

for some $\varepsilon > 0$, and define

$$c = \sum_{v \in Q_0^{\mathcal{U}}} |\dim \mathbf{S}_v - \dim \mathbf{T}_v|.$$

Then, $W_1(\mu_{\mathbf{S}}, \mu_{\mathbf{T}})$ is no larger than

$$\frac{4|Q_1|}{n} \left[(1 + \cos^2 \varepsilon) c + \left(2 \sec \varepsilon \tan \varepsilon + (1 + \sec^2 \varepsilon) \tan^2 \varepsilon + \sec^2 \varepsilon - \cos^2 \varepsilon \right) (n - c) \right]$$

where $n = \max\{\dim_{\text{Tot}} \mathbf{S}_\bullet, \dim_{\text{Tot}} \mathbf{T}_\bullet\}$.

Proof. Assume that $\dim_{\text{Tot}} \mathbf{S}_\bullet \geq \dim_{\text{Tot}} \mathbf{T}_\bullet$. Define the transformation η with components $\eta_v = \pi_v^{\mathbf{T}} \iota_v^{\mathbf{S}}$ for each $v \in Q_0$. Then $\cos \varepsilon \leq \|\eta\| \leq 1$ and $1 \leq \|\eta^+\| \leq$

$(\cos \varepsilon)^{-1}$. In order to bound $\partial(\eta)$ note that

$$\|\iota_v^{\mathbf{S}} \pi_v^{\mathbf{S}} - \iota_v^{\mathbf{T}} \pi_v^{\mathbf{T}}\|_2 = \sin \theta_{v,\max} \leq \sin \varepsilon,$$

where $\theta_{v,\max}$ is the largest principal angle between $\mathbf{S}_{|v|}$ and $\mathbf{T}_{|v|}$. This is known as the projection metric [102] where the equality can be seen from Theorem 2.6.1 in [31]. This then implies that, for each edge e from u to v , we have $\partial(\eta, e) \leq 2 \sin \varepsilon$ thus $\partial(\eta) \leq 2\sqrt{|Q_1|} \sin \varepsilon$. Similarly, $\partial(\eta^*) \leq 2\sqrt{|Q_1|} \sin \varepsilon$. Applying Theorem 3.4.4 together with these bounds provide the result. $\square$

3.5 Eigenvalue Inequalities for Quiver Operations

So far, we have described our problem as computing sections of a quiver representation. In this section we seek to simplify the representation and the underlying quiver whilst preserving its space of sections and bounding the change in its eigenvalues. We will establish similar eigenvalue inequalities when the quiver itself is deformed via certain natural operations. The most convenient framework for extracting such inequalities is to upgrade the given quiver representation to a sheaf, as described below.

3.5.1 Sheaves over Quivers

Fix a quiver $Q = (s, t : Q_1 \rightarrow Q_0)$.

Definition 3.5.1. An $\mathbf{FdHilb}(\mathbb{k})$ -valued **sheaf** $\mathcal{A}_\bullet$ over Q consists of:

- A finite-dimensional Hilbert space $\mathcal{A}_v$ for each vertex $v \in Q_0$,
- a finite-dimensional Hilbert space $\mathcal{A}_e$ for each edge $e \in Q_1$, and

- for each edge $e \in Q_1$, a pair of linear maps

$$\mathcal{A}_{s(e) \rightarrow e} : \mathcal{A}_{s(e)} \rightarrow \mathcal{A}_e \quad \text{and} \quad \mathcal{A}_{e \leftarrow t(e)} : \mathcal{A}_{t(e)} \rightarrow \mathcal{A}_e.$$

Such sheaves are closely related to representations of quivers. Every representation $\mathbf{A}_\bullet$ over Q induces a sheaf $\mathcal{A}_\bullet$ over Q by setting $\mathcal{A}_{s(e) \rightarrow e} = \mathbf{A}_e$ and $\mathcal{A}_{e \leftarrow t(e)} = \text{Id}_{t(e)}$. Conversely, every sheaf evidently induces a representation on the subdivision of Q where every edge $e : u \rightarrow v$ is decomposed into the zigzag $u \rightarrow e \leftarrow v$. Thus, there is a natural notion of Laplacian for sheaves over a quiver, a special case⁴ of which appears in [37].

Definition 3.5.2. Suppose $\mathcal{A}_\bullet$ is a $\mathbf{FdHilb}(\mathbb{k})$ -valued sheaf over a quiver Q . Define $\text{Tot}(\mathcal{A}_\bullet) = \prod_{v \in Q_0} \mathcal{A}_v$ and $\text{Tar}(\mathcal{A}_\bullet) = \prod_{e \in Q_1} \mathcal{A}_e$.

1. The **boundary operator** of $\mathcal{A}_\bullet$ is the linear map $B_{\mathcal{A}} : \text{Tot}(\mathcal{A}_\bullet) \rightarrow \text{Tar}(\mathcal{A}_\bullet)$ of $\mathcal{A}_\bullet$ given in component form by

$$(B_{\mathcal{A}})_{e,v} = \begin{cases} \mathcal{A}_{v \rightarrow e} - \mathcal{A}_{e \leftarrow v} & \text{if } s(e) = t(e) = v, \\ \mathcal{A}_{v \rightarrow e} & \text{if } s(e) = v \text{ and } t(e) \neq v, \\ -\mathcal{A}_{e \leftarrow v} & \text{if } s(e) \neq v \text{ and } t(e) = v, \\ 0 & \text{otherwise.} \end{cases}$$

2. The **Laplacian** of $\mathcal{A}_\bullet$ is then defined as

$$L_{\mathcal{A}} = B_{\mathcal{A}}^* B_{\mathcal{A}}.$$

3. Finally, the **Dirichlet energy** for $\mathcal{A}$ is the function $\mathcal{E}_{\mathcal{A}} : \text{Tot}(\mathcal{A}_\bullet) \rightarrow \mathbb{R}$ given

⁴The only difference between our quiver sheaf Laplacian and the sheaf Laplacians of [37] over one-dimensional cell complexes is that we allow edges to have the same source and target vertex. On the other hand, cellular sheaves and their Laplacians are also defined over regular cell complexes of dimension > 1 .

by

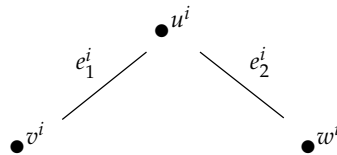
$$\mathcal{E}_{\mathcal{A}}(x) = \sum_{e \in Q_1} \left\| \mathcal{A}_{s(e) \rightarrow e}(x_{s(e)}) - \mathcal{A}_{e \leftarrow t(e)}(x_{t(e)}) \right\|_{\mathcal{A}_e}^2.$$

If one constructs a sheaf $\mathcal{A}_\bullet$ from a quiver representation $\mathbf{A}_\bullet$ of Q as described above, then the sheaf boundary operator $B_{\mathcal{A}}$ coincides exactly with the quiver boundary operator $B_{\mathbf{A}}$; therefore the Laplacians of $\mathbf{A}$ and $\mathcal{A}$, as well as their spectra, also coincide in this case.

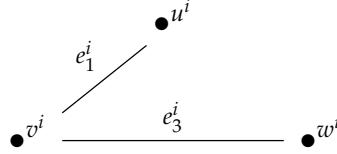
3.5.2 Operations on the Quiver

This section describes how the spectrum of a sheaf, or indeed a representation, over a quiver changes under various combinatorial operations on the quiver. Removing an edge or a vertex gives bounds on the change in spectrum, as Proposition A.1.1 describes, however these may reduce the dimension of the space of sections. Propositions 3.5.5, 3.5.7, and 3.5.9 describe operations on the quiver, dependent on the sheaf, that preserve the space of sections. These are described for sheaves over a quiver but all these operations also preserve quiver representations. These intermediate results are used to simplify quivers (while bounding the change in eigenvalues) in our proofs of Theorems 4.2.14 and 4.2.15.

Definition 3.5.3. For a quiver Q , define an *edge pairing* as a collection P of pairs of distinct neighbouring edges $(e_1^1, e_2^1), \dots, (e_1^m, e_2^m)$ of Q such no two edges are repeated, i.e. $e_j^i \neq e_l^k$ unless $i = k$ and $j = l$. To simplify notation assume that each pair (e_1^i, e_2^i) , the edges e_1^i and e_2^i are directed away from their common vertex u^i . Label the other vertices for e_1^i and e_2^i , possibly the same vertex, as v^i and w^i respectively.



We then define Q^P as the P -homotopic quiver to Q where Q^P has the same nodes and edges as Q except with each edge e_2^i replaced by an edge e_3^i between v^i and w^i .



For $\mathcal{A}_\bullet$ is a sheaf over Q and a pairing P of Q , say that $\mathcal{A}_\bullet$ is compatible with P if for each pair $(e_1^i, e_2^i) \in P$ we have that $\mathcal{A}_{e_1^i} = \mathcal{A}_{e_2^i}$ and $\mathcal{A}_{u^i \rightarrow e_1^i} = \mathcal{A}_{u^i \rightarrow e_2^i}$. Such a compatible sheaf induces a sheaf on Q^P as follows: define the sheaf $\mathcal{A}_\bullet^P$ on Q^P as the same as $\mathcal{A}_\bullet$ on the common edges of Q and Q^P and otherwise by $\mathcal{A}_{e_3^i}^P = \mathcal{A}_{e_2^i}$, $\mathcal{A}_{e_3^i \leftarrow e_3^i}^P = \mathcal{A}_{e_1^i \leftarrow e_1^i}$ and $\mathcal{A}_{w^i \rightarrow e_3^i}^P = \mathcal{A}_{e_2^i \leftarrow w^i}$ for each i . The next proposition shows the i -th eigenvalue of $\mathcal{A}_\bullet^P$ is bounded by a scalar multiple of the i -th eigenvalue of $\mathcal{A}_\bullet$. First we will need a lemma.

Lemma 3.5.4. Suppose $X, Y, Z \in \mathbb{R}^n$. Then

$$\|X - Z\|^2 + \|Z - Y\|^2 \leq \varphi^2(\|X - Y\|^2 + \|Y - Z\|^2)$$

where φ is the golden ratio $\varphi = \frac{1+\sqrt{5}}{2}$. This is also true for any metric space (M, d) and can be written as

$$d(X, Z)^2 \leq \varphi^2 d(X, Y)^2 + \varphi d(Y, Z)^2.$$

Proof. Let us abbreviate

$$x = \|Y - Z\| \quad y = \|X - Z\| \quad z = \|X - Y\|.$$

If any of x, y or z are zero, then the proposition is immediately true, so assume that

x, y and z are strictly positive. By the triangle inequality, we have

$$\frac{x^2 + y^2}{x^2 + z^2} \leq \frac{x^2 + (x + z)^2}{x^2 + z^2}.$$

We then want to know when the expression on the right side is maximised. As $x, z > 0$, there exists $u > 0$ for which $x = uz$. Substituting this into the right side gives

$$\frac{2u^2 + 2u + 1}{u^2 + 1},$$

which is maximised when $u = \varphi$. Thus, the maximum value of the ratio is φ^2 . If X, Y, Z are colinear with Y between X and Z and $\|Y - Z\| = \varphi\|X - Y\|$, then the maximum is attained, which shows that this inequality is sharp. $\square$

Proposition 3.5.5. *For Q a quiver and P an edge pairing of Q , suppose that $\mathbf{A}_\bullet$ is a sheaf on Q compatible with P . Then the eigenvalues of $\mathcal{A}_\bullet$ and $\mathcal{A}_\bullet^P$ are related by*

$$\varphi^{-2}\lambda_j(\mathcal{A}_\bullet) \leq \lambda_j(\mathcal{A}_\bullet^P) \leq \varphi^2\lambda_j(\mathcal{A}_\bullet)$$

and vice-versa for $j = 1, \dots, n$ where n is the total dimension of $\mathbf{A}_\bullet$.

Proof. Using Lemma 3.5.4 we have

$$\begin{aligned} \mathcal{E}_{\mathcal{A}}(x) &= \sum_{(e_1^i, e_2^i) \in P} \left[\left\| \mathcal{A}_{e_1^i \leftarrow v^i} x_{v^i} - \mathcal{A}_{u^i \rightarrow e_1^i} x_{u^i} \right\|_{\mathcal{A}_{e_1^i}}^2 + \left\| \mathcal{A}_{u^i \rightarrow e_2^i} x_{u^i} - \mathcal{A}_{e_2^i \leftarrow w^i} x_{w^i} \right\|_{\mathcal{A}_{e_2^i}}^2 \right] \\ &\quad + \sum_{\text{unpaired } e \in Q_1} \left\| \mathcal{A}_{s(e) \rightarrow e} x_{s(e)} - \mathcal{A}_{e \leftarrow t(e)} x_{t(e)} \right\|_{\mathcal{A}_{s(e)}}^2 \\ &\leq \varphi^2 \sum_{(e_1^i, e_2^i) \in P} \left[\left\| \mathcal{A}_{u^i \rightarrow e_1^i} x_{u^i} - \mathcal{A}_{e_1^i \leftarrow v^i} x_{v^i} \right\|_{\mathcal{A}_{e_1^i}}^2 + \left\| \mathcal{A}_{e_3^i \leftarrow v^i}^P x_{v^i} - \mathcal{A}_{w^i \rightarrow e_3^i}^P x_{w^i} \right\|_{\mathcal{A}_{e_3^i}}^2 \right] \\ &\quad + \sum_{\text{unpaired } e \in Q_1} \left\| \mathcal{A}_{s(e) \rightarrow e} x_{s(e)} - \mathcal{A}_{e \leftarrow t(e)} x_{t(e)} \right\|_{\mathcal{A}_{s(e)}}^2 \\ &\leq \varphi^2 \mathcal{E}_{\mathcal{A}^P}(x). \end{aligned}$$

Then by Theorem A.3.2

$$\lambda_i(\mathcal{A}_\bullet) = \min_{\dim V=i} \left[\max_{x \in V \setminus 0} \frac{\mathcal{E}_{\mathcal{A}}(x)}{\langle x, x \rangle} \right] \leq \min_{\dim V=i} \left[\max_{x \in V \setminus 0} \frac{\varphi^2 \mathcal{E}_{\mathcal{A}^P}(x)}{\langle x, x \rangle} \right] = \varphi^2 \lambda_i(\mathcal{A}_\bullet^P).$$

proving the proposition. $\square$

We note that Proposition 3.5.5 directly applies to graphs. Although we make no use of the following corollary, we hope that it will be of independent interest to spectral graph theorists.

Corollary 3.5.6. *Suppose $G = (V, E)$ is a multigraph with n vertices. Suppose we have an edge pairing $P = (e_1^1, e_1^1), \dots, (e_1^m, e_2^m)$ on G as in Definition 3.5.3. Let $\widehat{G}$ with the same nodes and edges as G except with each edge e_2^i replaced by an edge e_3^i between v^i and w^i . Then the eigenvalues of $L_{\widehat{G}}$ and L_G are related by*

$$\varphi^{-2} \lambda_j(L_G) \leq \lambda_j(L_{\widehat{G}}) \leq \varphi^2 \lambda_j(L_G)$$

and vice-versa for $j = 1, \dots, n$.

Proposition 3.5.7. *Suppose $\mathcal{A}_\bullet$ is a sheaf over a quiver Q with total dimension n . Given a subquiver Q' , let $\mathcal{A}'_\bullet$ denote the restriction of $\mathcal{A}_\bullet$ to Q' .*

3.5.7(a) *If $e \in Q_1$ is an edge such that $s(e) = t(e)$ and $\mathcal{A}_{s(e)} = \mathcal{A}_{t(e)} = \text{Id}$ then $\lambda_i(\mathcal{A}'_\bullet) = \lambda_i(\mathcal{A}_\bullet)$ for $i = 1, \dots, n$ where $\mathcal{A}'_\bullet$ is $\mathcal{A}_\bullet$ restricted to the subquiver Q' of Q with the edge e removed.*

3.5.7(b) *Suppose we have an edge pairing $P = (e_1^1, e_2^1), \dots, (e_1^m, e_2^m)$ of Q such that for each i , the edges e_1^i, e_2^i have the same source and target vertices. Assume that for each i , there exists a linear map $M_i : \mathcal{A}_{e_2^i} \rightarrow \mathcal{A}_{e_1^i}$ satisfying both*

$$\mathcal{A}_{s(e_1^i) \rightarrow e_1^i} = M_i \mathcal{A}_{s(e_2^i) \rightarrow e_2^i} \quad \text{and} \quad \mathcal{A}_{e_1^i \leftarrow t(e_1^i)} = M_i \mathcal{A}_{e_2^i \leftarrow t(e_2^i)}$$

Let Q' be the subquiver of Q with all the edges $e_2^1, \dots, e_2^n$ removed, and let $\mathcal{A}'_\bullet$ be the restriction of $\mathcal{A}_\bullet$ to Q' . Then, we have

$$(1 + \Gamma_P)^{-1} \lambda_i(\mathcal{A}_\bullet) \leq \lambda_i(\mathcal{A}'_\bullet) \leq \lambda_i(\mathcal{A}_\bullet)$$

for all $i = 1, \dots, n$, where $\Gamma_P := \max_i \{\|M_i\|^2\}$.

Proof. For 3.5.7(a), observe that $\mathcal{E}_{\mathcal{A}}(x) = \mathcal{E}_{\mathcal{A}'}(x)$ and thus the result follows from Theorem A.3.2. For 3.5.7(b), we have

$$\begin{aligned} \mathcal{E}_{\mathcal{A}}(x) &\leq \sum_{(e_1^i, e_2^i) \in P} (1 + \|M_i\|^2) \left\| \mathcal{A}_{s(e_1^i) \rightarrow e_1^i} x_{s(e_1^i)} - \mathcal{A}_{e_1^i \leftarrow t(e_1^i)} x_{t(e_1^i)} \right\|^2 \\ &\quad + \sum_{\text{unpaired } e \in Q_1} \left\| \mathcal{A}_{s(e) \rightarrow e} x_{s(e)} - \mathcal{A}_{e \leftarrow t(e)} x_{t(e)} \right\|_{\mathcal{A}_{s(e)}}^2 \\ &\leq (1 + \Gamma_P) \mathcal{E}_{\mathcal{A}'}(x). \end{aligned}$$

and the result follows from Theorem A.3.2. □

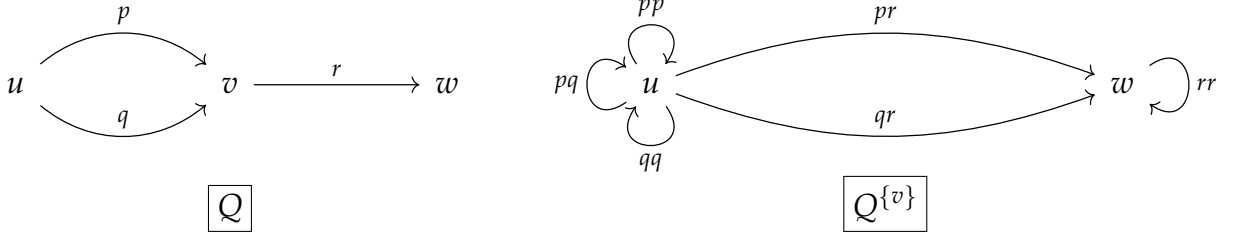
In the following, at the risk of causing ambiguity with self-loops, we will ignore the direction of edges in and use the relation $v \sim e$ if either $s(e) = v$ or $t(e) = v$.

Definition 3.5.8. Let Q be a quiver and let V be a subset of vertices such that there are no edges in Q between vertices of V . The **V-reduced quiver** Q^V is defined as follows. For each vertex $v \in V$,

1. first remove v and any associated edges from Q ;
2. then, for each pair of edges (p, q) in Q such that $u \sim p \sim v \sim q \sim w$ for vertices $u, w \in Q_0 \setminus V$, add a new edge pq between u and w .

(We note in the second step above that if $u = w$, then pq forms a self-loop.) The resulting quiver is the desired Q^V .

Figure 3.1: An example of the reduction of a quiver Q with $V = \{v\}$.



We now describe an analogue of the Kron reduction of a graph [20] for sheaves over a quiver. As discussed in [37], Kron reduction does not work in general for sheaves over a (loopless) graph. However, as the next proposition shows, if one restricts the type of vertex removed, dependent on the sheaf, and moreover allows for loops in the underlying quiver, then Kron reduction is possible.

Proposition 3.5.9. *Given a sheaf $\mathcal{A}_\bullet$ over a quiver Q , assume that $V \subset Q_0$ is a subset of vertices such that there are no edges in Q between vertices of V . Further suppose that for each $v \in V$ there exists some real number $\alpha_v > 0$ such that for each edge $e \sim v$, we have*

$$\text{either } \mathcal{A}_{v \rightarrow e}^* \mathcal{A}_{v \rightarrow e} = \alpha_v \text{Id}_v \quad \text{or} \quad \mathcal{A}_{e \leftarrow v}^* \mathcal{A}_{e \leftarrow v} = \alpha_v \text{Id}_v$$

as appropriate. Then there exists a sheaf $\mathcal{A}_\bullet^V$ over Q^V such that $\Gamma(\mathcal{A}_\bullet^V) \cong \Gamma(\mathcal{A}_\bullet)$ and

$$\lambda_i(\mathcal{A}_\bullet) \leq \lambda_i(\mathcal{A}_\bullet^V) \leq \lambda_{i+r}(\mathcal{A}_\bullet)$$

for $i = 1, \dots, n - r$ where n is the total dimension of $\mathcal{A}_\bullet$ and $r = \sum_{v \in V} \dim \mathcal{A}_v$.

Proof. Consider the Laplacian $L_{\mathcal{A}}$. It can be written in a block form as

$$L_{\mathcal{A}} = \begin{pmatrix} Q_0 \setminus V & V \\ X & Y^* \\ Y & D \end{pmatrix} \begin{matrix} Q_0 \setminus V \\ V \\ V \end{matrix}.$$

As there are no edges between vertices of V , D is a diagonal matrix where the (v, v) block is $d_v \alpha_v \text{Id}$ where d_v is the degree of v . Thus D is invertible, and we can form the Schur complement

$$L_{\mathcal{A}}/D := X - Y^* D^{-1} Y.$$

A vector $\begin{bmatrix} x & z \end{bmatrix}^T$ satisfies

$$\begin{bmatrix} X & Y^\dagger \\ Y & D \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

if and only if $Xx + Y^\dagger z = 0$ and $Yx + Dz = 0$. As D is invertible this can be rearranged to $Xx - YD^{-1}Y^\dagger x = 0$ hence $\ker(L_{\mathcal{A}}/D) \cong \ker L_A$. For the interlacing, we can directly apply Theorem A.3.5.

We will now define the sheaf $\mathcal{A}_\bullet^V$. As in Definition 3.5.8, we will ignore orientation of the edges and write $u \sim e$ if either $s(e) = v$ or $t(e) = v$ and $\mathcal{A}_{v \sim e}$ for the corresponding assigned linear map. If $p \neq q$, define $\mathcal{A}_{u \sim pq}^V = (d_v \alpha_v)^{-1/2} \mathcal{A}_{v \sim p}^* \mathcal{A}_{u \sim p}$ and likewise for w . Otherwise, if $p = q$, define $\mathcal{A}_{u \rightarrow pp}^V = Z_p \mathcal{A}_{u \sim p}$ and $\mathcal{A}_{pp \leftarrow u}^V = \mathbf{0}$ for a linear map Z_p which we will now define. First observe that for every non-zero vector x , we have

$$\frac{1}{\alpha_v} x^* \mathcal{A}_{v \sim p} \mathcal{A}_{v \sim p}^* x \leq \frac{1}{\alpha_v} \|\mathcal{A}_{v \sim p}^*\|^2 \|x\|^2 = \frac{1}{\alpha_v} \|\mathcal{A}_{v \sim p}^* \mathcal{A}_{v \sim p}\| \|x\|^2 = \|x\|^2$$

thus there exists some Z_p such that

$$\text{Id} - \frac{1}{\alpha_v} \mathcal{A}_{v \sim p} \mathcal{A}_{v \sim p}^* = Z_p^* Z_p.$$

since the left-hand side is positive semi-definite.

Now we claim that $L_{\mathcal{A}^V} = L_{\mathcal{A}}/D$. To do so, consider vertices $u \in Q_0 \setminus V$ and $v \in V$. For each edge $u \sim p \sim v$, the contribution to the diagonal block indexed by

(u, u) of $L_{\mathcal{A}}/D$ is

$$\mathcal{A}_{u \sim p}^* \mathcal{A}_{u \sim p} - \frac{1}{d_v \alpha_v} \sum_{u \sim q \sim v} \mathcal{A}_{u \sim p}^* \mathcal{A}_{v \sim p} \mathcal{A}_{v \sim q}^* \mathcal{A}_{u \sim q}.$$

We need to show this is equal to the contribution from the new edges of $\mathcal{A}_{\bullet}^Q$. For each vertex $u \in Q_0 \setminus V$, the contribution to the diagonal block is

$$\begin{aligned} & \sum_{v \in V} \sum_{u \sim p \sim v} \left[\mathcal{A}_{u \sim p}^* \mathcal{A}_{u \sim p} - \frac{1}{d_v \alpha_v} \sum_{u \sim q \sim v} \mathcal{A}_{u \sim p}^* \mathcal{A}_{v \sim p} \mathcal{A}_{v \sim q}^* \mathcal{A}_{u \sim q} \right] \\ &= \sum_{v \in V} \left[\sum_{u \sim p \sim v} \left[\mathcal{A}_{u \sim p}^* \mathcal{A}_{u \sim p} - \frac{d_v^u}{d_v \alpha_v} \mathcal{A}_{u \sim p}^* \mathcal{A}_{v \sim p} \mathcal{A}_{v \sim p}^* \mathcal{A}_{u \sim p} \right] \right. \\ & \quad \left. + \sum_{\substack{u \sim p \sim v \\ u \sim q \sim v \\ p \neq q}} \frac{1}{d_v \alpha_v} (\mathcal{A}_{v \sim p}^* \mathcal{A}_{u \sim p} - \mathcal{A}_{v \sim q}^* \mathcal{A}_{u \sim q})^* (\mathcal{A}_{v \sim p}^* \mathcal{A}_{u \sim p} - \mathcal{A}_{v \sim q}^* \mathcal{A}_{u \sim q}) \right] \end{aligned}$$

where d_v^u is the number of edges between u and v . The term on the last line corresponds to the new self-loops on u given by v . By the defining property of Z_p , the term

$$\sum_{u \sim p \sim v} \left[\mathcal{A}_{u \sim p}^* \mathcal{A}_{u \sim p} - \frac{d_v^u}{d_v \alpha_v} \mathcal{A}_{u \sim p}^* \mathcal{A}_{v \sim p} \mathcal{A}_{v \sim p}^* \mathcal{A}_{u \sim p} \right]$$

equals

$$\sum_{u \sim p \sim v} \left[\frac{d_v - d_v^u}{d_v \alpha_v} \mathcal{A}_{u \sim p}^* \mathcal{A}_{v \sim p} \mathcal{A}_{v \sim p}^* \mathcal{A}_{u \sim p} + \mathcal{A}_{u \sim p}^* Z_p^* Z_p \mathcal{A}_{u \sim p} \right]$$

and thus corresponds with the contribution of the self-loop pp and the new edges to different vertices, as required. $\square$

Chapter 4

Feature Selection and Compatibility

4.1 Introduction

Feature Selection

When working with data described by a large number of features, such as gene expression data, it is often desirable, even necessary, to reduce the number of features considered. We saw this in chapter 2: our aim was to reduce our set of features, the expressions of genes, to a smaller set that best described the biological variation. There are also practical reasons to reduce the dimensionality of the data: many computational and statistical procedures perform poorly in higher dimensions.

There are several ways of quantifying the relevance of a given feature — one may use domain-specific context, optimise a statistical quantity such as variance, or exploit the presence of labelled training data, and so on. This process of feature selection forms a cornerstone of data analysis. Examples of well-known feature selectors include principal components [61], LASSO [86] and recursive feature elimination [34], as well as the Laplacian score and its refinements from Chapter 1.

Data Divided

In many situations, we are provided with a natural grouping of the dataset, such as the biological cell types, or disease type, or the origin tissue of a biological sample. These groupings may overlap, i.e. not form a partition: as discussed in Chapter 2, cell types are not precisely defined.

If we are going to perform feature selection on such a dataset, ideally we will perform it on the group level. This is common practice in single-cell analysis (e.g. [71]), the primary reason being that if we only perform feature selection on the entire dataset, then features relevant for groups within the dataset, particularly rare groups, will not be best represented. For example, we may have identified a rare malignant cell type whose expression for some gene G is abnormally high; if this cell type is rare enough, the variance of the expression of G will not be high enough for it to be selected on the whole dataset.

Performing feature selection per group still leaves us with a problem. The goal of feature selection was to *reduce* the number of features, yet we now have a set of features per group. We want to be able to somehow combine features that are the same, or similar, across different groups, and compare how they relate to the topological structure of the underlying groups.

Consider the following toy example presented in situations I and II in Figure 4.1. In both situations we have two features, Feature 1 and Feature 2, defined on a dataset which has two overlapping groups, A and B . Now suppose we rank the relevance of features by their variance. In situation I, Feature 1 is most relevant for subsets A , B , and $A \cap B$, but would not be more relevant than Feature 2 on their union $A \cup B$. This is a similar phenomenon to Simpson's paradox in statistics, where splitting the data into certain groups can reverse a trend between two variables. Another form of inconsistency is exhibited in Figure 4.1.II where Feature 1 is most relevant for A , and Feature 2 for B , but both are equally relevant (or

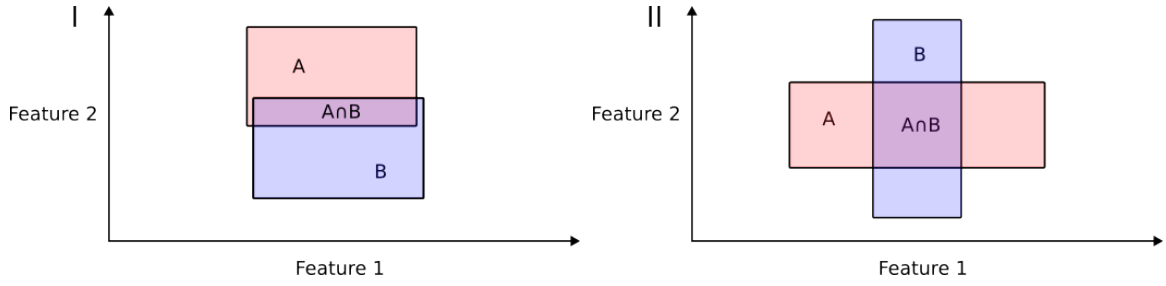


Figure 4.1: Two features defined on a dataset equipped with a decomposition into two overlapping subsets.

irrelevant) for $A \cap B$ and $A \cup B$. Both these examples illustrate that there is not a simple relationship between features relevant to two groups and their intersection or union.

For data with a grouping, any feature selection process (which remains agnostic to the decomposition) is liable to select features that are irrelevant to certain subsets while ignoring features that are relevant to certain subsets. This will not only effect which features we consider most relevant but will potentially bias downstream analysis.

In this chapter we will present a principled framework for selecting features on a finite set X that are consistent with respect to a decomposition of X into a *cover* $\mathcal{U}$. See Figure 1 on page 14 for an overview of the pipeline. In Chapter 5 we apply this methodology to the motivating dataset for this problem, single-cell chromatin accessibility.

4.2 Compatible Features

4.2.1 Overview

In this section, we present a principled framework for selecting features on a finite set X that are consistent with respect to a decomposition of X into a *cover* $\mathcal{U}$. A cover is a collection of subsets of X whose union is X . Crucially, the subsets

of X which comprise $\mathcal{U}$ are allowed to admit arbitrary overlaps. This is a reasonable, and perhaps necessary stipulation; in single-cell data analysis, as discussed in Chapter 2, cell types are not precisely defined. A more appropriate modelling approach is to allow some notion of continuity, as will be witnessed by the cover $\mathcal{U}$.

Given the immense diversity of available feature selectors and their inner workings, we adopt an abstract approach, treating the feature selector as a ‘black-box’. For our purposes, a feature selector $\mathcal{S}$ is any deterministic process that

1. accepts as input some finite collection $\mathcal{F}$ of functions (features), $X \rightarrow \mathbb{R}$, and
2. outputs a subspace $\mathcal{S}(\mathcal{F})$ of the $\mathbb{R}$ -linear span $\mathbf{V}(\mathcal{F})$ of $\mathcal{F}$.

Importantly, we will also assume that $\mathbf{V}(\mathcal{F})$ has a fixed inner product structure $\langle -, - \rangle$ that reflects a reasonable notion of similarity between features $X \rightarrow \mathbb{R}$.

For the purposes of these introductory remarks, let us call a nonempty subset $\sigma \subset \mathcal{U}$ *admissible* whenever the intersection $|\sigma| := \bigcap_{U \in \sigma} U$ is nonempty. Our quest here is to isolate the largest subspace of $\mathcal{S}(\mathcal{F})$ that is compatible with respect to the decomposition of X into admissible subsets. Two forms of compatibility are studied here: *local* and *global*. Given an admissible subset $\sigma \subset \mathcal{U}$, a relevant feature $v : |\sigma| \rightarrow \mathbb{R}$ is locally compatible with $\mathcal{U}$ if for every admissible $\tau \subset \mathcal{U}$ with $|\tau| \supset |\sigma|$, there is some relevant feature $v^+ : |\tau| \rightarrow \mathbb{R}$ whose restriction to $|\sigma|$ coincides with v . In other words, a locally compatible feature is one which continues to be selected as we pass to larger admissible subsets. Conversely, consider a collection of features $\{v_\sigma\}$, one for each admissible σ . Such a collection is globally compatible with $\mathcal{U}$ if for each pair $|\sigma| \subset |\tau|$ of admissible subsets, the orthogonal projection of features relevant to $|\sigma|$ onto $|\tau|$ maps v_σ onto v_τ . The question addressed in this chapter is:

Fix a feature selector $\mathcal{S}$ and a cover $\mathcal{U}$ of X . Given an input $\mathcal{F}$ family of functions $X \rightarrow \mathbb{R}$, to what extent can we efficiently discover all of the locally

and globally compatible features output by $\mathcal{S}$?

We approach this problem by recasting to the problem of finding global sections of quiver representations. The presence of noise severely impairs the compatibility framework described above. The main difficulty is that one never obtains an exact equality between (projections of) features selected over an admissible set $|\tau|$ and features selected over an admissible subset $|\sigma|$. The best-case scenario is that features relevant to $|\sigma|$ lie *near*¹ the restriction of some feature relevant to $|\tau|$. In fact, exact equality may not even be desirable — often, features are highly correlated, and collapsing together such features will further reduce the dimension of the space of selected features. Thus, there are many compelling reasons to seek a less rigid compatibility framework. Here, we use the notion of approximate global section, introduced in Chapter 3.

4.2.2 From Feature Selection to Quiver Representations

Given a finite set X let X^* be the $\mathbb{R}$ -vector space which consists of all functions $X \rightarrow \mathbb{R}$. Denote by $\text{Sub}(X^*)$ the collection of all subsets of X^* (not necessarily subspaces), and for each such subset $\mathcal{F}$ write $\mathbf{V}(\mathcal{F})$ to indicate the subspace of X^* given by the $\mathbb{R}$ -linear span of $\mathcal{F}$. Let $\mathbf{Gr}_k(X^*)$ be the space of k -dimensional subspaces of X^* , known as the Grassmannian, for each $k \geq 0$, and consider the disjoint union

$$\mathbf{Gr}_\infty(X^*) := \coprod_{k \geq 0} \mathbf{Gr}_k(X^*).$$

Here is the main object of study in this chapter.

Definition 4.2.1. A *feature selector* on X is any map $\mathcal{S} : \text{Sub}(X^*) \rightarrow \mathbf{Gr}_\infty(X^*)$, such that for each input $\mathcal{F} \subset X^*$, the corresponding output $\mathcal{S}(\mathcal{F})$ is a subspace of $\mathbf{V}(\mathcal{F})$.

¹Here proximity is determined by the inner product metric on $\mathbf{V}(\mathcal{F})$.

We now fix a feature selector $\mathcal{S}$ on X ; readers may wish to keep the following concrete example in mind throughout the remainder of §4.2.2.

Example 4.2.2. Among the most well-known and ubiquitous feature selectors is *principal components analysis* [43] — here,

- X is a finite origin-centered subset of some real Euclidean space $\mathbb{R}^n$,
- $\mathbb{k}$ is the field $\mathbb{R}$ of real numbers, and
- $\mathcal{F}$ is the set $\{x \mapsto \langle x, p \rangle\}$ of maps given by taking inner products with the unit vectors p .

For a fixed $k \ll n$, the output $\mathcal{S}(\mathcal{F})$ is the vector space spanned by the k inner product functions (or equivalently, the k unit vectors) along which the variance in X is maximised.

If $Y \subset X$ is a subset, we can restrict any function in X^* to Y^* — for a subset of functions $\mathcal{F} \subset X^*$ we will denote the restrictions of all functions in $\mathcal{F}$ to Y as $\mathcal{F}|_Y$. Our goal here is to find a principled framework for relating the vector subspaces $\mathcal{S}(\mathcal{F}|_Y)$ to each other for a fixed $\mathcal{F}$ as Y ranges over subsets of X . Already for principal components analysis, it is clear that there need not be any coherent relationship between $\mathcal{S}(\mathcal{F}|_Y)$ and $\mathcal{S}(\mathcal{F})$ for arbitrary $Y \subset X$ — for instance, neither one is guaranteed to be a subspace of the other within $\mathbf{V}(\mathcal{F})$. It therefore becomes necessary to constrain the family of subsets under consideration. Here we will describe how to relate the subspaces $\mathcal{S}(\mathcal{F}|_Y)$ for Y ranging over the lattice of subsets generated from a chosen *cover* $\mathcal{U}$ of X . We recall that any such $\mathcal{U}$ is a finite collection of nonempty subsets satisfying $X = \bigcup_{U \in \mathcal{U}} U$.

Definition 4.2.3. The *quiver associated to a cover* $\mathcal{U}$ of X is the quiver $Q^{\mathcal{U}}$ whose

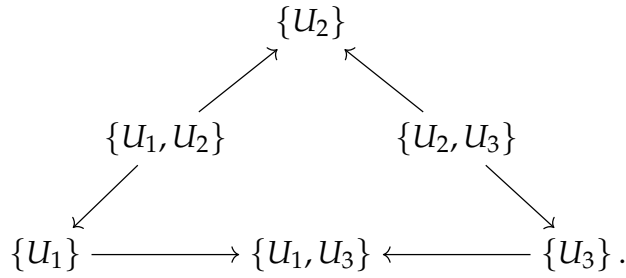
1. vertices are subsets $\sigma \subset \mathcal{U}$ whose *support* $|\sigma| := \bigcap_{U \in \sigma} U$ is nonempty; and,

2. there is a unique directed edge $\sigma \rightarrow \tau$ whenever σ properly contains τ (and hence $|\sigma| \subset |\tau|$).

Example 4.2.4. Suppose $X = \{x_1, x_2, x_3\}$ and define the elements of $\mathcal{U}$ as

$$U_a = \{x_1, x_2\}, \quad U_b = \{x_1, x_3\}, \quad U_c = \{x_2, x_3\}.$$

Then $Q^{\mathcal{U}}$ is



Given such a quiver $Q^{\mathcal{U}}$, a feature selector $\mathcal{S}$ defines a subspace arrangement

$$\mathbf{S}^{\mathcal{F}, \mathcal{U}} = \left\{ S(\mathcal{F}|_{|\sigma|}) \right\}_{\sigma \in Q_0^{\mathcal{U}}}$$

of $Q^{\mathcal{U}}$ in $\mathbf{V}(\mathcal{F})$. As described in Definition 3.1.2, this induces a representation $\mathbf{S}_{\bullet}^{\mathcal{F}, \mathcal{U}}$ of $Q^{\mathcal{U}}$ where the map from $\sigma \supset \tau$ is associated to the orthogonal projection

$$\mathbf{S}_{\sigma \rightarrow \tau}^{\mathcal{F}, \mathcal{U}} := \pi_{\tau}^{\mathcal{S}} \iota_{\sigma}^{\mathcal{S}} : S(\mathcal{F}|_{|\sigma|}) \rightarrow S(\mathcal{F}|_{|\tau|})$$

where the maps $\pi_{\tau}^{\mathcal{S}}$ is the orthogonal projection of $\mathbf{V}(\mathcal{F})$ onto $S(\mathcal{F}|_{|\tau|})$ and $\iota_{\sigma}^{\mathcal{S}}$ is the inclusion of $S(\mathcal{F}|_{|\sigma|})$ into $\mathbf{V}(\mathcal{F})$. We will call this the $\mathcal{S}$ -representation of $Q^{\mathcal{U}}$ along $\mathcal{F}$.

Remark 4.2.5. It may be worth noting that the vertices of $Q^{\mathcal{U}}$ are simplices in the nerve of the cover $\mathcal{U}$ [21], and every edge $\sigma \rightarrow \tau$ in $Q^{\mathcal{U}}$ corresponds to a face relation (where the simplex τ lies in the boundary of the simplex σ). In particular, it follows from the definition of $Q^{\mathcal{U}}$ above that the existence of adjacent edges $\sigma \rightarrow \tau \rightarrow \gamma$ forces the existence of the edge $\sigma \rightarrow \gamma$. We call $\mathbf{A}_{\bullet}$ a **sheaf** on the

nerve of $\mathcal{U}$ if it satisfies the following *associativity criterion*: for every such pair of adjacent edges, the map $\mathbf{A}_{\sigma \rightarrow \gamma}$ must equal the composite $\mathbf{A}_{\tau \rightarrow \gamma} \circ \mathbf{A}_{\sigma \rightarrow \tau}$. In general, we do not expect this associative property to hold when $\mathbf{A}_\bullet$ arises from a feature selector. The next section contains our attempt to rectify this shortcoming.

4.2.3 Compatibility and Sections

Here we use the $\mathcal{S}$ -representation to define and study two notions of compatibility for a feature selector $\mathcal{S}$ on X against a cover $\mathcal{U}$ when invoked with input $\mathcal{F} \subset X^*$:

1. Fix a vertex $\sigma \in Q_0^{\mathcal{U}}$. An element $v_\sigma \in \mathbf{S}_\sigma^{\mathcal{F}, \mathcal{U}}$ is *locally compatible* with $\mathcal{U}$ if

$$\iota_\sigma^{\mathcal{S}}(v_\sigma) = \iota_\tau^{\mathcal{S}} \circ \mathbf{S}_{\sigma \rightarrow \tau}^{\mathcal{F}, \mathcal{U}}(v_\sigma)$$

holds for every edge $\sigma \rightarrow \tau$ in Q_1 .

2. A collection $\{v_\sigma \in \mathbf{S}_\sigma^{\mathcal{F}, \mathcal{U}} \mid \sigma \in Q_0^{\mathcal{U}}\}$ is *globally compatible* with $\mathcal{U}$ if the equality

$$\mathbf{S}_{\sigma \rightarrow \tau}^{\mathcal{F}, \mathcal{U}}(v_\sigma) = v_\tau$$

holds for every edge $\sigma \rightarrow \tau$ in $Q^{\mathcal{U}}$.

This notion of local compatibility checks whether or not (the orthogonal projections of) a given feature v_σ which has been selected by $\mathcal{S}$ over $|\sigma|$ continue to be selected over all greater subsets $|\tau| \supset |\sigma|$ generated by the cover $\mathcal{U}$. In this sense, local compatibility tests the robustness of selected features as we decrease the depth of the cover. Already for principal component analysis, it is clear that not every feature will be locally compatible, i.e., in general $\mathbf{S}_{\sigma \rightarrow \tau}^{\mathcal{F}, \mathcal{U}}$ has a non-trivial kernel. Global compatibility, on the other hand, tests features horizontally across the quiver — to be globally compatible, the vectors of a family $\{v_\sigma\}$ must map

coherently onto each other under the linear maps of $\mathbf{S}_{\bullet}^{\mathcal{F}, \mathcal{U}}$ as we traverse zigzag paths in $Q^{\mathcal{U}}$ of the form

$$\sigma_1 \rightarrow \tau_1 \leftarrow \sigma_2 \rightarrow \tau_2 \leftarrow \cdots \leftarrow \sigma_k \rightarrow \tau_k.$$

Global sections of a quiver representation, as defined in Definition 3.1.3, are closely related to the compatibility criteria from Definition 4.2.6. In particular, if $\mathbf{S}$ is a feature selector on a set X , then the sections of $\mathbf{S}_{\bullet}^{\mathcal{F}, \mathcal{U}}$ are exactly globally compatible features of $\mathcal{S}$ against the cover $\mathcal{U}$. Sections can also be used to compute locally compatible features, as we will now describe.

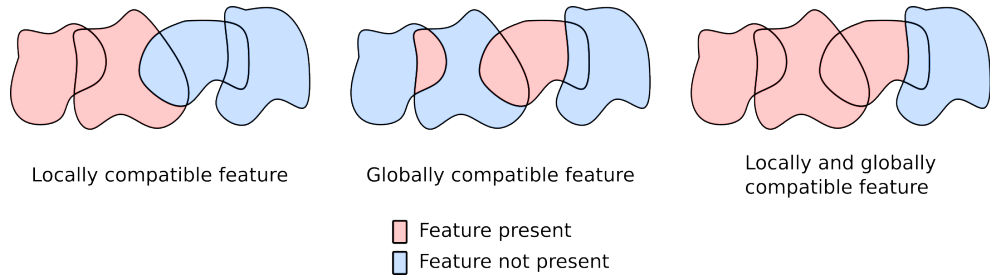


Figure 4.2: Example of different compatibility conditions.

4.2.4 Local Compatibility

Definition 4.2.6. A feature selector $\mathcal{S}$ on X is **locally compatible** with $\mathcal{U}$ on $\mathcal{F} \subset X^*$ if the following triangle of vector spaces commutes for each edge $\sigma \rightarrow \tau$ of $Q^{\mathcal{U}}$:

$$\begin{array}{ccc} \mathbf{S}_{\sigma}^{\mathcal{F}, \mathcal{U}} & \xrightarrow{\mathbf{S}_{\sigma \rightarrow \tau}^{\mathcal{F}, \mathcal{U}}} & \mathbf{S}_{\tau}^{\mathcal{F}, \mathcal{U}} \\ & \searrow \iota_{\sigma}^{\mathcal{S}} & \swarrow \iota_{\tau}^{\mathcal{S}} \\ & \mathbf{V}(\mathcal{F}) & \end{array}$$

Equivalently, the identity $\iota_{\sigma}^{\mathcal{S}} = \iota_{\tau}^{\mathcal{S}} \circ \pi_{\tau}^{\mathcal{S}} \circ \iota_{\sigma}^{\mathcal{S}}$ holds for every edge $\sigma \rightarrow \tau$ of $Q^{\mathcal{U}}$.

We can rephrase this condition in terms of transformations to a constant representation:

Proposition 4.2.7. *The feature selector $\mathcal{S}$ is locally compatible with $\mathcal{U}$ on $\mathcal{F} \subset X^*$ if and only if the collection of maps*

$$\iota^{\mathcal{S}} = \left\{ \iota_{\sigma}^{\mathcal{S}} : \mathbf{S}_{\sigma}^{\mathcal{F}, \mathcal{U}} \hookrightarrow \mathbf{V}(\mathcal{F})_{\sigma} = \mathbf{V}(\mathcal{F}) \right\}$$

indexed by vertices σ of $Q^{\mathcal{U}}$ prescribe a morphism of quiver representations between $\mathbf{S}_{\bullet}^{\mathcal{F}, \mathcal{U}}$ and the constant representation $\mathbf{V}(\mathcal{F})_{\bullet}$. In other words, the defect of ι is 0.

We will now describe how we can compute the locally compatible features by computing the global sections of a particular quiver representation.

Definition 4.2.8. Let Q be a quiver. For each vertex $u \in Q_0$ define the **floret** of based at u to be the following quiver $F^{Q,u}$:

1. the set of vertices is the disjoint union

$$F_0^{Q,u} = \{u\} \sqcup \{v_e \mid (e : u \rightarrow v) \in Q_1\}.$$

In other words, there is a distinguished copy of u , which we label u_0 henceforth, and an additional object denoted v_e for every edge in Q_1 from u to v .

2. Define the set of edges of $F^{Q,u}$ as

$$F_1^{Q,u} = \{e_{LD}, e_{DL} : u_0 \rightarrow v_e \mid (e : u \rightarrow v) \in Q_1\}.$$

In other words, for each edge $e : u \rightarrow v$ in Q there are two distinct edges e_{LD} and e_{DL} from u_0 to v_e in $F^{Q,u}$.

Now suppose $\mathbf{A}_{\bullet}$ and $\mathbf{A}'_{\bullet}$ are representations of the quiver Q with a transformation

$\tau : \mathbf{A}_\bullet \rightarrow \mathbf{A}'_\bullet$. Define the following representation $\mathbf{F}_\bullet^{\tau,u}$ of FQ,u :

$$\begin{aligned} \mathbf{F}_{u_0}^{\tau,u} &= \mathbf{A}_u & \mathbf{F}_{v_e}^{\tau,u} &= \mathbf{A}'_v \\ \mathbf{F}_{e_{LD}}^{\tau,u} &= \mathbf{A}'_e \circ \tau_u & \mathbf{F}_{e_{DL}}^{\tau,u} &= \tau_{t(e)} \circ \mathbf{A}_e. \end{aligned}$$

See Figure 4.2.8 for an illustration.

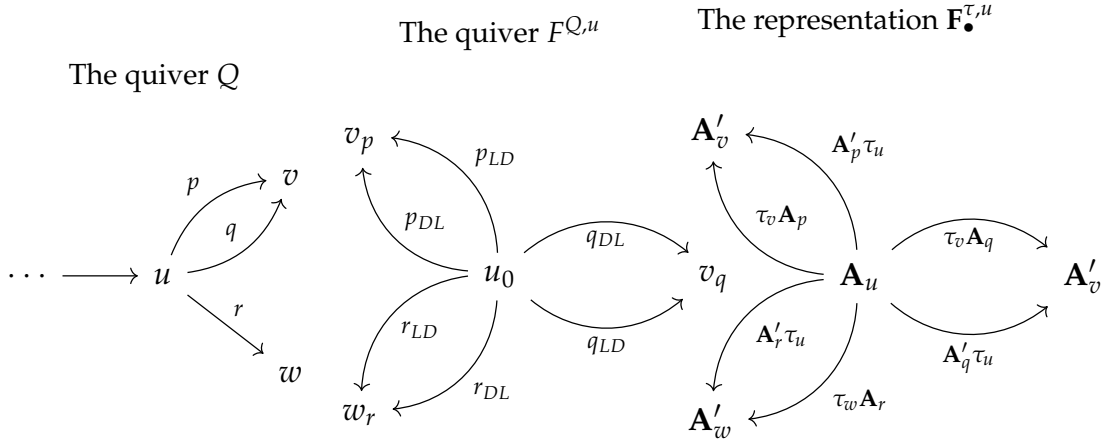


Figure 4.3: An illustration of the floret based at u and its representation.

Constructing the floret now allows us to compute locally compatible features: in the case where the transformation is $\iota^{\mathcal{S}} : \mathbf{S}_\bullet^{\mathcal{F},\mathcal{U}} \rightarrow \mathbf{V}(\mathcal{F})_\bullet$ as in Proposition 4.2.7, the sections $\Gamma(\mathbf{F}_\bullet^{\iota^{\mathcal{S}}\sigma})$ are exactly the locally compatible features at σ .

4.2.5 Bicompatible features

In Section 3.2 below, we will consider the problem of discovering features which satisfy approximate versions of the local and global compatibility criteria. For this purpose, it will be helpful to construct a single quiver whose global sections correspond to features which are both locally and globally compatible. The first step in this construction is a method for gluing quivers and their representations. Let $Q = (s, t : Q_1 \rightarrow Q_0)$ and $Q' = (s', t' : Q'_1 \rightarrow Q'_0)$ be two quivers.

Definition 4.2.9. For each subset $R \subset Q_0 \times Q'_0$, let $Q \cup_R Q'$ be the **merged quiver** whose

1. vertex set is the quotient $V := (Q_0 \sqcup Q'_0)/R$,
2. edge set is the disjoint union $E := Q_1 \sqcup Q'_1$,

and source/target maps are given as follows. Let ρ be the composite $Q_0 \hookrightarrow Q_0 \sqcup Q'_0 \twoheadrightarrow V$, and define $\rho' : Q'_0 \rightarrow V$ similarly. If $e \in E$ lies in Q_1 , then its source and target vertices are $\rho \circ s(e)$ and $\rho \circ t(e)$; and if e lies in Q'_1 , then its source and target vertices are $\rho' \circ s'(e)$ and $\rho' \circ t'(e)$ respectively.

This is a particular case of the pushout in the category of quivers.

Fix representations $\mathbf{A}_\bullet$ of a quiver Q and $\mathbf{A}'_\bullet$ of another quiver Q' .

Definition 4.2.10. Let $R \subset Q_0 \times Q'_0$ be any subset for which we have $\mathbf{A}_i = \mathbf{A}'_j$ whenever (i, j) lies in R . The **merged representation** of $\mathbf{A}_\bullet$ and $\mathbf{A}'_\bullet$, denoted $(\mathbf{A} \cup_R \mathbf{A}')_\bullet$, is the following representation of $Q \cup_R Q'$. To vertices, it assigns

$$(\mathbf{A} \cup_R \mathbf{A}')_i := \begin{cases} \mathbf{A}_i & i \in Q_0 \text{ and } (i, j) \notin R \text{ for any } j \in Q'_0 \\ \mathbf{A}'_i & i \in Q'_0 \text{ and } (j, i) \notin R \text{ for any } j \in Q_0 \\ \mathbf{A}_i = \mathbf{A}'_j & \text{if } (i, j) \in R \\ \mathbf{A}_j = \mathbf{A}'_i & \text{if } (j, i) \in R \end{cases}$$

And on edges, we have

$$(\mathbf{A} \cup_R \mathbf{A}')_e := \begin{cases} \mathbf{A}_e & \text{if } e \in Q_1, \\ \mathbf{A}'_e & \text{if } e \in Q'_1. \end{cases}$$

One may embed both $\text{Tot}(\mathbf{A}_\bullet)$ and $\text{Tot}(\mathbf{A}'_\bullet)$ within the total space of the merged representation in the natural way, and it is readily seen that the space of sections

of $(\mathbf{A} \cup_R \mathbf{A}')_\bullet$ is (isomorphic to) the intersection $\Gamma(\mathbf{A}_\bullet) \cap \Gamma(\mathbf{A}'_\bullet)$.

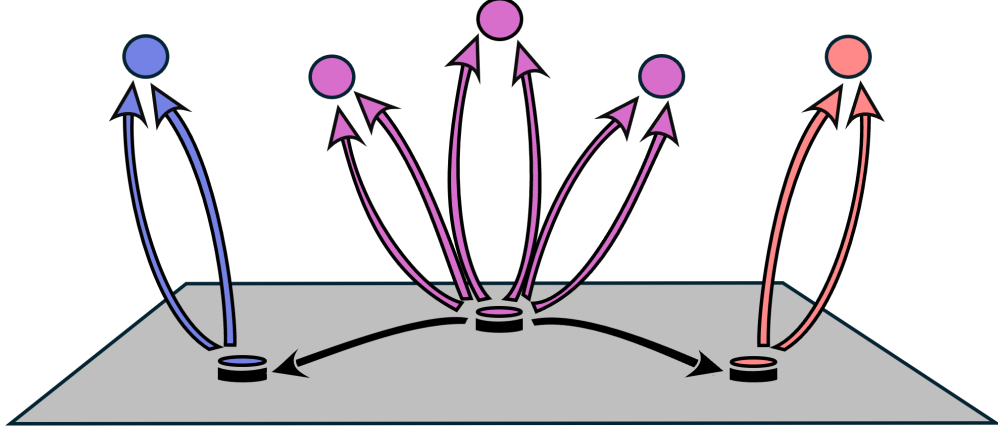


Figure 4.4: Illustration of the merged quiver

Returning to the case of interest, let $\mathcal{S}$ be a feature selector on a finite set X ; consider a cover $\mathcal{U}$ of X and a subset $\mathcal{F} \subset X^*$. Let $\widehat{Q}^{\mathcal{U}}$ be the quiver given by the disjoint union of florets $F^{Q,\sigma}$ indexed by vertices $\sigma \in Q_0^{\mathcal{U}}$, and let $\widehat{\mathbf{S}}^{\mathcal{F},\mathcal{U}}$ be the representation of $\widehat{Q}^{\mathcal{U}}$ given by the corresponding representations $\mathbf{F}^{\mathcal{S},\sigma}$. We define $R \subset Q_0^{\mathcal{U}} \times \widehat{Q}_0^{\mathcal{U}}$ as the collection

$$R := \left\{ (\sigma, \sigma_0) \mid \sigma \in Q_0^{\mathcal{U}} \right\},$$

where σ_0 is the central vertex of $F^{Q,\sigma}$.

Definition 4.2.11. The *combined $\mathcal{S}$ -representation* (of $Q^{\mathcal{U}} \cup_R \widehat{Q}^{\mathcal{U}}$, along $\mathcal{F}$) is the merged representation $\mathbf{N}_{\bullet}^{\mathcal{S},\mathcal{F},\mathcal{U}} := \mathbf{S}_{\bullet}^{\mathcal{F},\mathcal{U}} \cup_R \widehat{\mathbf{S}}_{\bullet}^{\mathcal{F},\mathcal{U}}$.

By design, the sections of $\mathbf{N}_{\bullet}^{\mathcal{S},\mathcal{F},\mathcal{U}}$ correspond to relevant features which are both locally and globally compatible with $\mathcal{U}$. We call these sections the *bicompatible features* of $\mathcal{S}$. This construction can be rephrased much more generally with categories; this is explained in Appendix Section A.2.

Example 4.2.12. Suppose $\mathcal{U} = \{U_1, U_2\}$ is a cover of X , such that $U_1 \cap U_2 \neq \emptyset$.

Define $\tau_1 = \{U_1\}$, $\tau_2 = \{U_2\}$ and $\sigma = \{U_1, U_2\}$. The quiver $Q^{\mathcal{U}}$ is

$$\tau_1 \longleftarrow \sigma \longrightarrow \tau_2$$

and the corresponding quiver $Q^{\mathcal{U}} \cup_R \widehat{Q}^{\mathcal{U}}$ is

$$\begin{array}{ccccc} \tau_1 & \longleftarrow & \sigma & \longrightarrow & \tau_2 \\ & & \swarrow & \searrow & \\ & & v_{\sigma \rightarrow \tau_1} & & v_{\sigma \rightarrow \tau_2} \end{array}$$

Consider a bicompatible feature, i.e., a section $x \in \Gamma(\mathbf{N}_{\bullet}^{\mathcal{S}, \mathcal{F}, \mathcal{U}})$. By definition, we have $\iota_{\tau_2}^{\mathcal{S}}(x_{\tau_2}) = \iota_{\sigma}^{\mathcal{S}}(x_{\sigma}) = \iota_{\tau_1}^{\mathcal{S}}(x_{\tau_1})$, hence

$$\Gamma(\mathbf{N}_{\bullet}^{\mathcal{S}, \mathcal{F}, \mathcal{S}}) \cong \iota^{\mathcal{S}}(\mathbf{S}_{|\tau_1|}^{\mathcal{F}, \mathcal{U}}) \cap \iota^{\mathcal{S}}(\mathbf{S}_{|\sigma|}^{\mathcal{F}, \mathcal{U}}) \cap \iota^{\mathcal{S}}(\mathbf{S}_{|\tau_2|}^{\mathcal{F}, \mathcal{U}}).$$

Thus, bicompatible features are the ones which are relevant for every $U \in \mathcal{U}$.

4.2.6 Opposite Representations

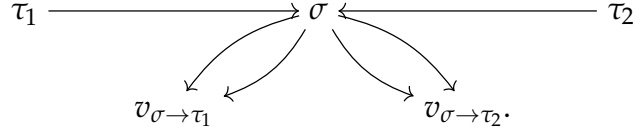
If one is also interested in features which are relevant for some – but not all – $U \in \mathcal{U}$, then it becomes necessary to consider a different notion of bicompatibility. Given a quiver $Q = (s, t : Q_1 \rightarrow Q_0)$, its *opposite* Q^{op} is the quiver with the same vertices and edges but with source and target maps interchanged, i.e., $s^{\text{op}} := t$ and $t^{\text{op}} := s$. Every representation $\mathbf{A}_{\bullet}$ of Q valued in $\mathbf{FdHilb}(\mathbb{k})$ induces a unique *opposite representation* $\mathbf{A}_{\bullet}^{\text{op}}$ of Q^{op} , which is obtained by taking the adjoint of every edge map: namely, $(\mathbf{A}^{\text{op}})_e := \mathbf{A}_e^*$ for each $e \in Q_1$.

When $\mathbf{A}_{\bullet}$ is the $\mathcal{S}$ -representation of a feature selector $\mathcal{S}$ along some input $\mathcal{F}$, the fact that $\iota_{\sigma}^{\mathcal{S}}$ and $\pi_{\sigma}^{\mathcal{S}}$ form an adjoint pair for each vertex $\sigma \in Q_0^{\mathcal{U}}$ implies that for every edge $\sigma \rightarrow \tau$ in $Q^{\mathcal{U}}$, we have

$$(\mathbf{S}^{\mathcal{F}, \mathcal{U}})_{\tau \rightarrow \sigma}^{\text{op}} = (\pi_{\tau}^{\mathcal{S}} \circ \iota_{\sigma}^{\mathcal{S}})^* = \pi_{\sigma}^{\mathcal{S}} \circ \iota_{\tau}^{\mathcal{S}}.$$

Definition 4.2.13. The **mixed S-representation** (of $(Q^{\mathcal{U}})^{\text{op}} \cup_R \widehat{Q}^{\mathcal{U}}$, along $\mathcal{F}$) is the merged representation $\mathbf{M}_{\bullet}^{\mathcal{S}, \mathcal{F}, \mathcal{U}} := \left((\mathbf{S}^{\mathcal{F}, \mathcal{U}})^{\text{op}} \cup_R \widehat{\mathbf{S}}^{\mathcal{F}, \mathcal{U}} \right)_{\bullet}$.

Let us revisit Example 4.2.12 with a view towards understanding the sections of $\mathbf{M}_{\bullet}^{\mathcal{S}, \mathcal{F}, \mathcal{U}}$. The corresponding quiver $(Q^{\mathcal{U}})^{\text{op}} \cup_R \widehat{Q}^{\mathcal{U}}$ is



Now the sections are given by the direct sum

$$\Gamma(\mathbf{M}_{\bullet}^{\mathcal{S}, \mathcal{F}, \mathcal{U}}) \cong \left(\iota_{\tau_1}^{\mathcal{S}}(\mathbf{S}_{|\tau_1|}^{\mathcal{F}, \mathcal{U}}) \cap \iota_{\sigma}^{\mathcal{S}}(\mathbf{S}_{|\sigma|}^{\mathcal{F}, \mathcal{U}})^{\perp} \right) \oplus \left(\iota_{\tau_2}(\mathbf{S}_{|\tau_2|}^{\mathcal{F}}) \cap \iota_{\sigma}(\mathbf{S}_{|\sigma|}^{\mathcal{F}})^{\perp} \right) \oplus \Gamma(\mathbf{N}_{\bullet}^{\mathcal{S}, \mathcal{F}}).$$

In this case, $\iota_{\tau_1}^{\mathcal{S}}(\mathbf{S}_{|\tau_1|}^{\mathcal{F}, \mathcal{U}}) \cap \iota_{\sigma}^{\mathcal{S}}(\mathbf{S}_{|\sigma|}^{\mathcal{F}, \mathcal{U}})^{\perp}$ corresponds to features that are relevant to U_1 but not to $U_1 \cap U_2$. Note that

$$\left(\iota_{\tau_1}^{\mathcal{S}}(\mathbf{S}_{|\tau_1|}^{\mathcal{F}, \mathcal{U}}) \cap \iota_{\sigma}^{\mathcal{S}}(\mathbf{S}_{|\sigma|}^{\mathcal{F}, \mathcal{U}})^{\perp} \right) \oplus \left(\iota_{\tau_2}(\mathbf{S}_{|\tau_2|}^{\mathcal{F}}) \cap \iota_{\sigma}(\mathbf{S}_{|\sigma|}^{\mathcal{F}})^{\perp} \right)$$

might not be trivial, i.e. there could be a feature relevant to both U_1 and U_2 but not to $U_1 \cap U_2$, and hence this feature corresponds to two independent sections. Unlike before, a feature being globally compatible is not necessarily locally compatible.

4.2.7 Reduction of the Feature Selector Problem

Using the quiver operations in §3.5, we can reduce the number of vertices in the quiver whilst preserving the space of sections; this allows us to reduce the difficulty of computing approximate sections.

Theorem 4.2.14. Let $\mathcal{S}$ be a feature selector on X , $\mathcal{U}$ a cover of X such that $Q^{\mathcal{U}}$ is (weakly) connected, and $\mathcal{F} \subset X^*$. Fix a vertex $\sigma_0 \in Q_0^{\mathcal{F}}$. Then there exists a quiver Q' and a representation $\mathbf{A}'_{\bullet}$ of Q' satisfying the following properties:

1. The quiver Q' has only one vertex and $|Q_0^{\mathcal{F}}|$ edges.

2. The Laplacian of $\mathbf{A}'_{\bullet}$ is:

$$\sum_{\tau \in Q_0^{\mathcal{U}}} [\text{Id}_{\sigma_0} - \pi_{\sigma_0}^S \iota_{\tau}^S \pi_{\tau}^S \iota_{\sigma_0}^S]$$

3. There exist constants $C_1, C_2 > 0$ that depend on $\mathcal{S}, \mathcal{U}$ and $\mathcal{F}$ such that

$$C_1 \lambda_i(\mathbf{N}_{\bullet}^{\mathcal{S}, \mathcal{F}, \mathcal{U}}) \leq \lambda_i(\mathbf{A}'_{\bullet}) \leq C_2 \lambda_{i+k}(\mathbf{N}_{\bullet}^{\mathcal{S}, \mathcal{F}, \mathcal{U}})$$

for $i = 1, \dots, n - k$.

Here n is the total dimension of $\mathbf{N}_{\bullet}^{\mathcal{S}, \mathcal{F}, \mathcal{U}}$ and $k := n - \dim \mathbf{S}_{\sigma_0}^{\mathcal{F}, \mathcal{U}}$.

Proof. We view $\mathbf{N}_{\bullet}^{\mathcal{S}, \mathcal{F}, \mathcal{U}}$ as a sheaf $\mathcal{A}_{\bullet}$ over a quiver Q , defined as follows:

$$\begin{aligned} Q_0 &= \{v_{\sigma} : \sigma \in Q_0^{\mathcal{U}}\} \cup \{v_{\sigma\tau} : (\sigma \rightarrow \tau) \in Q_1^{\mathcal{U}}\}, \\ Q_1 &= \{e_{\sigma\tau}, r_{\sigma\tau}^{DL}, r_{\sigma\tau}^{LD} : (\sigma \rightarrow \tau) \in Q_1^{\mathcal{U}}\}, \\ s(e_{\sigma\tau}) &= v_{\sigma}, \quad t(e_{\sigma\tau}) = v_{\tau}, \\ s(r_{\sigma\tau}^{DL}) &= s(r_{\sigma\tau}^{LD}) = v_{\sigma}, \quad t(r_{\sigma\tau}^{DL}) = t(r_{\sigma\tau}^{LD}) = v_{\sigma\tau}. \end{aligned}$$

This is precisely the merged quiver on which $\mathbf{N}_{\bullet}^{\mathcal{S}, \mathcal{F}}$ is defined. The desired sheaf $\mathcal{A}_{\bullet}$ assigns the following data to vertices and edges of Q for each $\sigma \in Q_0^{\mathcal{U}}$:

$$\begin{aligned} \mathcal{A}_{v_{\sigma} \rightarrow e_{\sigma\tau}} &= \pi_{\tau}^{\mathcal{S}} \circ \iota_{\sigma}^{\mathcal{S}}, & \mathcal{A}_{e_{\sigma\tau} \leftarrow v_{\tau}} &= \text{Id}, & \mathcal{A}_{v_{\sigma} \rightarrow r_{\sigma\tau}^{DL}} &= \iota_{\tau}^{\mathcal{S}} \circ \pi_{\tau}^{\mathcal{S}} \circ \iota_{\sigma}^{\mathcal{S}}, \\ \mathcal{A}_{r_{\sigma\tau}^{DL} \leftarrow v_{\sigma}} &= \text{Id}, & \mathcal{A}_{v_{\sigma} \rightarrow r_{\sigma\tau}^{LD}} &= \iota_{\sigma}^{\mathcal{S}}, & \mathcal{A}_{r_{\sigma\tau}^{LD} \leftarrow v_{\sigma}} &= \text{Id}. \end{aligned}$$

Use Proposition 3.5.7(b) to get a new copy $e_{\sigma\tau}^2$ of the edge $e_{\sigma\tau}$ with the maps multiplied by $\iota_{\tau}^{\mathcal{S}}$. Use this edge with Proposition 3.5.5 to move the edge $r_{\sigma\tau}^{DL}$ to an edge between v_{τ} and $v_{\sigma\tau}$, then using the edge $r_{\sigma\tau}^{LD}$ with Proposition 3.5.5 move this edge

between v_τ and $v_{\sigma\tau}$ to an edge between v_σ and v_τ which has maps ι_σ and ι_τ respectively. Call this edge $p_{\sigma\tau}$. Using Proposition 3.5.7(b) can remove both $e_{\sigma\tau}$ and $e_{\sigma\tau}^2$ since π_τ^S is the left inverse of ι_τ^S . Finally, we can remove the edge $r_{\sigma\tau}^{LD}$ with Proposition 3.5.9 as $(\iota_\sigma^S)^* \iota_\sigma^S = \text{Id}$. Now choose some vertex σ_0 . As the nerve is connected, and each edge from each vertex has the same restriction maps to that vertex we can keep applying Proposition 3.5.5 until all edges have σ_0 as a boundary. There might be double edges, but these can be removed with Proposition 3.5.7(b). Then we apply Proposition 3.5.9 to reduce to just the vertex σ_0 . The Laplacian is now

$$\sum_{\tau \in Q_0^{\mathcal{U}}} \left[\text{Id}_{\sigma_0} - \pi_{\sigma_0}^S \iota_\tau^S \pi_\tau^S \iota_{\sigma_0}^S \right]$$

and the eigenvalue relationship follows by tracking all the uses of Propositions 3.5.7(a), 3.5.7(b), 3.5.5, 3.5.9. $\square$

Theorem 4.2.15. *Let $\mathcal{S}$ be a feature selector on a set X equipped with a cover $\mathcal{U}$. For each subset $\mathcal{F} \subset X^*$, there exists a sheaf $\mathcal{A}'$ over a quiver Q' with Laplacian*

$$(L_{\mathcal{A}'})_{\sigma, \tau} = \begin{cases} \sum_{\tau \subset \sigma} (2\text{Id}_\sigma - \pi_\sigma^S \iota_\tau^S \pi_\tau^S \iota_\sigma^S) + \sum_{\sigma \subset \gamma} \pi_\sigma^S \iota_\gamma^S \pi_\gamma^S \iota_\sigma^S & \text{if } \sigma = \tau, \\ -\pi_\sigma^S \iota_\tau^S & \text{if } \tau \subset \sigma \text{ or } \sigma \subset \tau, \\ 0 & \text{otherwise.} \end{cases}$$

such that there is an isomorphism $\Gamma(\mathcal{A}'_\bullet) \cong \Gamma(\mathbf{M}_\bullet^{\mathcal{S}, \mathcal{F}, \mathcal{U}})$; and moreover, we have

$$\lambda_i(\mathbf{M}_\bullet^{\mathcal{S}, \mathcal{F}, \mathcal{U}}) \leq \lambda_i(\mathcal{A}'_\bullet) \leq \lambda_{i+k}(\mathbf{N}_\bullet^{\mathcal{S}, \mathcal{F}, \mathcal{U}})$$

for all $i \in \{1, \dots, n - k\}$. Here $\mathbf{M}_\bullet^{\mathcal{S}, \mathcal{F}, \mathcal{U}}$ is the mixed representation of Definition 4.2.13, while n is its total dimension and $k := n - \sum_{\sigma \in Q_0} \dim \mathbf{S}_\sigma^{\mathcal{F}, \mathcal{U}}$.

Proof. Consider the representation $\left((\mathbf{S}^{\mathcal{F}, \mathcal{U}})^{\text{op}} \cup_R \widehat{\mathbf{S}}^{\mathcal{F}, \mathcal{U}} \right)_\bullet$ as defined in Definition 4.2.13,

and view it as a sheaf $\mathcal{A}_\bullet$ over a quiver Q . We can immediately remove identity self-loops using Proposition 3.5.7(a). The quiver Q is as follows:

$$\begin{aligned} Q_0 &= \left\{ v_\sigma : \sigma \in Q_0^{\mathcal{U}} \right\} \cup \left\{ v_{\sigma\tau} : (\sigma \rightarrow \tau) \in Q_1^{\mathcal{U}} \right\}, \\ Q_1 &= \left\{ e_{\sigma\tau}, r_{\sigma\tau}^{DL}, r_{\sigma\tau}^{LD} : (\sigma \rightarrow \tau) \in Q_1^{\mathcal{U}} \right\}, \\ s(e_{\sigma\tau}) &= v_\tau, \quad t(e_{\sigma\tau}) = v_\sigma, \\ s(r_{\sigma\tau}^{DL}) &= s(r_{\sigma\tau}^{LD}) = v_\sigma, \quad t(r_{\sigma\tau}^{DL}) = t(r_{\sigma\tau}^{LD}) = v_{\sigma\tau}. \end{aligned}$$

The sheaf $\mathcal{A}_\bullet$ consists of the following for each $\sigma \in Q^{\mathcal{U}}$:

$$\begin{aligned} \mathcal{A}_{v_\tau \rightarrow e_{\sigma\tau}} &= \pi_\sigma^S \circ \iota_\tau^S, & \mathcal{A}_{e_{\sigma\tau} \leftarrow v_\sigma} &= \text{Id}, \\ \mathcal{A}_{v_\sigma \rightarrow r_{\sigma\tau}^{DL}} &= \iota_\tau^S \circ \pi_\tau^S \circ \iota_\sigma^S, & \mathcal{A}_{r_{\sigma\tau}^{DL} \leftarrow v_\sigma} &= \text{Id}, \\ \mathcal{A}_{v_\sigma \rightarrow r_{\sigma\tau}^{LD}} &= \iota_\sigma^S, & \mathcal{A}_{r_{\sigma\tau}^{LD} \leftarrow v_\sigma} &= \text{Id}. \end{aligned}$$

We can use Proposition 3.5.9 to remove each $v_{\sigma\tau}$. In doing so, we add three new self-loops to v_σ ; two identities and a self loop with maps ι_σ^S and $\iota_\tau^S \circ \pi_\tau^S \circ \iota_\sigma^S$. The two identity self loops can be removed with Proposition 3.5.7(a). $\square$

4.2.8 Remarks on computation

In Theorems 4.2.14 and 4.2.15, we have assumed that L is Hermitian with respect to some orthonormal basis. Often, however, the bases we would prefer to use are not orthogonal with respect to our chosen inner product. We will see such an example in §5, where computing inner products is straightforward whilst finding an orthonormal basis is computationally infeasible. It is then advantageous then to compute eigenvectors of a Laplacian without such a change of basis. Suppose, for each $\sigma \in \mathcal{U}$, we have some basis $\mathcal{B}_\sigma$ of $\mathbf{S}_{|\sigma|}^{\mathcal{F}, \mathcal{U}}$ and a Hermitian positive definite matrix M_σ representing the inner product in this basis. As M_σ is positive definite

there exists an invertible W_σ such that $M_\sigma = W_\sigma^* W_\sigma$. Then, in this basis, we now have for $\sigma, \tau \in \mathcal{U}$

$$\begin{aligned}\pi_\sigma^S \iota_\tau^S &= W_\sigma^{-*} M_{\sigma\tau} W_\tau^{-1}, \\ \pi_\sigma^S \iota_\tau^S \pi_\tau^S \iota_\sigma^S &= W_\sigma^{-*} M_{\sigma\tau} M_\tau^{-1} M_{\tau\sigma} W_\sigma^{-1}\end{aligned}$$

where $M_{\sigma\tau}$ is the matrix of inner products between $\mathcal{B}_\sigma$ and $\mathcal{B}_\tau$. Now computing eigenvectors of a Laplacian L in this basis can be realised as the generalised eigenvector problem

$$Lx = \lambda M_{\mathcal{U}} x$$

where $M_{\mathcal{U}}$ is the block diagonal matrix with blocks M_σ for $\sigma \in \mathcal{U}$.

As the matrices M_τ are Hermitian and positive-definite, the products

$$M_{\sigma\tau} M_\tau^{-1} M_{\tau\sigma}$$

can be quickly computed using the Cholesky decomposition of M_τ . Furthermore, as L is Hermitian and usually sparse, there exist computationally efficient methods for calculating its smallest eigenvalues and their eigenvectors, e.g. variations on the Lanczos method [47], combined with the Cholesky decomposition (see e.g. [41]) of $L + \alpha \mathbf{I}$ for some small $\alpha > 0$.

Code Availability

The code for constructing covers of data and for computing the approximate sections of the mixed and combined representations, given a sparse inner product matrix between selected features, is available at:

https://github.com/osumray/harmonic_feature_selection

Chapter 5

Single-Cell Chromatin Accessibility

We will now apply the method of finding consistent features of a feature selector to the analysis of single-cell chromatin accessibility.

5.1 Chromatin accessibility

In Chapter 2 we introduced the *expression* of a gene, which we measure in transcriptomics, by the mRNA count from that gene. This gives us insight into which genes are active within a cell, but then the questions become, why is that gene active, and what makes a cell type display a particular pattern of gene expression?

For a gene, the transcription and regulation thereof is a highly complex subject, involving an orchestra of different proteins and enzymes. To start transcription, the enzyme RNA polymerase must bind to a specific region of DNA near the gene, called the promoter. This binding is facilitated or repressed by transcription factors. These transcription factors can bind to the promoter region itself, or to nearby (*cis*) regulatory elements such as enhancers, or even to distant regulatory elements. The salient point is that transcription involves proteins *binding* to DNA. It turns out that DNA can be physically organised to either block or allow the binding of such proteins, thus affecting the transcription of genes.

DNA is physically organised as *chromatin*, a macromolecule consisting of DNA and structural proteins. The fundamental unit of chromatin is a *nucleosome*, which consists of approximately 147 *base-pairs* (namely, the usual A, C, G and T nucleotides) wound around eight histone molecules [16]. This is commonly referred to as the ‘beads on a string’ model. The density of nucleosomes varies considerably across the genome; low-density ‘open’ regions are more accessible to transcription factors and other binding proteins, while high-density regions are less accessible, thereby reducing gene transcription. Therefore, understanding the relative accessibility of various parts of the genome is important in the study of gene regulation and its variation across different cell types.

5.2 Measuring Chromatin Accessibility: ATAC-seq

ATAC-seq There are several technologies for measuring chromatin accessibility [57], with perhaps the most popular being ATAC-seq¹ [10]. The ATAC-seq protocol works as follows: an enzyme, Tn5 transposase, cleaves the cell DNA with open regions, producing a collection of fragments of DNA. These are then amplified, sequenced, and aligned to the genome. The ends of these fragments correspond to the Tn5 transposase cuts and thus indicate the regions of open chromatin. Recently, this technology has been adapted to allow single-cell sequencing scATAC-seq, using microfluidics [71], the same method that is used in some RNA-seq protocols. There are even technologies that allow ‘multiomics’, the sequencing of both chromatin accessibility and transcriptomics simultaneously.

The output of scATAC-seq may be viewed from a mathematical perspective as follows. Consider a set X of cells $\{x_1, \dots, x_N\}$, with N being on the order of thousands. Each cell x_i has a genome consisting of M base pairs, where M is ap-

¹The Assay for Transposase-Accessible Chromatin using Sequencing

proximately three billion (in the human genome). The output of ATAC-seq on X is an $N \times M$ integer matrix. For each $1 \leq i \leq N$ and $1 \leq j \leq M$, the entry $\alpha_{i,j}$ is a number in $\{0, 1, 2\}$ representing the accessibility for cell x_i at the j -th genomic position. A 1 represents a fragment end at that position on one chromosome, a 2 represents two fragment ends on both chromosomes, and a 0 corresponds to no fragment ends. Often, the chromosomal difference is ignored, giving a binary matrix.

Peak Calling The dimensionality of ATAC-seq data is immense; the human genome contains approximately three billion base pairs, and theoretically each Tn5 transposase cleavage event could occur at any of these positions along the genome. Therefore, the use of a feature selector is crucial to working with ATAC-seq data.

One possible method is to simply tile the genome into equally sized windows, however, this loses fidelity of the data. The more common method is a class of algorithms called *peak-callers* [100]. A popular method is the MACS2 algorithm [106], originally designed for ChIP-seq, an earlier technology. The MACS2 algorithm works approximately as follows: The input to the MACS2 algorithm is the collection of fragment endpoints – in the case of single-cell ATAC-seq, this is the collection of all fragments from all cells, disregarding the origin cells of the fragments. This defines a density of Tn5 transposase cuts along the genome for the population of cells. Significant regions, known as peaks, are identified as regions of the Tn5 transposase density lies above a certain threshold. This threshold is chosen by first fitting a Poisson distribution to a window around each genome position. The relevant threshold is the density with a user-supplied (corrected) p-value according to this local Poisson model. Within each peak region, the *summit* is defined as the base-pair position with the highest significance under this model.

Stochasticity in ATAC-seq The notion of open chromatin is not exact. Indeed, the Tn5 transposase cleaves the DNA wherever it can access it, with some preference for certain sequence motifs. Therefore there is a random element to the summits called in the peak calling process. In practice, e.g. in [71], the summits are extended into an interval along the genome.

Cell Types The purpose of single-cell ATAC-seq is to understand the differences in gene regulation (via chromatin accessibility) of different cell types within a sample. If we are interested in a rare population of cells, their ATAC-seq profiles will have a small contribution to the Tn5 transposase density used in the MACS2 algorithm, therefore open regions for these cells will not be called when performing feature selection on the whole sample population. The solution, as used in [71], is to first separate the cells into their cell types, then apply the peak-calling algorithm. Unless the classification of cells into cell type is known *a priori*, it must be inferred from the data, e.g. by clustering and labelling. This leads to the problem of conglomerating the different peaks or summits found from these different cell type populations. The solution used in [71] is to take the union of all peaks from different clusters, and remove overlapping regions by keeping the most significant peaks.

5.3 Analysing scATAC-seq with Quiver Laplacians

In the remaining section, we propose the methodology in Section 4.2 as a principled alternative approach to handling peak-calling in single-cell ATAC-seq data. Figure 2 on page 20 gives an overview of how the pipeline is applied to single-cell chromatin accessibility data.

1. Firstly, we will construct a cover of the single-cell data by clustering the data,

and then expanding these clusters so that they may overlap. Each of these clusters represent a group of similar cell states, a cell type. By introducing overlaps we allow for cell types to not be strictly defined, and furthermore, model the underlying topology of cell states with the cover.

2. We model the stochasticity by using an inner product between genomic positions, using a Gaussian kernel. In particular, the accessible genomic positions are only known approximately. Thus, the accessibility status of nearby positions is not independent; we capture this phenomenon via the inner product structure on feature space. This will provide a quiver representation which describes how summits, the open chromatin positions, found by peak-calling on one cluster maps to summits found from a nearby cluster.
3. By using the mixed representation of Definition 4.2.13, we will compute combinations of summits that are approximately locally and globally compatible.

5.3.1 Pre-processing the data

We use the single-cell ATAC-seq data of [71], which consist of 28,274 tumour-infiltrating T cells. The pre-processing steps are the same as those of *ibid*:

1. We align fragments to the GRCH37/HG19 assembly, the human reference genome used in *ibid*.
2. We filter sample profiles to keep those with at least 1,000 fragments and a transcription start site (TSS) enrichment score of 8.
3. We create a count matrix by tiling the genome with 2,500 base-pair windows and counting overlaps of endpoints of fragments with the windows.
4. Next, we binarise this matrix and retain only the top 20,000 most accessible sites; call the resulting $28,724 \times 20,000$ boolean matrix B .

5. We then create a frequency-inverse document frequency matrix, apply truncated singular value decomposition to obtain the 2nd to 25th singular vectors. This results in a reduced matrix E of size 28724×24 .

5.3.2 Generating the quiver

We build an undirected graph G from the reduced matrix E as follows. Its vertices are the cells C_i , which correspond to rows of E . Treating each such row as a vector in $\mathbb{R}^{24}$, let $\delta_{i,j} \geq 0$ be the Euclidean distance between rows i and j for each $i < j$. We construct the 15 nearest neighbour graph generated by the $\delta_{i,j}$ distances. Using the Leiden algorithm [87] the vertex set V of G is partitioned into finitely many communities, $V = \bigsqcup_i V_i$. As in *ibid*, communities of size < 200 are excluded. Deviating from [71], we then generate a cover $\mathcal{U}$ of X from the surviving communities as follows: each cover element $U_i \in \mathcal{U}$ has the form $V_i \cup W_i$, where $W_i \subset (V - V_i)$ consists of those vertices which share an edge with some vertex of V_i . The resulting quiver $Q^{\mathcal{U}}$ has 87 vertices, and is depicted in Figure 5.1A. Similarly, Figure 5.1B displays the contents of each vertex relative to the overall distribution, using the cell types identified in [71].

5.3.3 Building the representation

As in [71], we will use the MACS2 peak-calling algorithm to perform feature selection on each of the subpopulations of the dataset. In order to collate the summits across the different subpopulations, our analysis deviates from [71]. The paper [71] used an iterative removal process for nearby summits (or overlapping peaks): the most significant summit² is kept and any peaks that overlapped are removed. Instead, we use the notion of compatible features as developed in Chapter 4, as we will now describe.

²as measured with the MACS2 algorithm

Having constructed $Q^{\mathcal{U}}$, we apply our feature selector $\mathcal{S}$ – which in this case is the MACS2 peak-calling algorithm [106] – to the fragment profiles of cells corresponding to each individual vertex. In the notation of Section 4.2.2, the input $\mathcal{F}$ is given by the columns of the binarised matrix B ; explicitly, the map $f_j : X \rightarrow \mathbb{R}$ associated to the j -th column of E sends each x_i to the entry $B_{ij} \in \{0, 1\}$. Calling peaks on this set of features produces a vector space $\mathbf{S}_{|\sigma|} \subset \mathbf{V}(\mathcal{F})$ for every vertex σ of $Q^{\mathcal{U}}$. We used the same parameters as in [71], and obtained several hundred thousand summits for each vertex as a result. For the purposes of this demonstration, we retained only the most significant 1,000 summits per vertex. Thus the total dimension of the selected vector spaces is 87,000. Since peak-calling on different sub-populations may change the true summit location on the genome slightly, we follow [103] and use a Gaussian kernel between summits (with bandwidth equal to 1,000 base-pairs) as the inner product on $\mathbf{V}(\mathcal{F})$. This allows us to build the mixed representation $\mathbf{M}^{\mathcal{S}, \mathcal{F}, \mathcal{U}}$ (see Definition 4.2.13).

5.4 Results

We computed the first 3,000 eigenvectors of the Laplacian of $\mathbf{M}^{\mathcal{S}, \mathcal{F}, \mathcal{U}}$. Since the total space admits a distinguished basis whose vectors correspond to summits (i.e., certain genomic positions $\{p_k\}$), we may associate to each eigenvector $v = \sum_k \alpha_k p_k$ the subset of supported genomic positions $P_v := \{k \mid \alpha_k \neq 0\}$ ³. Figure 5.1D shows that for an overwhelming majority of eigenvectors v , the diameter of P_v for most v approximately equals 147 base-pairs, which is the average length of DNA around a nucleosome. A small fraction of the eigenspaces had dimension > 1 ; this resulted in a mixing of genomic positions, as shown by the blue outliers in Figure 5.1D. Nevertheless, in almost all cases, we observed that **the genomic supports of ei-**

³Or more precisely, $|\alpha_k| > 10^{-5}$, the tolerance parameter for ARPACK-NG [49] in our experiment.

genvectors are tightly aligned across the vertices of $Q^{\mathcal{U}}$.

We then performed a more detailed analysis of the sparsity pattern of the 3,000 eigenvectors. The outcome of this analysis has been summarised in Figure 5.1E, which plots the 3,000 eigenvectors against all 87 vertices of $Q^{\mathcal{U}}$ and highlights those vertices along which a given eigenvector has a non-zero component. We observe that the eigenvectors fall broadly into two classes. The first class, which is considerably larger, consists of those **eigenvectors whose genomic support contains summits which are shared by most, if not all vertices**: see, for instance, the components of eigenvector 11 in Figure 5.2. The second class contains **eigenvectors whose genomic support contains summits which are shared by a small fraction of the vertices**. A prime example of this type is eigenvector 2702, which has nonzero components along only 3 of the 87 vertices, and has also been depicted in Figure 5.2.

Finally, we note that (at least for these two examples) **the size of the support of the eigenvector relates to biological function**. Eigenvector 11, for instance, is concentrated in a regulatory region a few hundred base pairs downstream of the gene *FAM72D*. This gene is involved in the cell cycle [23], and is therefore relevant to most subpopulations of cells. On the other hand, eigenvector 2702 lies just downstream of the gene *STAT3*, which is a transcription factor known to be an important regulator for differentiation and proliferation in Th1, Th17, and Treg cells [64]. We also note that this eigenvector was only supported on the vertices 8, (8, 11), and 11 (Figure 5.2). These vertices are particularly enriched in Th1, Th17 and Treg cells (see Figure 5.1B). Indeed, Figure 5.1E reveals that many other eigenvectors are also uniquely supported on the vertices 8, (8, 11), 11 and (1, 11). Since vertex 11 is relatively disconnected from the other vertices, our analysis highlights it as an unusual sub-population of X .

Figure 5.1: Overview of Laplacian from 4.2.15 applied to tumour-infiltrating T cells. (A) Quiver constructed from cover, with vertex size corresponding to the cardinality of the covering set. (B) Log likelihood of each cell type label within each vertex. (C) First 3,000 eigenvalues. (D) Range of genomic positions on each chromosome corresponding to non-zero components of the first 3,000 eigenvectors. Blue points correspond to positions across multiple chromosomes. (E) Binary matrix corresponding to the existence of a non-zero component in each vertex for the first 3,000 eigenvectors.

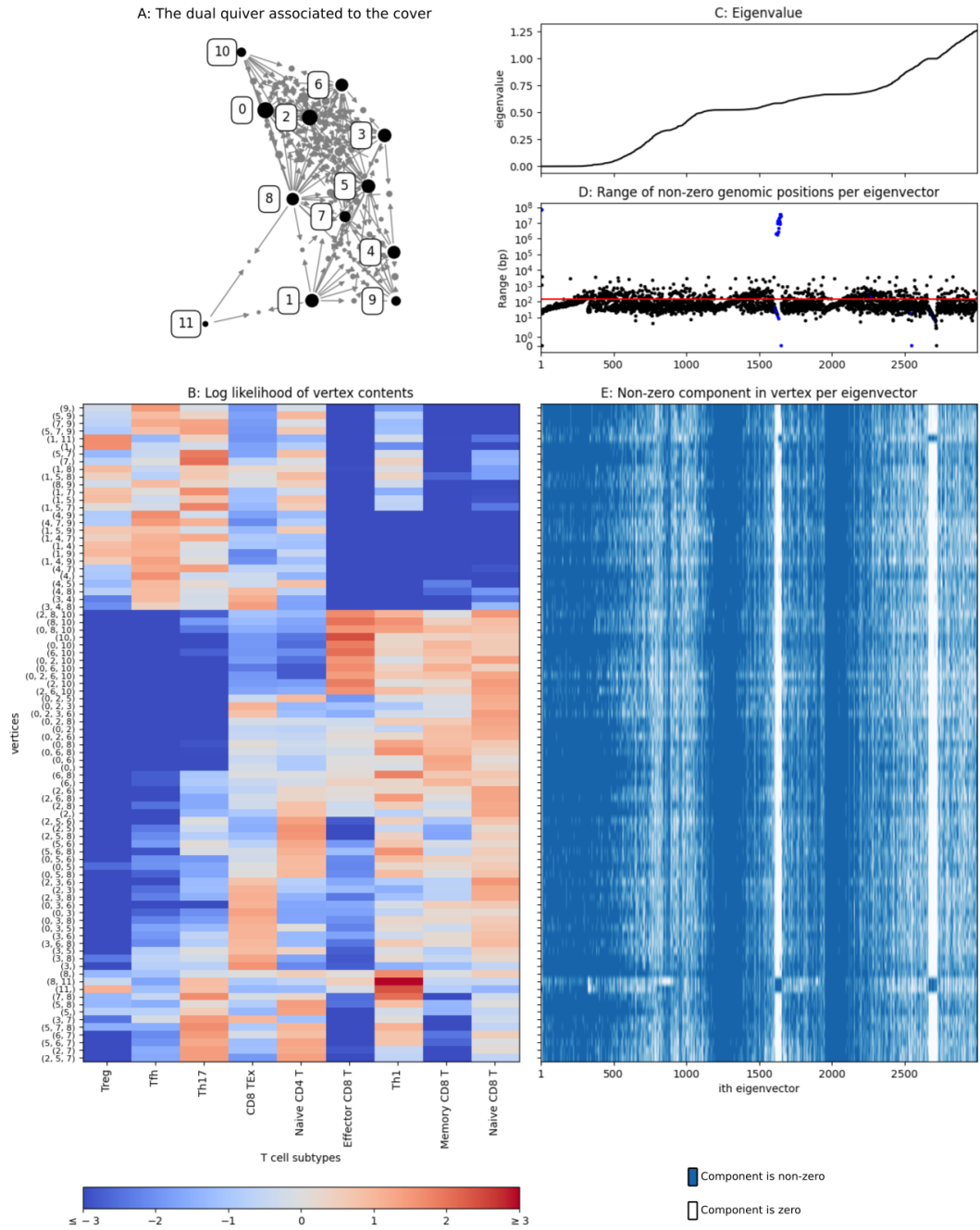
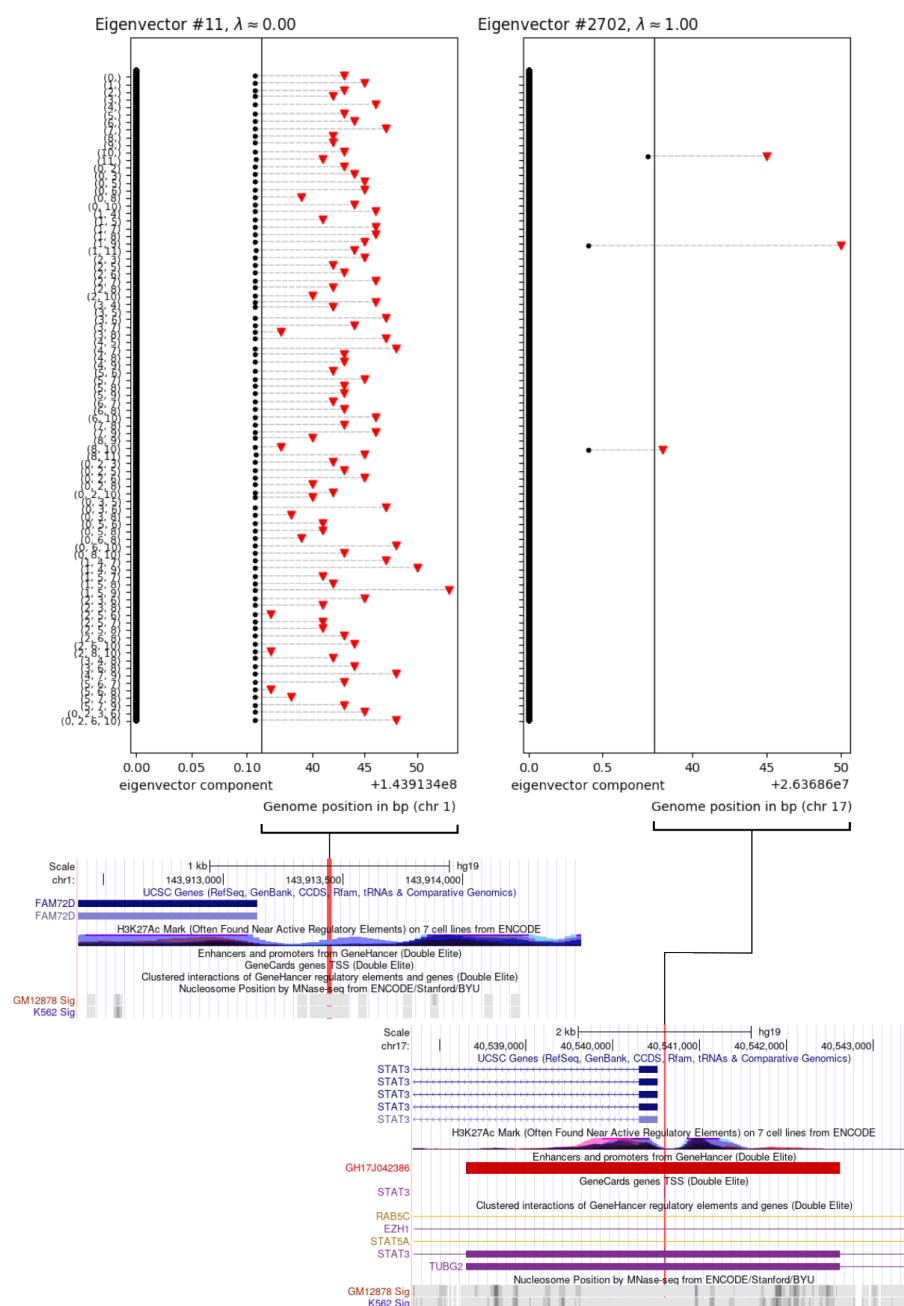


Figure 5.2: Examples of eigenvectors. Eigenvector #11 has a non-zero component at each vertex, whilst eigenvector #2702 has only non-zero components in vertices (8,),(11,) and (8, 11).



Chapter 6

Quiver Laplacians and Persistence Modules

6.1 Background on Persistent Homology

Topological Data Analysis (TDA) is a nascent field in data analysis that seeks to analyse the topology of data by using methods from fields such as algebraic topology. One of the main tools of TDA is persistent homology. Typically we are given a set of data points $X \subset \mathbb{R}^n$, known as a point cloud. Directly looking at the topology of a point cloud is not very interesting, being a collection of 0-dimensional objects. In order to infer the underlying topology of the object that these points were sampled from, a topological model, typically a simplicial complex, is constructed from the point cloud. For example, in the Vietoris Rips complex $VR_\epsilon(X)$ of a point cloud, the d -dimensional simplices are formed by tuples of points that are within distance ϵ of each other. From such a simplicial complex, its homology (usually over a field) can be computed. However, we chose some parameter ϵ to construct the simplicial complex. What if we chose a different parameter?

Definition 6.1.1. Suppose K is a finite simplicial complex. A *filtration* of K is a

function $h : K \rightarrow \mathbb{R}$ such that h is monotonic with respect to the face relation of K : if σ is a face of τ then $h(\sigma) \leq h(\tau)$.

Given a filtration h of K and a value $t \in \mathbb{R}$ the collection

$$K_{h \leq t} = \{\sigma \in K : h(\sigma) \leq t\}$$

is a subcomplex of K by property of h being monotonic.

For a finite point cloud $X \subset \mathbb{R}^n$, let $K = 2^X$ be the power set of X . Define $h : K \rightarrow \mathbb{R}$ by

$$h(\sigma) = \max_{x, y \in \sigma} \|x - y\|_2.$$

Then h is a filtration of K and for each $\varepsilon \geq 0$ we have that

$$K_{h \leq \varepsilon} = \text{VR}_\varepsilon(X).$$

Another common way to build a filtration of a simplicial complex K is to define h on the vertices K^0 of K , then extend this function to all of K in the minimal way for it to become a filtration of K :

$$h(\sigma) := \max_{v \in \sigma} h(v).$$

This is called a *lower-star filtration* of K .

Given a filtration h of K each sequence of values $t_1 \leq \dots \leq t_n \in \mathbb{R}$ corresponds to a sequence of nested subcomplex of K

$$K_{h \leq t_1} \subseteq \dots \subseteq K_{h \leq t_n}.$$

One can think of this as a functor from the poset $[n]$ to **Top**, the category of topological spaces. Applying the degree p homology functor $H_p(-; \mathbb{k})$, for a field

F , gives a functor $[n] \rightarrow \mathbf{FdVect}_{\mathbb{K}}$:

$$H_p(K_{h \leq t_1}; \mathbb{K}) \hookrightarrow \cdots \hookrightarrow H_p(K_{h \leq t_n}; \mathbb{K}).$$

In the language of quiver representations, this sequence of vector spaces is a $\mathbf{FdVect}_{\mathbb{K}}$ -valued representation over the quiver A_n :

$$\bullet_1 \rightarrow \cdots \rightarrow \bullet_n.$$

Such a representation arising from a filtered simplicial complex is known as a *persistence module*. The degree p persistence module of a filtered complex is the solution to the question of changing parameters we raised earlier: the persistence module contains the information of how topological information, e.g. connected components, loops, voids, etc., change with different parameters. This is crucial to the study of TDA, as *a priori* we have no way of knowing which choice of ε returns a correct topological model $\text{VR}_{\varepsilon}(X)$.

Computationally, however, the persistence module is unwieldy. The data of a persistence module is a collection of vector spaces and linear maps, which not only is unilluminating to the topology of our data, but also difficult to perform statistics or comparison on collections of these persistence modules. The solution is to break each persistence module into simpler parts — indecomposables. A quiver representation $\mathbf{I}_{\bullet}$ is indecomposable if

$$\mathbf{I}_{\bullet} \cong \mathbf{I}_{\bullet}^1 \oplus \mathbf{I}_{\bullet}^2$$

implies that either $\mathbf{I}_{\bullet}^1$ or $\mathbf{I}_{\bullet}^2$ is the zero representation. A standard result [9] is that

every quiver representation $\mathbf{A}_\bullet$ valued in $\mathbf{FdVect}$ is decomposable as

$$\mathbf{A}_\bullet \cong \bigoplus_{j \in I} \mathbf{I}_\bullet^j$$

where each $\mathbf{I}_\bullet^j$ is indecomposable, and the multiplicity of each isomorphism class of $\mathbf{I}_\bullet^j$ in the decomposition is unique for each $\mathbf{A}_\bullet$. Thus, $\mathbf{A}_\bullet$ can be described entirely by the indecomposables that appear in this decomposition. The question is then, what are these indecomposables?

Theorem 6.1.2 (Gabriel, [24]). *Every indecomposable representation of the quiver A_n valued in $\mathbf{FdVect}_\mathbb{k}$ is isomorphic to an interval module $\mathbf{I}[b, d]$ defined on vertices by*

$$\mathbf{I}[b, d]_v = \begin{cases} \mathbb{k} & \text{if } b \leq v \leq d, \\ 0 & \text{otherwise,} \end{cases}$$

and on edges by

$$\mathbf{I}[b, d]_e = \begin{cases} 1_\mathbb{k} & \text{if } b \leq s(e), t(e) \leq d, \\ 0 & \text{otherwise.} \end{cases}$$

This is the core result that allows persistent homology to work; each persistence module can be completely described combinatorially by a collection of intervals $[b_i, d_i]$, known as its *barcode*, each describing where a topological feature first appears, its birth b_i , and where this topological feature last appears, its death d_i .

6.1.1 Multiparameter persistence

Persistent homology paired with e.g. the Vietoris-Rips complex allows us to study the topology of a point cloud at different scales ε . Whilst there is a notion of stability for persistent homology [80] the analysis can be confounded when certain types of noise are added. The ideal solution would be to allow the variation of multiple

parameters simultaneously. For example, controlling the distance in which points are joined but also only allowing points with a certain density. This gives rise to multifiltered simplicial complexes and multiparameter persistence modules, representations of the product of quivers A_n .

In the same fashion as single-parameter persistent homology, we would like to decompose such representations into indecomposables. Unfortunately, Gabriel's theorem (Theorem 6.1.2 being a version thereof) tells us that only a handful of quiver types have finitely many indecomposables. There are other 'tame' quivers which have countable many families of indecomposables, with each family parameterised by $\mathbb{k}$. All other quivers are called 'wild': for a given quiver, and a choice of dimension for each vertex, the number of parameters for the space of indecomposables of those dimension is based on the dimension vector of the representation. Thus, attempting to classify all indecomposables is infeasible.

6.2 Harmonic Persistent Homology

While topological data analysis attempts to understand the topology of data, by including a parameter, such as the maximum distance between points in the Vietoris-Rips complex, we invariably include geometric information. It is then reasonable to include more geometric information about our data.

In this section, we describe how one can construct persistence modules with an inner product. The notion of harmonic persistent homology was introduced in [5], which we will now describe.

Let K be a simplicial complex with chain groups

$$C_p(K; \mathbb{k}) = \bigoplus_{\sigma \in K[p]} \mathbb{k}\sigma$$

where K^p denotes the set of p -simplices of K and where $\mathbb{k} = \mathbb{R}$ or $\mathbb{C}$.

Suppose that we also endow each $C_p(K; \mathbb{k})$ with an inner product $\langle -, - \rangle_{C_p(K; \mathbb{k})}$. If $\mathbb{k} = \mathbb{R}$ or $\mathbb{C}$ we will call K a Euclidean or unitary simplicial complex respectively, and simply write the chain groups as $C_p(K)$. In the standard inner product, we treat all p -simplices as orthogonal and of unit length and set

$$\langle \sigma_i, \sigma_j \rangle_{C_p(K)} = \delta_{i,j}$$

for all $\sigma_i, \sigma_j \in K^p$.

If we give the simplicial complex a local orientation by ordering the vertices of the simplicial complex, we can write each p -simplex σ as an ordered p -tuple $[v_0, v_1, \dots, v_n]$ of vertices of K . From this, we can then define the boundary map: for each $p \in \mathbb{Z}$ define $\partial_p(K) : C_p(K) \rightarrow C_{p-1}(K)$ by

$$\partial_p(K)(\sigma) = \sum_{0 \leq j \leq p} [v_0, \dots, \widehat{v}_j, \dots, v_p]$$

where $\widehat{v}_j$ denotes the omission of v_j .

The p -cycles are defined by $Z_p(K) = \ker(\partial_p(K))$ and the p -boundaries by $B_p(K) = \text{im}(\partial_{p+1}(K))$. The p -th homology group is then defined as

$$H_p(K) = Z_p(K) / B_p(K).$$

Definition 6.2.1 (Harmonic homology subspace, [5]). Let K be a Euclidean or unitary simplicial complex. For $p \geq 0$ the *harmonic homology subspace* $\mathcal{H}_p(K)$ is defined as

$$\mathcal{H}_p(K) = Z_p(K) \cap B_p(K)^\perp.$$

There is an isomorphism between each $H_p(K)$ and $\mathcal{H}_p(H)$ given by

$$\pi_p(K) : z + B_p \mapsto \text{proj}_{B_p(K)^\perp}(z).$$

Suppose K_1 is a subcomplex of K_2 where the inner product on K_1 is the restriction of the inner product on K_2 , thus $C_p(K_1)$ is an isometrically embedded subspace of $C_p(K_2)$. Let B_p^1, B_p^2 be the boundary spaces of K_1, K_2 respectively. Then as $B_p^1 \leq B_p^2$, we have $(B_p^1)^\perp \geq (B_p^2)^\perp$. Let $\text{proj}_{B_p(K_2)^\perp}$ be the orthogonal projection of $C_p(K_2)$ onto $(B_p^2)^\perp$. Then as $Z_p^1 \leq Z_p^2$, the image of $\text{proj}_{B_p(K_2)^\perp} \upharpoonright_{\mathcal{H}_p(K_1)}$ is contained in $\mathcal{H}_p(K_2)$. Furthermore, as detailed in [5], the maps

$$\begin{array}{ccc} H_p(K_1) & \xrightarrow{\quad\quad\quad} & H_p(K_2) \\ \downarrow \pi_p(K_1) & & \downarrow \pi_p(K_2) \\ \mathcal{H}_p(K_1) & \xrightarrow{\text{proj}_{B_p(K_2)^\perp} \upharpoonright_{\mathcal{H}_p(K_1)}} & \mathcal{H}_p(K_2) \end{array}$$

commute, where the top map is the induced map on homology from the inclusion $K_1 \hookrightarrow K_2$. As homology is functorial, this implies that the assignment $K_i \mapsto \mathcal{H}_p(K_i)$ is also functorial from the poset category of subcomplexes of K , and that there is a natural isomorphism between the homology groups and the harmonic subspaces. Note that the proof of the above commuting diagram relies on the fact that $C_p(K_1) \hookrightarrow C_p(K_2)$ is an isometry; the harmonic homology subspace is not functorial for general simplicial maps.

This allows to define the *harmonic persistence module* [5] of a filtered Euclidean or unitary simplicial complex:

$$\mathcal{H}_p(K_{h \leq t_1}) \longrightarrow \cdots \longrightarrow \mathcal{H}_p(K_{h \leq t_n}).$$

6.3 Indecomposables of Unitary Quiver Representations

As with quiver representations over **FdVect**, we can decompose representations over **FdHilb** (such as Euclidean and Unitary representations) into indecomposables. However, in order to preserve this metric information, we must insist that the representations are isometrically isomorphic. Now we wish to understand which representations are unitarily indecomposable, at least for the quiver type A_n , as in single-parameter persistent homology.

Unfortunately, the situation for unitary representations is as dire as in multi-parameter persistence. Indeed, as described in [74], the only tame quivers are the two quivers

$$\bullet \qquad \bullet \longrightarrow \bullet.$$

The indecomposable of $\bullet$ is simply the underlying field $\mathbb{C}$ via an orthonormal basis, and the indecomposables of $\bullet \rightarrow \bullet$ are found via singular value decomposition of the linear map:

$$\mathbb{C} \xrightarrow{\sigma \text{Id}} \mathbb{C}.$$

As an example of how much stricter unitary (or orthogonal) decomposition is, consider the following filtered simplicial complex:

$$\bullet^0 \xrightarrow{2} \bullet^1$$

where 0, 1, and 2 indicate when the simplex first appears in the filtration. The degree 0 harmonic persistence module of this filtered complex (assuming all the

simplices are orthogonal) is

$$\mathbb{R} \xrightarrow{\begin{bmatrix} 1 \\ 0 \end{bmatrix}} \mathbb{R}^2 \xrightarrow{\begin{bmatrix} \sqrt{2} & \sqrt{2} \end{bmatrix}} \mathbb{R}.$$

While under non-orthogonal equivalence, as in usual persistent homology, this module decomposes into

$$\begin{array}{ccccc} \mathbb{R} & \xrightarrow{1} & \mathbb{R} & \xrightarrow{1} & \mathbb{R} \\ \oplus & & \oplus & & \oplus \\ 0 & \longrightarrow & \mathbb{R} & \longrightarrow & 0, \end{array}$$

i.e. into two interval modules $I[0, 2]$ and $I[1, 1]$. However, up to orthogonal equivalence, this module is actually indecomposable!

One of the main results of [74] is that for any quiver $Q \notin \{\bullet, \bullet \rightarrow \bullet\}$ there exists an unitarily indecomposable representation of Q with dimensions $z \in \mathbb{Z}^{|Q_0|}$ if and only if $zM_Q \geq z$ where M_Q is $|Q_0| \times |Q_0|$ matrix with entry $m_{u,v}$ being the number of edges between u and v in Q . For example, for the quiver

$$\bullet \rightarrow \bullet \rightarrow \bullet$$

we have that

$$M_Q = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

thus we have an unitarily indecomposable representation of Q for each tuple of integers z_1, z_2, z_3 such that $z_1, z_3 \leq z_2 \leq z_1 + z_3$.

The bottom line is that, just as in multiparameter persistence, searching for a complete invariant for unitary or Euclidean representations is, in a sense, hopeless. We must instead search for other invariants that inform us of the information

within a unitary persistence module.

6.4 The Spectra of Persistence Modules

In Chapter 3 we introduced a Laplacian for quiver representations and noted that the spectrum of the quiver Laplacian is invariant under unitary morphisms and therefore is a suitable candidate for an invariant of unitary quiver representations. Certainly it is not complete – the quiver Laplacian is a generalisation of the graph Laplacian and the existence of non-isomorphic yet isospectral graphs is well-known – the discrete analogue of the question ‘can one hear the shape of a drum’. However, the spectrum of a quiver Laplacian can give more information about a filtered simplicial complex than persistent homology is able to do, as we will now illustrate.

Consider the family of bump functions $\psi_c : \mathbb{R}^2 \rightarrow [0, 1]$ indexed by $c \geq 1$ given by

$$\psi_c(x, y) = \varphi(c\sqrt{x^2 + y^2} + \frac{1}{2} - \frac{1}{2c})$$

where

$$\varphi(r) = \begin{cases} 1 & \text{if } r \leq 0 \\ 1 - \frac{e^{-1/r}}{e^{-1/r} + e^{-1/(1-r)}} & \text{if } 0 < r < 1 \\ 0 & r \geq 1. \end{cases}$$

Figure 6.1 shows ψ_1 and ψ_2 restricted to the unit square.

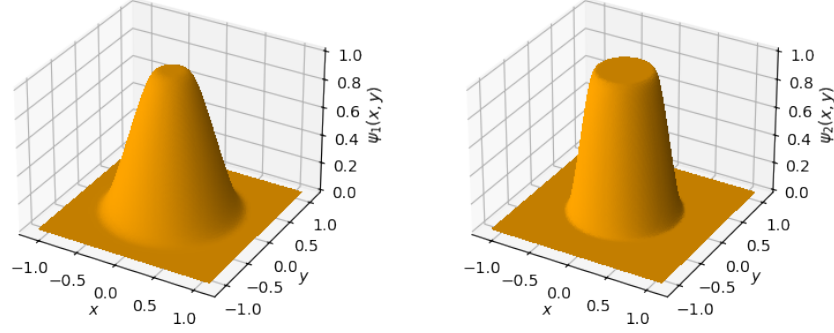


Figure 6.1: The functions ψ_1 and ψ_2 .

Let K be sufficiently fine triangulation of the unit square and $0 = t_1 \leq \dots \leq t_n = 1$ a sufficiently fine discretisation of $[0, 1]$. If we give K the lower-star filtration defined by ψ_c its barcode in degree 1 will consist of a just a single interval $[0, t_n]$ for any value of $c \geq 1$. Thus persistent homology cannot tell these bump functions apart. However, being able to separate these may be desirable in many contexts: if the filtration comes from a greyscale image, for example, we may want to understand how ‘sharp’ maxima and minima are, rather than just their intensity values.

The harmonic persistence modules, on the other hand, will be different. Figure 6.2 shows the harmonic persistence module for the lower-star filtration of ψ_1 on a triangulation of the square $K = [-1.1, 1.1]$. In each subplot of Figure 6.2B, for times $t_i \in [0, 1]$, we have plotted $(|v_\sigma^i|)_{\sigma \in K^1}$ for the unit-norm generator v^i of $\mathcal{H}_1(K_{\varphi_1 \leq t_i})$. For every $t_i \in [0, 1)$ the space $\mathcal{H}_1(K_{\varphi_1 \leq t_i})$ is 1 dimensional, thus the maps for each $t_i \rightarrow t_{i+1}$ are the scalars $\langle v^i, v^{i+1} \rangle_{C_1(K_{\varphi \leq t_{i+1}})}$. These maps depend on c : if the slope of ψ_c is greater, i.e. c is larger, then the change in sublevel sets from $\psi_c \leq t_i$ to $\psi_c \leq t_{i+1}$ is smaller, hence the norm of the map $\mathcal{H}_1(K_{\psi_c \leq t_i}) \rightarrow$

$\mathcal{H}_1(K_{\psi_c \leq t_{i+1}})$ is smaller.

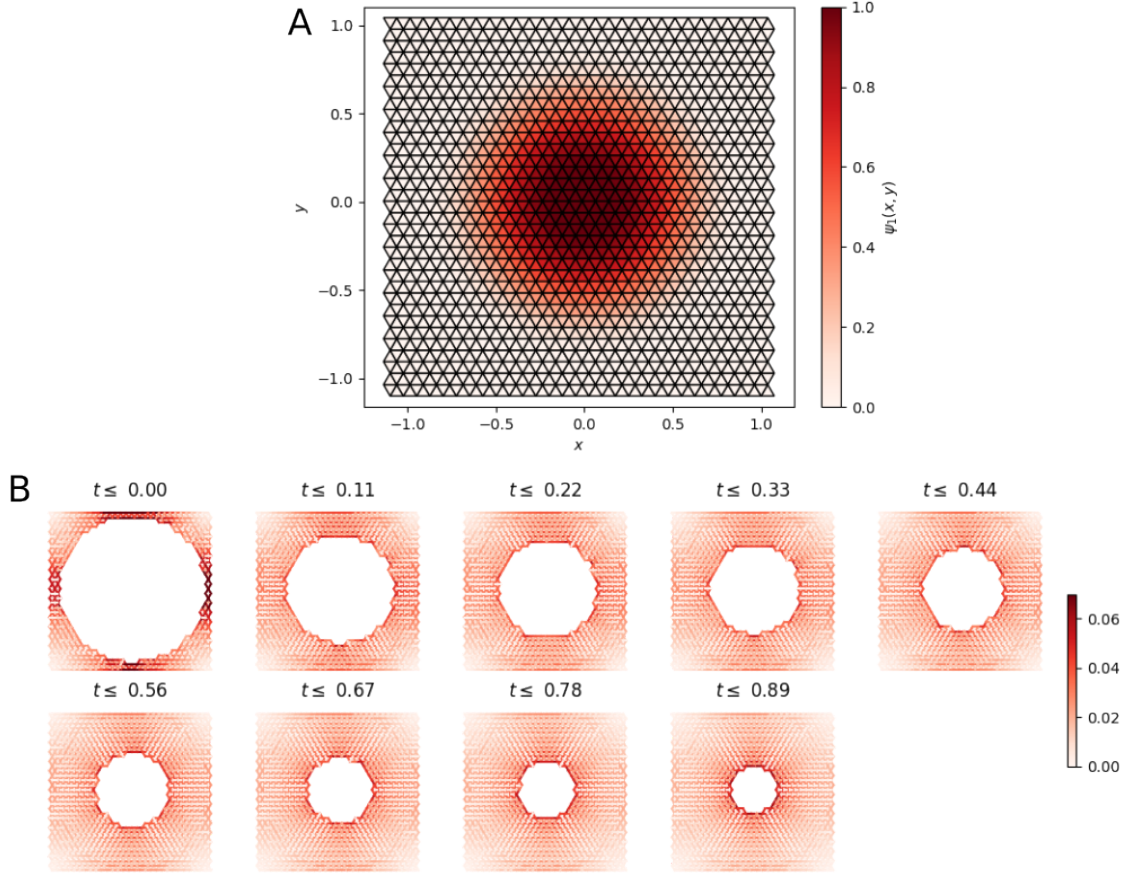


Figure 6.2: Harmonic representatives of $\mathcal{H}_1(K_{\psi_1 \leq t})$. (A) The function ψ_1 induces a filtration on a triangulation of $[-1.1, 1.1]$. (B) Harmonic representatives of the degree 1 homology spaces at filtration t . Plotted is the absolute value of each component on the 1-simplices.

This can be directly seen in the continuous case on $[0, 1]$. The continuous analogue on $[0, a]$ of the degree 0 harmonics of the graph Laplacian of an A_n graph are the solutions to

$$\Delta f = f'' = 0$$

subject to Neumann boundary conditions $f'(0) = f'(1) = 0$. These are just the constant functions on $[0, a]$ hence the space $\mathcal{H}_0([0, a])$ is generated by $g_a(x) = a^{-1/2} \mathbb{1}_{[0, a]}(x)$. Suppose we choose a monotonically increasing differentiable function $f : [0, 1] \rightarrow [0, 1]$ as a filtration on $[0, 1]$. Define $\mathcal{H}_0(t_0 \leq t_0 + t)$ to be the map

$\mathcal{H}_0([0, 1]_{f \leq t_0}) \rightarrow \mathcal{H}_0([0, 1]_{f \leq t_0+t})$ for $t_0, t_0 + t \in (0, 1)$. Computing this yields

$$\begin{aligned} \mathcal{H}_0(t_0 \leq t_0 + t) &= \langle g_{f^{-1}(t_0)}, g_{f^{-1}(t_0+t)} \rangle \\ &= \int_0^1 f^{-1}(t_0)^{-1/2} f^{-1}(t_0 + t)^{-1/2} \mathbb{1}_{[0, f^{-1}(t_0)]}(x) \mathbb{1}_{[0, f^{-1}(t_0+t)]}(x) \, dx \\ &= f^{-1}(t_0)^{1/2} f^{-1}(t_0 + t)^{-1/2} \end{aligned}$$

thus

$$\frac{d}{dt} \mathcal{H}_0(t_0 \leq t_0 + t) = -\frac{1}{2} f^{-1}(t_0)^{\frac{1}{2}} f^{-1}(t_0 + t)^{-\frac{3}{2}} \left[\frac{df}{dt} \Big|_{f^{-1}(t_0+t)} \right]^{-1}.$$

In Figure 6.3 we compare the spectra of the quiver Laplacian of the sublevel set filtration of ψ_c on K for 10 different values of $1 \leq c \leq 2$. In this example, we have discretised the range $[0, 1]$ of ψ_c into 10 equal steps. The first 9 $H_1(K_{\psi_c \leq t_i})$ have rank 1 and the last, where $t_i = 1$ has rank 0. Figure 6.3A shows the 9 eigenvalues associated to each Laplacian, exhibiting that the spectrum of their Laplacians are distinct for different values of c . In Figure 6.3B, we plot the component of the first eigenvector, a global section, in each $H_1(K_{\psi_c \leq t_i})$. Since a lower value of c means the change in the harmonic representatives is greater, the global section changes more rapidly for lower c . Hence this is a collection of filtered complexes for which ordinary persistent homology could not tell apart, was able to be distinguished via the quiver Laplacian.

6.4.1 Multiparameter persistence

The quiver Laplacian is defined for a representation over any quiver representation, thus we can similarly compute the spectra of multiparameter persistence modules. The standard invariant for multiparameter persistence modules is the rank invariant: For a quiver representation $\mathbf{A}_\bullet$ over a poset P the rank invariant [9]

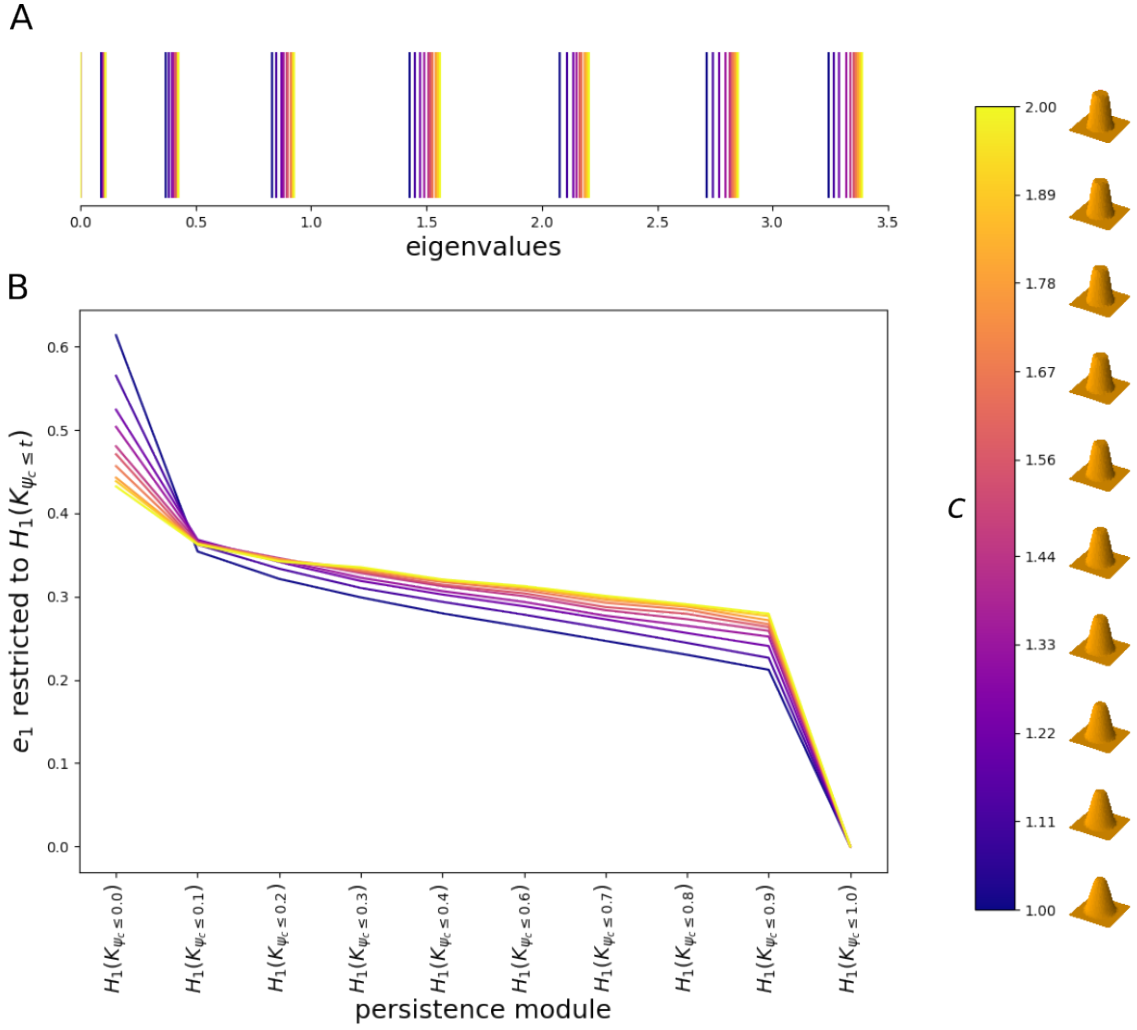


Figure 6.3: Comparison of the spectra of persistence modules. (A) The eigenvalues of the persistence module obtained from each ψ_c , coloured by value of c . (B) The global section for each value of c .

is the collection of $\dim \mathbf{A}_{p \rightarrow q}$ for $p \leq q \in P$. Apart from when $P = \mathbb{N}$ or $\mathbb{R}$, as in ordinary persistence, the rank invariant is not complete. Consider the following two multiparameter persistence modules from [22]:

$$\begin{array}{ccc}
 \mathbb{R} & \longrightarrow & 0 \\
 \uparrow [1 \ 0] & & \uparrow \\
 \mathbb{R}^2 & \longrightarrow & \mathbb{R}
 \end{array}
 \qquad
 \begin{array}{ccc}
 \mathbb{R} & \longrightarrow & 0 \\
 \uparrow [0 \ 1] & & \uparrow \\
 \mathbb{R}^2 & \longrightarrow & \mathbb{R}.
 \end{array}
 \tag{6.1}$$

Both have the same rank invariant yet are not isomorphic. They can be decom-

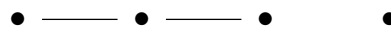
posed (orthogonally) into

$$\begin{array}{ccc} \mathbb{R} & & \mathbb{R} \\ \uparrow & & \uparrow \\ \begin{bmatrix} 1 & 0 \end{bmatrix} & & 1 \\ \mathbb{R}^2 & \xrightarrow{\begin{bmatrix} 1 & 0 \end{bmatrix}} & \mathbb{R} \\ & & \mathbb{R} \end{array} \quad \cong \quad \begin{array}{ccc} \mathbb{R} & & \mathbb{R} \\ \uparrow & & \uparrow \\ 1 & & 1 \\ \mathbb{R} & \xrightarrow{1} & \mathbb{R} \\ & & \mathbb{R} \end{array} \oplus \begin{array}{ccc} \mathbb{R} & & \mathbb{R} \\ \uparrow & & \uparrow \\ 1 & & 1 \\ \mathbb{R} & \xrightarrow{1} & \mathbb{R} \\ & & \mathbb{R} \end{array}$$

and

$$\begin{array}{ccc} \mathbb{R} & & \mathbb{R} \\ \uparrow & & \uparrow \\ \begin{bmatrix} 0 & 1 \end{bmatrix} & & 1 \\ \mathbb{R}^2 & \xrightarrow{\begin{bmatrix} 1 & 0 \end{bmatrix}} & \mathbb{R} \\ & & \mathbb{R} \end{array} \quad \cong \quad \begin{array}{ccc} \mathbb{R} & & \mathbb{R} \\ \uparrow & & \uparrow \\ 1 & & 1 \\ \mathbb{R} & \xrightarrow{1} & \mathbb{R} \\ & & \mathbb{R} \end{array} \oplus \begin{array}{ccc} \mathbb{R} & & \mathbb{R} \\ \uparrow & & \uparrow \\ 1 & & 1 \\ \mathbb{R} & \xrightarrow{1} & \mathbb{R} \\ & & \mathbb{R} \end{array}$$

respectively. Thus the quiver Laplacian of the first is the graph Laplacian of the graph



and so has eigenvalues 0,0,1,3 while the quiver Laplacian of the second is the graph Laplacian of the graph



thus has eigenvalues 0,0,2,2, hence the persistence modules in 6.1 can be distinguished with the quiver Laplacian, but not with the rank invariant.

Appendix A

Appendix

A.1 Removing Edges and Vertices

This appendix extends known results from graphs, e.g. see [91] and [99] respectively, involving eigenvalue interlacings after removing vertices and edges to sheaves over a quiver.

Proposition A.1.1. *Let $\mathcal{A}_\bullet$ be a sheaf over a quiver Q , and let n be its total dimension.*

A.1.1(a) Let $R \subset Q_1$ be a subset of edges of Q . Let Q' be the quiver with edges $Q_1 \setminus R$ and $\mathcal{A}'$ the restriction of $\mathcal{A}$ to Q' . Then

$$\lambda_{k+i}(\mathcal{A}_\bullet) \leq \lambda_{k+i}(\mathcal{A}'_\bullet) \leq \lambda_{k+r+i}(\mathcal{A}_\bullet)$$

for $i = 1, \dots, n - k - r$ where $k = \dim \Gamma(\mathcal{A}_\bullet)$ and $r = \sum_{e \in R} \dim \mathcal{A}_e$.

A.1.1(b) Suppose $V \subset Q_0$ is a subset of vertices of Q . Let Q' be the quiver with vertices $Q_0 \setminus V$ and all edges $e \in Q_1$ such that either $s(e) \in V$ or $t(e) \in V$, and $\mathcal{A}'_\bullet$ the restriction of $\mathcal{A}_\bullet$ to Q' . Then

$$\lambda_i(\mathcal{A}_\bullet) - w_1 \leq \lambda_i(\mathcal{A}'_\bullet) \leq \lambda_{k+i}(\mathcal{A}_\bullet) - w_2$$

for $i = 1, \dots, n - k$ where $k = \sum_{u \in Q_0 \setminus V} \dim \mathcal{A}_u$ and

$$w_1 = \max_{v \in V} \left\{ \sum_{\substack{s(e)=v \\ t(e) \notin V}} \sigma_{\max}(\mathcal{A}_{v \rightarrow e})^2 + \sum_{\substack{t(e)=v \\ s(e) \notin V}} \sigma_{\max}(\mathcal{A}_{e \leftarrow v})^2 \right\},$$

$$w_2 = \min_{v \in V} \left\{ \sum_{\substack{s(e)=v \\ t(e) \notin V}} \sigma_{\min}(\mathcal{A}_{v \rightarrow e})^2 + \sum_{\substack{t(e)=v \\ s(e) \notin V}} \sigma_{\min}(\mathcal{A}_{e \leftarrow v})^2 \right\}.$$

Proof of Proposition A.1.1(a). Define $L_{\mathcal{A}, \text{op}} = B_{\mathcal{A}} B_{\mathcal{A}}^*$. Let $k = \dim \ker L_{\mathcal{A}}$ and $c = \dim \ker L_{\mathcal{A}, \text{op}}$. Then for $i = 1, \dots, n - k$, we have that

$$\lambda_{k+i}(L_{\mathcal{A}}) = \lambda_{c+i}(L_{\mathcal{A}, \text{op}}).$$

Now, removing the set of edges $R \subset Q_1$ corresponds to the $r \times r$ principal submatrix of $L_{\mathcal{A}, \text{op}}$, where $r := \sum_{e \in R} \dim \mathcal{A}_e$. By Theorem A.3.3 we have

$$\lambda_i(L_{\mathcal{A}, \text{op}}) \leq \lambda_i(L_{\mathcal{A}', \text{op}}) \leq \lambda_{i+r}(L_{\mathcal{A}, \text{op}})$$

where $i \in \{1, \dots, m - r\}$ thus

$$\lambda_{k+i}(\mathcal{A}_{\bullet}) \leq \lambda_{k+i}(\mathcal{A}'_{\bullet}) \leq \lambda_{k+r+i}(\mathcal{A}_{\bullet})$$

for $i \in \{1, \dots, n - k - r\}$. □

Proof of Proposition A.1.1(b). The boundary matrix $B_{\mathcal{A}}$ has block form

$$\begin{bmatrix} B_{11} & \mathbf{0} \\ B_{21} & B_{22} \end{bmatrix}$$

where $B_{11} = B_{\mathcal{A}'}$. Here, $B_{\mathcal{A}}$ has row blocks indexed by edges between vertices in V and column blocks indexed by vertices in V . Then Laplacian $L_{\mathcal{A}} = B_{\mathcal{A}}^* B_{\mathcal{A}}$ has

block form

$$\begin{bmatrix} B_{11}^* B_{11} + B_{21}^* B_{21} & B_{21}^* B_{22} \\ B_{22}^* B_{21} & B_{22}^* B_{22} \end{bmatrix}.$$

Define $L_{Q^0 \setminus V} = B_{11}^* B_{11} + B_{21}^* B_{21}$ which is a principal submatrix of $L_{\mathcal{A}}$. Note that $B_{21}^* B_{21}$ is a diagonal matrix: The v, v block is

$$\sum_{\substack{s(e)=v \\ t(e) \notin V}} \mathcal{A}_{v \rightarrow e}^* \mathcal{A}_{v \rightarrow e} + \sum_{\substack{t(e)=v \\ s(e) \notin V}} \mathcal{A}_{e \leftarrow v}^* \mathcal{A}_{e \leftarrow v}.$$

We now have that

$$\begin{aligned} \lambda_i(\mathcal{A}'_{\bullet}) &= \lambda_i(B_{11}^* B_{11}) = \lambda_i(L_{Q^0 \setminus V} - B_{21}^* B_{21}) \leq \lambda_i(L_{Q^0 \setminus V}) + \lambda_{n-k}(-B_{21}^* B_{21}) \\ &= \lambda_i(L_{Q^0 \setminus V}) - \lambda_1(B_{21}^* B_{21}) \\ &= \lambda_i(L_{Q^0 \setminus V}) - \min_{v \in V} \lambda_1 \left(\sum_{\substack{s(e)=v \\ t(e) \notin V}} \mathcal{A}_{v \rightarrow e}^* \mathcal{A}_{v \rightarrow e} + \sum_{\substack{t(e)=v \\ s(e) \notin V}} \mathcal{A}_{e \leftarrow v}^* \mathcal{A}_{e \leftarrow v} \right) \\ &\leq \lambda_i(L_{Q^0 \setminus V}) - \min_{v \in V} \left\{ \sum_{\substack{s(e)=v \\ t(e) \notin V}} \lambda_1(\mathcal{A}_{v \rightarrow e}^* \mathcal{A}_{v \rightarrow e}) + \sum_{\substack{t(e)=v \\ s(e) \notin V}} \lambda_1(\mathcal{A}_{e \leftarrow v}^* \mathcal{A}_{e \leftarrow v}) \right\} \\ &= \lambda_i(L_{Q^0 \setminus V}) - \min_{v \in V} \left\{ \sum_{\substack{s(e)=v \\ t(e) \notin V}} \sigma_{\min}(\mathcal{A}_{v \rightarrow e})^2 + \sum_{\substack{t(e)=v \\ s(e) \notin V}} \sigma_{\min}(\mathcal{A}_{e \leftarrow v})^2 \right\} \\ &\leq \lambda_{i+k}(L_{\mathcal{A}}) - \min_{v \in V} \left\{ \sum_{\substack{s(e)=v \\ t(e) \notin V}} \sigma_{\min}(\mathcal{A}_{v \rightarrow e})^2 + \sum_{\substack{t(e)=v \\ s(e) \notin V}} \sigma_{\min}(\mathcal{A}_{e \leftarrow v})^2 \right\}. \end{aligned}$$

The first two inequalities are from Theorem A.3.4 and the final inequality is due to Theorem A.3.3. The right inequality is proven similarly. $\square$

A.2 Naturalisers

The following appendix describes a universal framework for constructing the combined representations as in Chapter 4 but over a general category, via the notion of

‘naturalising’ a transformation between functors.

Let Q be a quiver with vertex set Q_0 , edge set Q_1 and maps $s, t : Q_1 \rightarrow Q_0$ delineating the source and target vertex of each edge. A *path* of Q is any finite sequence of edges

$$p = (e_1, e_2, \dots, e_k) \quad \text{with} \quad s(e_i) = t(e_{i-1}) \text{ for all } i.$$

We may extend s and t from edges to such paths by setting $s(p) := s(e_1)$ and $t(p) := t(e_k)$. By convention, we also assume that for each vertex $v \in Q_0$ there is a distinguished *empty* path 1_v with $s(1_v) = v = t(1_v)$.

Definition A.2.1. The **free category** $\mathbf{Fr}(Q)$ of a quiver Q is defined as follows: its object set is Q_0 , and morphisms $v \rightarrow w$ are all paths p with $s(p) = v$ and $t(p) = w$. These morphisms are composed via concatenation of paths, with 1_v serving as the identity morphism for each object v .

Let $\mathbf{A}_\bullet$ be any representation of Q valued in an arbitrary category $\mathcal{C}$ — namely, each vertex $v \in Q_0$ is assigned an object $\mathbf{A}_v$ of $\mathcal{C}$, while each edge $e \in Q_1$ is assigned a morphism $\mathbf{A}_e : \mathbf{A}_{s(e)} \rightarrow \mathbf{A}_{t(e)}$ of $\mathcal{C}$. It is known that any such representation canonically induces a functor $\mathbf{Fr}(\mathbf{A}_\bullet) : \mathbf{Fr}(Q) \rightarrow \mathcal{C}$, and a morphism $\eta_\bullet : \mathbf{A}_\bullet \rightarrow \mathbf{A}'_\bullet$ of Q -representations similarly induces a natural transformation $\mathbf{Fr}(\eta_\bullet) : \mathbf{Fr}(\mathbf{A}_\bullet) \Rightarrow \mathbf{Fr}(\mathbf{A}'_\bullet)$ of functors — see [53, Ch II.7] for details.

Remark A.2.2. The setting of immediate interest to us here is the following special case of the general free category framework described above. As mentioned in the prologue to Proposition 4.2.7, we begin with a feature selector $\mathcal{S}$ in the sense of Definition 4.2.1. Given a cover $\mathcal{U}$ of the underlying set X and an inner product structure on the function space X^* , for every input subset $\mathcal{F} \subset X^*$ we construct two representations $\mathbf{S}_\bullet^{\mathcal{F}, \mathcal{U}}$ and $\mathbf{V}(\mathcal{F})_\bullet$ of the quiver $Q^\mathcal{U}$. Each of these induces a functor from the free category $\mathbf{Fr}(Q^\mathcal{U})$ valued in the category $\mathbf{Hilb}(\mathbb{k})$ of Hil-

bert spaces over $\mathbb{k}$. By Definition 4.2.6, our feature selector $\mathcal{S}$ is locally compatible (with $\mathcal{U}$, on $\mathcal{F}$) if and only if the maps $\iota_{\bullet}^{\mathcal{F}}$ induce a morphism of quiver representations, which in turn occurs only if $\mathbf{Fr}(\iota_{\bullet}^{\mathcal{S}})$ constitutes a natural transformation $\mathbf{Fr}(\mathbf{S}_{\bullet}^{\mathcal{F}, \mathcal{U}}) \Rightarrow \mathbf{Fr}(\mathbf{V}(\mathcal{F})_{\bullet})$ in the category of functors $\mathbf{Fr}(Q^{\mathcal{U}}) \rightarrow \mathbf{Hilb}(\mathbb{k})$.

Having related our compatibility criterion from Definition 4.2.6 to the existence of a natural transformation $\mathbf{Fr}(\mathbf{S}_{\bullet}^{\mathcal{F}, \mathcal{U}}) \Rightarrow \mathbf{Fr}(\mathbf{V}(\mathcal{F})_{\bullet})$, we now seek the optimal natural transformation supported by the collection of maps $\eta_{\bullet}^{\mathcal{S}}$. This optimality takes the form of a general universal property, which applies to the following class of map-collections.

Definition A.2.3. Let $\mathcal{I}$ and $\mathcal{C}$ be a pair of categories, and $F, G : \mathcal{I} \rightarrow \mathcal{C}$ a pair of functors. A **transformation** from F to G is any collection τ of morphisms in $\mathcal{C}$ of the form $\tau_i : F_i \rightarrow G_i$, indexed by objects i of $\mathcal{I}$ (and subject to no further constraints).

These collections τ have also been called *unnatural transformations* in [1, Example 13.1] and *assignments* in [12, Definition 3.3]. We avoid the former terminology on account of not wanting to make natural transformations special cases of unnatural ones, and the latter terminology since the word assignment is already quite overloaded. In any event, we recall that a transformation τ constitutes a *natural* transformation of functors $F \Rightarrow G$ whenever the diagram

$$\begin{array}{ccc} F_i & \xrightarrow{\tau_i} & G_i \\ F_{\varphi} \downarrow & & \downarrow G_{\varphi} \\ F_j & \xrightarrow{\tau_j} & G_j \end{array}$$

commutes in $\mathcal{C}$ for each morphism $\varphi : i \rightarrow j$ in $\mathcal{I}$. The following result is purely category-theoretic; it describes a wide variety of scenarios where an optimal natural transformation may be extracted from an arbitrary τ . (We expect both the

statement and its proof to be familiar to experts, but we were unable to locate an account in the literature and have therefore included it here for completeness.)

Lemma A.2.4. *Assume that $F, G : \mathcal{I} \rightarrow \mathcal{C}$ are functors from a category $\mathcal{I}$ to a category $\mathcal{C}$, and that τ is a transformation from F to G . If either*

1. *$\mathcal{C}$ is complete (i.e., admits all limits), or*
2. *$\mathcal{I}$ is finite and $\mathcal{C}$ is finitely complete,*

then there exists a universal triple (E, ϵ, δ) consisting of a functor $E : \mathcal{I} \rightarrow \mathcal{C}$ along with natural transformations $\epsilon : E \Rightarrow F$ and $\delta : E \Rightarrow G$ so that the diagram

$$\begin{array}{ccc}
 & E_i & \\
 \epsilon_i \swarrow & & \searrow \delta_i \\
 F_i & \xrightarrow{\tau_i} & G_i
 \end{array} \tag{A.1}$$

commutes in $\mathcal{C}$ for every object i in $\mathcal{I}$.

We call the universal triple (E, ϵ, δ) a **naturaliser** for τ — if τ itself is a natural transformation, then up to isomorphism its naturaliser is simply (F, id, τ) . In general, the E_i will be subobjects of F_i and $\delta : E \Rightarrow G$ will serve as the optimal natural transformation which approximates τ . This optimality takes the form of the following universal property. Consider any other triple (E', ϵ', δ') such that diagrams analogous to (A.1) commute for every object i of $\mathcal{I}$. There is a unique natural transformation $\zeta : E' \Rightarrow E$ which makes this diagram commute:

$$\begin{array}{ccccc}
 & E' & & & \\
 \epsilon' \swarrow & \Downarrow \zeta & \searrow \delta' & & \\
 F & \xrightarrow{\epsilon} E & \xrightarrow{\delta} G & &
 \end{array} \tag{A.2}$$

In other words, *every method of upgrading τ to a natural transformation necessarily factors through its naturaliser*. We now assume the setting and notation of Lemma A.2.4, and dedicate the next subsection to its proof.

A.2.1 Proof of Lemma A.2.4

The following construction plays an essential role in our argument.

Definition A.2.5. The **floret** at a given object i of $\mathcal{J}$ is a functor $\mathcal{B}^i : \mathcal{J}^i \rightarrow \mathcal{C}$ whose source category $\mathcal{J}^i$ is defined as follows:

1. the set of objects is the disjoint union

$$\{i\} \sqcup \{j_\psi \mid \psi : i \rightarrow j \text{ in } \mathcal{J}\}.$$

In other words, there is a distinguished copy of i , which we label i_0 henceforth, and an additional object denoted j_ψ for every morphism $\psi : i \rightarrow j$ in $\mathcal{J}$.

2. for each $\psi : i \rightarrow j$ in $\mathcal{J}$, there are two morphisms in $\mathcal{J}^i$ which will be denoted $\psi_{DL}, \psi_{LD} : i_0 \rightarrow j_\psi$; other than these, $\mathcal{J}^i$ only has identity morphisms.

The floret $\mathcal{B}^i$ acts on objects as

$$i_0 \mapsto F_i \quad \text{and} \quad j_\psi \mapsto G_j.$$

On morphisms, we have

$$\varphi_{DL} \mapsto \tau_j \circ F_\varphi \quad \text{and} \quad \varphi_{LD} \mapsto G_\varphi \circ \tau_i.$$

We note in passing that there are no nontrivial composites of morphisms in $\mathcal{J}^i$, so we are not required to spell out any composition laws. Moreover, each $\mathcal{J}^i$ is finite

when $\mathcal{I}$ is finite; the completeness assumptions from the statement of Lemma A.2.4 therefore guarantee that $E_i := \lim \mathcal{B}^i$ exists.

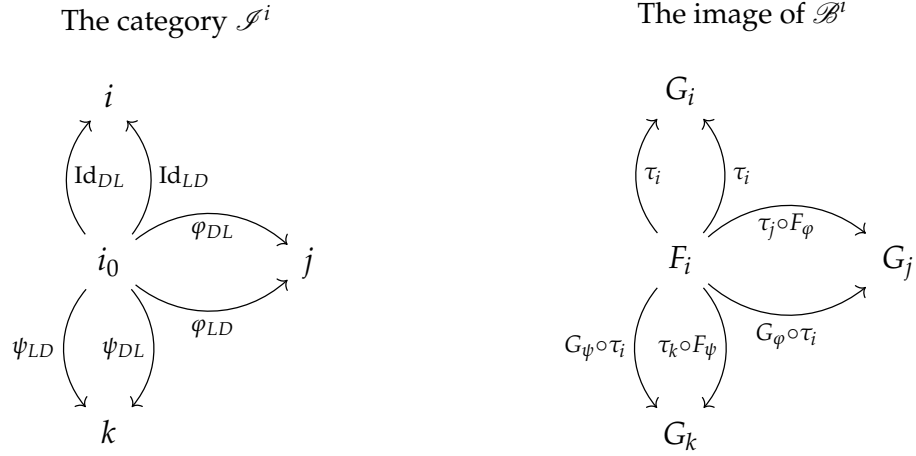


Figure A.1: An illustration of the floret functor at $i \in \mathcal{I}$, assuming that the only non-identity morphisms in $\mathcal{I}$ sourced at i are $\varphi : i \rightarrow j$ and $\psi : i \rightarrow k$

Remark A.2.6. Since E_i is a cone over $\mathcal{B}^i$ by definition, there exists

- a morphism $u_i : E_i \rightarrow F_i$, and
- for each $\psi : i \rightarrow k$ in $\mathcal{I}$, a morphism $u_\psi : E_i \rightarrow G_k$,

so that the following diagram $\mathbf{D}(\psi)$ commutes in $\mathcal{C}$:

$$\begin{array}{ccc}
 & E_i & \\
 u_i \swarrow & & \searrow u_\psi \\
 F_i & \xrightarrow{G_\psi \circ \tau_i} & G_k \\
 & \xleftarrow{\tau_k \circ F_\psi} &
 \end{array} \tag{A.3}$$

The morphisms $u_\bullet$ and commuting diagrams $\mathbf{D}(\bullet)$ will help establish the following intermediate result.

Proposition A.2.7. *The assignment $i \mapsto E_i$ extends to a functor $\mathcal{I} \rightarrow \mathcal{C}$.*

Proof. It suffices to show that given any morphism $\varphi : i \rightarrow j$ in $\mathcal{I}$, the object E_i forms a cone over the floret $\mathcal{B}^j$ in $\mathcal{C}$ — if this cone property holds, then the universality of limits yields the desired morphism $E_\varphi : E_i \rightarrow E_j$ and guarantees associativity of composition. Now given $\varphi : i \rightarrow j$ in $\mathcal{I}$, let v_j be the composite $F_\varphi \circ u_i : E_i \rightarrow F_j$ in $\mathcal{C}$. Similarly, for any morphism $\psi : j \rightarrow \ell$ (where ℓ is an arbitrary object of $\mathcal{I}$), the second assertion of Remark A.2.6 guarantees the existence of $u_{\psi \circ \varphi} : E_i \rightarrow G_\ell$, which we abbreviate to v_ψ . The collection of morphisms

$$\{v_j : E_i \rightarrow F_j\} \cup \{v_\psi : E_i \rightarrow G_\ell \mid \psi : j \rightarrow \ell \text{ in } \mathcal{I}\} \quad (\text{A.4})$$

now gives E_i the desired structure of a cone over the floret $\mathcal{B}^j$, as shown below. For each morphism $\psi : j \rightarrow \ell$ in $\mathcal{I}$, we have

$$\begin{aligned} (\tau_\ell \circ F_\psi) \circ v_j &= \tau_\ell \circ (F_\psi \circ v_i) && \text{by associativity,} \\ &= \tau_\ell \circ (F_\psi \circ F_\varphi \circ u_i) && \text{by definition of } v_j, \\ &= \tau_\ell \circ (F_{\psi \circ \varphi} \circ u_i) && \text{since } F \text{ is a functor,} \\ &= u_{\psi \circ \varphi} && \text{since } \mathbf{D}(\psi \circ \varphi) \text{ commutes,} \\ &= v_\psi && \text{by definition of } v_\psi. \end{aligned}$$

And similarly,

$$\begin{aligned} (G_\psi \circ \tau_j) \circ v_j &= (G_\psi \circ \tau_j) \circ (F_\varphi \circ u_i) && \text{by definition of } v_j, \\ &= G_\psi \circ ((\tau_j \circ F_\varphi) \circ u_i) && \text{by associativity,} \\ &= G_\psi \circ ((G_\varphi \circ \tau_i) \circ u_i) && \text{since } \mathbf{D}(\varphi) \text{ commutes,} \\ &= (G_{\psi \circ \varphi} \circ \tau_i) \circ u_i && \text{functoriality of } G \text{ plus associativity,} \\ &= u_{\psi \circ \varphi} && \text{since } \mathbf{D}(\psi \circ \varphi) \text{ commutes,} \\ &= v_\psi && \text{by definition of } v_\psi, \end{aligned}$$

as desired. □

The next step in our proof of Lemma A.2.4 is to describe the natural transformations $\epsilon : E \Rightarrow F$ and $\delta : E \Rightarrow G$. These are provisionally defined as follows: for each object $i \in \mathcal{I}$, let $\epsilon_i : E_i \rightarrow F_i$ be the morphism u_i from Remark A.2.6, and let $\delta_i : E_i \rightarrow G_i$ be the composite $\tau_i \circ u_i$.

Proposition A.2.8. *The morphisms $\{\epsilon_i\}$ prescribe a natural transformation $E \Rightarrow F$ of functors $\mathcal{I} \rightarrow \mathcal{C}$; and similarly, the morphisms $\{\delta_i\}$ form a natural transformation $E \Rightarrow G$.*

Proof. Fix a morphism $\varphi : i \rightarrow j$ in $\mathcal{I}$. To establish the naturality of ϵ , it suffices by definition to show that the following diagram commutes in $\mathcal{C}$:

$$\begin{array}{ccc} E_i & \xrightarrow{\epsilon_i} & F_i \\ E_\varphi \downarrow & & \downarrow F_\varphi \\ E_j & \xrightarrow{\epsilon_j} & F_j \end{array}$$

As established in the proof of Proposition A.2.7, the object E_i forms a cone in $\mathcal{C}$ over the floret $\mathcal{B}^j$ via the morphisms $v_\bullet$ from (A.4). Both composites $F_\varphi \circ \epsilon_i$ and $\epsilon_j \circ E_\varphi$ are readily seen to equal $v_j : E_i \rightarrow F_j$, and hence each other — the former equals v_j by definition, while the latter equals v_j by the universality of E_j as a cone over $\mathcal{B}^j$.

Turning now to δ , we examine the relevant diagram for $\varphi : i \rightarrow j$:

$$\begin{array}{ccc} E_i & \xrightarrow{\delta_i} & G_i \\ E_\varphi \downarrow & & \downarrow G_\varphi \\ E_j & \xrightarrow{\delta_j} & G_j \end{array}$$

A direct calculation shows that this diagram commutes:

$$\begin{aligned}
\delta_j \circ E_\varphi &= (\tau_j \circ u_j) \circ E_\varphi && \text{by definition of } \delta_j \\
&= u_{\text{Id}} \circ E_\varphi && \text{by (A.3) with } \psi = \text{Id} : j \rightarrow j \\
&= u_\varphi && \text{since } E_\varphi : E_i \rightarrow E_j \text{ is the universal cone map} \\
&= (G_\varphi \circ \tau_i) \circ u_i && \text{by (A.3) with } \psi = \varphi : i \rightarrow j \\
&= G_\varphi \circ \delta_i && \text{by associativity and definition of } \delta_i
\end{aligned}$$

□

The following result concludes our proof of Lemma A.2.4.

Proposition A.2.9. *The naturaliser (E, ϵ, δ) of τ is universal in the following sense. Consider any triple (E', ϵ', δ') which make the diagram analogous to (A.3) commute for each object i of $\mathcal{I}$. There exists a unique natural transformation $\xi' : E' \Rightarrow E$ for which (A.1) commutes.*

Proof. It suffices to show that E'_i is a cone over the floret $\mathcal{B}^i$ for each object i of $\mathcal{I}$, so that the existence of the desired natural transformation ξ is forced by the universal property of limits. To this end, note that we have a morphism $u'_i : E'_i \rightarrow F_i$ in $\mathcal{C}$ given by $u'_i = \epsilon'_i$. For each morphism $\varphi : i \rightarrow j$ we also have a morphism $u'_\varphi : E'_i \rightarrow G_j$ given by $G_\varphi \circ \delta'_i$. Two elementary calculations show that these $u'_\bullet$ maps endow E'_i with the structure of a cone over $\mathcal{B}^i$ in $\mathcal{C}$, as desired. □

Remark A.2.10. In the special case where the codomain $\mathcal{C}$ of the functors E and F is an additive category with kernels and pullbacks, we have

$$E_i \simeq \bigcap_{\psi : i \rightarrow j} \ker (G_\psi \circ \eta_i - \eta_j \circ F_\psi),$$

where the intersection ranges over all $\psi : i \rightarrow j$ in $\mathcal{I}$. In this case, it becomes pos-

sible to compute the naturaliser of a transformation τ from E to F rather explicitly as a sub-object $\epsilon_i : E_i \hookrightarrow F_i$ for each object i of $\mathcal{S}$.

A.3 Classical Results on Eigenvalues of Hermitian Matrices

This appendix contains classical results on the eigenvalues of Hermitian matrices.

In the following, for an $n \times n$ Hermitian matrix A , let $\lambda_1(A), \dots, \lambda_n(A)$ be the eigenvalues ordered in increasing fashion.

Theorem A.3.1 (Rayleigh, see [41] Theorem 4.2.2). *Let A be a Hermitian matrix. Then for integers $1 \leq i_1, \dots, i_k \leq n$ let $x_{i_1}, \dots, x_{i_k}$ be orthonormal, where x_{i_j} is an eigenvector corresponding to $\lambda_{i_j}(A)$. Define $S = \text{span} \{x_{i_1}, \dots, x_{i_k}\}$. Then for any $x \in S$*

$$\lambda_{i_1}(A) \leq \frac{x^* A x}{\|x\|^2} \leq \lambda_{i_k}(A).$$

Theorem A.3.2 (Courant-Fischer, see [41] Theorem 4.2.6). *Let A be Hermitian matrix. Then*

$$\lambda_k(A) = \min_{\dim V=k} \left[\max_{x \in V \setminus 0} \frac{\langle x, Ax \rangle}{\langle x, x \rangle} \right].$$

Theorem A.3.3 (Cauchy's interlacing theorem, see [41] Theorem 4.3.17). *Let A be a Hermitian matrix of size $n \times n$ and B an $m \times m$ principal submatrix of A . Then*

$$\lambda_i(A) \leq \lambda_i(B) \leq \lambda_{i+n-m}(A)$$

for $i = 1, \dots, m$.

Theorem A.3.4 (Weyl's inequalities, see [41] Theorem 4.3.1). *For A, B Hermitian matrices of size $n \times n$ the following inequalities hold:*

$$\lambda_i(A + B) \leq \lambda_{i+j}(A) + \lambda_{n-j}(B)$$

for $j = 0, \dots, n-1$ and

$$\lambda_i(A+B) \geq \lambda_{i-j+1}(A) + \lambda_j(B)$$

for $j = 1, \dots, i$.

Theorem A.3.5 (See [78] Theorem 5). *For H a Hermitian matrix of dimension n given in block form as*

$$H = \begin{bmatrix} A & B^* \\ B & D \end{bmatrix}$$

where D is non-singular let $H/D = A - BD^{-1}B^$ be the Schur complement of D . Then the eigenvalues of H and H/D are interlaced:*

$$\lambda_i(H) \leq \lambda_i(H/D) \leq \lambda_{i+r}(H)$$

for $i = 1, \dots, n-r$ where r is the rank of D .

Bibliography

- [1] Benedikt Ahrens et al. *The Univalence Principle*. 29th Aug. 2022. arXiv: [2102.06275](https://arxiv.org/abs/2102.06275) [cs, math]. URL: <http://arxiv.org/abs/2102.06275> (visited on 19/07/2024). Pre-published.
- [2] Marcus Alvarez et al. “Enhancing Droplet-Based Single-Nucleus RNA-seq Resolution Using the Semi-Supervised Machine Learning Classifier DIEM”. In: *Scientific reports* 10.1 (2020), pp. 1–16.
- [3] Babak Ayazifar. “Graph Spectra and Modal Dynamics of Oscillatory Networks”. Thesis. Massachusetts Institute of Technology, 2002. URL: <https://dspace.mit.edu/handle/1721.1/16913> (visited on 19/07/2024).
- [4] M Barahona. *The Stability of a Graph Partition*. URL: https://www.ma.imperial.ac.uk/~mpbara/Partition_Stability/ (visited on 23/05/2022).
- [5] Saugata Basu and Nathanael Cox. *Harmonic Persistent Homology*. 28th Aug. 2022. DOI: [10.48550/arXiv.2105.15170](https://doi.org/10.48550/arXiv.2105.15170). arXiv: [2105.15170](https://arxiv.org/abs/2105.15170) [math]. URL: <http://arxiv.org/abs/2105.15170> (visited on 02/06/2024). Pre-published.
- [6] Mikhail Belkin and Partha Niyogi. “Laplacian Eigenmaps for Dimensionality Reduction and Data Representation”. In: *Neural computation* 15.6 (2003), pp. 1373–1396.

- [7] Sabrina Bianca Bennstein et al. “Umbilical Cord Blood-Derived ILC1-like Cells Constitute a Novel Precursor for Mature KIR+ NKG2A-NK Cells”. In: *Elife* 9 (2020), e55232.
- [8] Jochem H Bernink et al. “Human Type 1 Innate Lymphoid Cells Accumulate in Inflamed Mucosal Tissues”. In: *Nature immunology* 14.3 (2013), pp. 221–229.
- [9] Magnus Bakke Botnan and Michael Lesnick. *An Introduction to Multiparameter Persistence*. 13th Mar. 2023. DOI: [10.48550/arXiv.2203.14289](https://doi.org/10.48550/arXiv.2203.14289). arXiv: [2203.14289](https://arxiv.org/abs/2203.14289) [cs, math]. URL: <http://arxiv.org/abs/2203.14289> (visited on 19/07/2024). Pre-published.
- [10] Jason D Buenrostro et al. “Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-binding Proteins and Nucleosome Position”. In: *Nature methods* 10.12 (2013), pp. 1213–1218.
- [11] Daniela Calvetti, Lothar Reichel and Danny Chris Sorensen. “An Implicitly Restarted Lanczos Method for Large Symmetric Eigenvalue Problems”. In: *Electronic Transactions on Numerical Analysis* 2.1 (1994), p. 21.
- [12] Erin W. Chambers et al. *Bounding the Interleaving Distance for Mapper Graphs with a Loss Function*. 19th Mar. 2024. arXiv: [2307.15130](https://arxiv.org/abs/2307.15130) [cs, math]. URL: <http://arxiv.org/abs/2307.15130> (visited on 19/07/2024). Pre-published.
- [13] Fan RK Chung. “The Laplacian of a Hypergraph.” In: *Expanding graphs* 10 (1992), pp. 21–36.
- [14] Fan RK Chung and Fan Chung Graham. *Spectral Graph Theory*. 92. American Mathematical Soc., 1997.
- [15] Lyora A Cohen et al. “Serum Ferritin Is Derived Primarily from Macrophages through a Nonclassical Secretory Pathway”. In: *Blood, The Journal of the American Society of Hematology* 116.9 (2010), pp. 1574–1584.

- [16] Geoffrey Cooper and Kenneth Adams. *The Cell: A Molecular Approach*. Oxford University Press, 2022.
- [17] Sarah Cormican and Matthew D Griffin. “Human Monocyte Subset Distinctions and Function: Insights from Gene Expression Analysis”. In: *Frontiers in Immunology* 11 (2020), p. 1070.
- [18] Adeline Crinier et al. “Single-Cell Profiling Reveals the Trajectories of Natural Killer Cell Differentiation in Bone Marrow and a Stress Signature Induced by Acute Myeloid Leukemia”. In: *Cellular & molecular immunology* 18.5 (2021), pp. 1290–1304.
- [19] 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 (2013), pp. 150–163. DOI: [10 . 1109 / TCSI . 2012 . 2215780](https://doi.org/10.1109/TCSI.2012.2215780).
- [20] Florian Dorfler and Francesco Bullo. “Synchronization of Power Networks: Network Reduction and Effective Resistance”. In: *IFAC Proceedings Volumes* 43.19 (2010), pp. 197–202.
- [21] Samuel Eilenberg and Norman Steenrod. *Foundations of Algebraic Topology*. Vol. 2193. Princeton University Press, 2015.
- [22] Marc Fersztand et al. *Harder-Narasimhan Filtrations of Persistence Modules*. 28th Mar. 2023. DOI: [10 . 48550 / arXiv . 2303 . 16075](https://doi.org/10.48550/arXiv.2303.16075). arXiv: [2303 . 16075 \[math\]](https://arxiv.org/abs/2303.16075). URL: <http://arxiv.org/abs/2303.16075> (visited on 19/07/2024). Pre-published.
- [23] Martin Fischer et al. “Integration of TP53, DREAM, MMB-FOXO1 and RB-E2F Target Gene Analyses Identifies Cell Cycle Gene Regulatory Networks”. In: *Nucleic acids research* 44.13 (2016), pp. 6070–6086.

- [24] Peter Gabriel. “Unzerlegbare Darstellungen I”. In: *manuscripta mathematica* 6.1 (1st Mar. 1972), pp. 71–103. ISSN: 1432-1785. DOI: [10.1007/BF01298413](https://doi.org/10.1007/BF01298413). URL: <https://doi.org/10.1007/BF01298413> (visited on 19/06/2024).
- [25] Tingran Gao, Jacek Brodzki and Sayan Mukherjee. “The Geometry of Synchronization Problems and Learning Group Actions”. In: *Discrete & Computational Geometry* 65.1 (2021), pp. 150–211.
- [26] Zhimin Geng et al. “Altered Monocyte Subsets in Kawasaki Disease Revealed by Single-Cell RNA-sequencing”. In: *Journal of Inflammation Research* 14 (2021), p. 885.
- [27] 10X Genomics. *10X Peripheral Blood Mononuclear Cells (PBMC) Data*. 1st June 2022. URL: https://cf.10xgenomics.com/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz (visited on 01/06/2022).
- [28] Robert Ghrist. “Barcodes: The Persistent Topology of Data”. In: *Bulletin of the American Mathematical Society* 45.1 (2008), pp. 61–75.
- [29] Robert Ghrist and Hans Riess. “Cellular Sheaves of Lattices and the Tarski Laplacian”. 2020. arXiv: [2007.04099](https://arxiv.org/abs/2007.04099).
- [30] Timothy E Goldberg. “Combinatorial Laplacians of Simplicial Complexes”. Senior thesis. Bard College, 2002.
- [31] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. 3rd ed. Johns Hopkins Studies in the Mathematical Sciences. Baltimore: Johns Hopkins University Press, 1996. 694 pp. ISBN: 978-0-8018-5413-2 978-0-8018-5414-9.
- [32] 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. DOI: [10.1371/journal.pcbi.1007509](https://doi.org/10.1371/journal.pcbi.1007509). URL: <https://journals.plos>.

- [org/ploscompbiol/article?id=10.1371/journal.pcbi.1007509](https://doi.org/10.1371/journal.pcbi.1007509) (visited on 20/10/2021).
- [33] Jiao Gu, Bobo Hua and Shiping Liu. “Spectral Distances on Graphs”. In: *Discrete Applied Mathematics* 190 (2015), pp. 56–74.
 - [34] Isabelle Guyon et al. “Gene Selection for Cancer Classification Using Support Vector Machines”. In: *Machine learning* 46 (2002), pp. 389–422.
 - [35] C Hafemeister and R Satija. *Using Sctransform in Seurat*. URL: https://satijalab.org/seurat/articles/sctransform_vignette.html (visited on 23/05/2022).
 - [36] Christoph Hafemeister and Rahul Satija. “Normalization and Variance Stabilization of Single-Cell RNA-seq Data Using Regularized Negative Binomial Regression”. In: *Genome biology* 20.1 (2019), pp. 1–15.
 - [37] Jakob Hansen and Robert Ghrist. “Toward a Spectral Theory of Cellular Sheaves”. In: *Journal of Applied and Computational Topology* 3.4 (2019).
 - [38] Yuhan Hao et al. “Integrated Analysis of Multimodal Single-Cell Data”. In: *Cell* (2021). DOI: [10.1016/j.cell.2021.04.048](https://doi.org/10.1016/j.cell.2021.04.048). URL: <https://doi.org/10.1016/j.cell.2021.04.048>.
 - [39] Xiaofei He, Deng Cai and Partha Niyogi. “Laplacian Score for Feature Selection”. In: *Advances in neural information processing systems* 18 (2005), pp. 1–8.
 - [40] Renee S. Hoekzema et al. “Multiscale Methods for Signal Selection in Single-Cell Data”. In: *Entropy* 24.8 (8 Aug. 2022), p. 1116. ISSN: 1099-4300. DOI: [10.3390/e24081116](https://www.mdpi.com/1099-4300/24/8/1116). URL: <https://www.mdpi.com/1099-4300/24/8/1116> (visited on 05/07/2024).

- [41] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Second edition, corrected reprint. New York, NY: Cambridge University Press, 2017. 643 pp. ISBN: 978-0-521-54823-6 978-0-521-83940-2.
- [42] Zhicheng Ji and Hongkai Ji. "TSCAN: Pseudo-time Reconstruction and Evaluation in Single-Cell RNA-seq Analysis". In: *Nucleic Acids Research* 44.13 (27th July 2016), e117. ISSN: 0305-1048. DOI: [10 . 1093 / nar / gkw430](https://doi.org/10.1093/nar/gkw430). URL: <https://doi.org/10.1093/nar/gkw430> (visited on 18/05/2022).
- [43] I T Jolliffe. *Principal Components Analysis, 2nd Ed.* Springer, 2002.
- [44] Cliff A Joslyn et al. "A Sheaf Theoretical Approach to Uncertainty Quantification of Heterogeneous Geolocation Information". In: *Sensors* 20.12 (2020), p. 3418.
- [45] Mark Kac. "Can One Hear the Shape of a Drum?" In: *The American Mathematical Monthly* 73.4 (Apr. 1966), p. 1. ISSN: 00029890. DOI: [10 . 2307 / 2313748](https://doi.org/10.2307/2313748). JSTOR: [2313748](https://www.jstor.org/stable/2313748). URL: <https://www.jstor.org/stable/2313748?origin=crossref> (visited on 13/09/2024).
- [46] Diether Lambrechts et al. "Phenotype Molding of Stromal Cells in the Lung Tumor Microenvironment". In: *Nature medicine* 24.8 (2018), pp. 1277–1289.
- [47] Cornelius Lanczos. "An Iteration Method for the Solution of the Eigenvalue Problem of Linear Differential and Integral Operators". In: *United States Governm. Press Office Los Angeles, CA* (1950).
- [48] Robin D Lee et al. "Single-Cell Analysis Identifies Dynamic Gene Expression Networks That Govern B Cell Development and Transformation". In: *Nature communications* 12.1 (2021), pp. 1–16.
- [49] Rich Lehoucq et al. "ARPACK-NG: Large Scale Eigenvalue Problem Solver". In: *Astrophysics Source Code Library* (2023), ascl–2306.

- [50] Ji-Ju Leu and Donna L George. “Hepatic IGFBP1 Is a Prosurvival Factor That Binds to BAK, Protects the Liver from Apoptosis, and Antagonizes the Proapoptotic Actions of P53 at Mitochondria”. In: *Genes & development* 21.23 (2007), pp. 3095–3109.
- [51] Tapio Lönnberg et al. “Single-Cell RNA-seq and Computational Analysis Using Temporal Mixture Modelling Resolves Th1/Tfh Fate Bifurcation in Malaria”. In: *Science Immunology* 2.9 (3rd Mar. 2017), eaal2192. ISSN: 2470-9468. DOI: [10.1126/sciimmunol.aal2192](https://doi.org/10.1126/sciimmunol.aal2192). pmid: 28345074.
- [52] Michael I. Love, Wolfgang Huber and Simon Anders. “Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2”. In: *Genome Biology* 15.12 (5th Dec. 2014), p. 550. ISSN: 1474-760X. DOI: [10.1186/s13059-014-0550-8](https://doi.org/10.1186/s13059-014-0550-8). URL: <https://doi.org/10.1186/s13059-014-0550-8> (visited on 19/07/2024).
- [53] Saunders Mac Lane. *Categories for the Working Mathematician*. 2nd ed. Graduate Texts in Mathematics 5. New York: Springer, 1998. 314 pp. ISBN: 978-0-387-98403-2.
- [54] Leland McInnes et al. “UMAP: Uniform Manifold Approximation and Projection”. In: *Journal of Open Source Software* 3.29 (2018), p. 861.
- [55] Marina Meilă. “Comparing Clusterings—an Information Based Distance”. In: *Journal of multivariate analysis* 98.5 (2007), pp. 873–895.
- [56] Facundo Mémoli, Zhengchao Wan and Yusu Wang. *Persistent Laplacians: Properties, Algorithms and Implications*. 26th July 2021. arXiv: [2012 . 02808](https://arxiv.org/abs/2012.02808) [cs, math]. URL: <http://arxiv.org/abs/2012.02808> (visited on 21/10/2021).
- [57] Liesbeth Minnoye et al. “Chromatin Accessibility Profiling Methods”. In: *Nature Reviews Methods Primers* 1.1 (21st Jan. 2021), pp. 1–24. ISSN: 2662-

8449. DOI: [10.1038/s43586-020-00008-9](https://doi.org/10.1038/s43586-020-00008-9). URL: <https://www.nature.com/articles/s43586-020-00008-9> (visited on 02/07/2024).
- [58] Laurel A Monticelli et al. “IL-33 Promotes an Innate Immune Pathway of Intestinal Tissue Protection Dependent on Amphiregulin–EGFR Interactions”. In: *Proceedings of the National Academy of Sciences* 112.34 (2015), pp. 10762–10767.
- [59] Tianhao Mu et al. “Embryonic Liver Developmental Trajectory Revealed by Single-Cell RNA Sequencing in the Foxa2eGFP Mouse”. In: *Communications Biology* 3.1 (Nov. 2020), pp. 1–12. ISSN: 2399-3642. DOI: [10.1038/s42003-020-01364-8](https://doi.org/10.1038/s42003-020-01364-8).
- [60] Alejandro Parada-Mayorga et al. “Quiver Signal Processing (QSP)”. 2020. arXiv: [2010.11525](https://arxiv.org/abs/2010.11525).
- [61] Karl Pearson F.R.S. “On Lines and Planes of Closest Fit to Systems of Points in Space”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572. DOI: [10.1080/14786440109462720](https://doi.org/10.1080/14786440109462720).
- [62] Gabriele Pizzolato et al. “Single-Cell RNA Sequencing Unveils the Shared and the Distinct Cytotoxic Hallmarks of Human TCRV δ 1 and TCRV δ 2 γ T Lymphocytes”. In: *Proceedings of the National Academy of Sciences* 116.24 (2019), pp. 11906–11915.
- [63] Xiaojie Qiu et al. “Reversed Graph Embedding Resolves Complex Single-Cell Trajectories”. In: *Nature Methods* 14.10 (Oct. 2017), pp. 979–982. ISSN: 1548-7105. DOI: [10.1038/nmeth.4402](https://doi.org/10.1038/nmeth.4402). URL: <https://www.nature.com/articles/nmeth.4402> (visited on 18/05/2022).
- [64] Cédric Rébé and François Ghiringhelli. “STAT3, a Master Regulator of Anti-Tumor Immune Response”. In: *Cancers* 11.9 (2019), p. 1280.

- [65] Benjamin Ricaud et al. “Fourier could be a data scientist: From graph Fourier transform to signal processing on graphs”. In: *Comptes Rendus. Physique* 20.5 (2019), pp. 474–488. ISSN: 1878-1535. DOI: [10.1016/j.crhy.2019.08.003](https://doi.org/10.1016/j.crhy.2019.08.003). URL: <https://comptes-rendus.academie-sciences.fr/physique/articles/10.1016/j.crhy.2019.08.003/> (visited on 18/07/2024).
- [66] Katharina Rindler et al. “Single-Cell RNA Sequencing Reveals Tissue Compartment-Specific Plasticity of Mycosis Fungoides Tumor Cells”. In: *Frontiers in Immunology* 12 (2021), p. 666935.
- [67] Abbas H. Rizvi et al. “Single-Cell Topological RNA-Seq Analysis Reveals Insights into Cellular Differentiation and Development”. In: *Nature Biotechnology* 35.6 (June 2017), pp. 551–560. ISSN: 1546-1696. DOI: [10.1038/nbt.3854](https://doi.org/10.1038/nbt.3854). pmid: [28459448](https://pubmed.ncbi.nlm.nih.gov/28459448/).
- [68] Wouter Saelens et al. “A Comparison of Single-Cell Trajectory Inference Methods”. In: *Nature Biotechnology* 37.5 (May 2019), pp. 547–554. ISSN: 1546-1696. DOI: [10.1038/s41587-019-0071-9](https://doi.org/10.1038/s41587-019-0071-9). URL: <https://www.nature.com/articles/s41587-019-0071-9> (visited on 07/03/2022).
- [69] NYU Satija Lab. *Differential Expression Testing*. URL: https://satijalab.org/seurat/articles/de_vignette.html (visited on 23/07/2022).
- [70] NYU Satija Lab. *Seurat – Guided Clustering Tutorial*. URL: https://satijalab.org/seurat/articles/pbm3k_tutorial.html (visited on 23/05/2022).
- [71] Ansuman T. Satpathy et al. “Massively Parallel Single-Cell Chromatin Landscapes of Human Immune Cell Development and Intratumoral T Cell Exhaustion”. In: *Nature Biotechnology* 37.8 (Aug. 2019), pp. 925–936. ISSN: 1546-1696. DOI: [10.1038/s41587-019-0206-z](https://doi.org/10.1038/s41587-019-0206-z). URL: <https://www.nature.com/articles/s41587-019-0206-z> (visited on 06/02/2023).

- [72] Michael T Schaub et al. “Markov Dynamics as a Zooming Lens for MultiscaleCommunity Detection: Non Clique-like Communitiesand the Field-of-View Limit”. In: *PLoS ONE* 7.2 (2012).
- [73] Anna Seigal, Heather A Harrington and Vidit Nanda. “Principal Components along Quiver Representations”. In: *Foundations of Computational Mathematics* 23.4 (2023), pp. 1129–1165.
- [74] Vladimir V. Sergeichuk. “Unitary and Euclidean Representations of a Quiver”. In: *Linear Algebra and its Applications* 278.1 (15th July 1998), pp. 37–62. ISSN: 0024-3795. DOI: [10.1016/S0024-3795\(98\)00006-8](https://doi.org/10.1016/S0024-3795(98)00006-8). URL: <https://www.sciencedirect.com/science/article/pii/S0024379598000068> (visited on 04/06/2024).
- [75] P. N. Shivakumar and Kim Ho Chew. “A Sufficient Condition for Nonvanishing of Determinants”. In: *Proceedings of the American Mathematical Society* 43.1 (1974), pp. 63–66. ISSN: 0002-9939. DOI: [10.2307/2039326](https://doi.org/10.2307/2039326). JSTOR: [2039326](https://www.jstor.org/stable/2039326). URL: <https://www.jstor.org/stable/2039326> (visited on 19/07/2024).
- [76] David I Shuman, Benjamin Ricaud and Pierre Vandergheynst. “Vertex-Frequency Analysis on Graphs”. In: *Applied and Computational Harmonic Analysis* 40.2 (1st Mar. 2016), pp. 260–291. ISSN: 1063-5203. DOI: [10.1016/j.acha.2015.02.005](https://doi.org/10.1016/j.acha.2015.02.005). URL: <https://www.sciencedirect.com/science/article/pii/S1063520315000214> (visited on 18/07/2024).
- [77] Gurjeet Singh, Facundo Mémoli and Gunnar Carlsson. “Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition”. In: (), p. 11.
- [78] Ronald L. Smith. “Some Interlacing Properties of the Schur Complement of a Hermitian Matrix”. In: *Linear Algebra and its Applications* 177 (1992-

- 12-01, 1992), pp. 137–144. ISSN: 0024-3795. DOI: [10.1016/0024-3795\(92\)90321-Z](https://doi.org/10.1016/0024-3795(92)90321-Z). URL: <https://www.sciencedirect.com/science/article/pii/S002437959290321Z>.
- [79] Silvia Sookoian et al. “Metastasis-Associated Lung Adenocarcinoma Transcript 1 as a Common Molecular Driver in the Pathogenesis of Nonalcoholic Steatohepatitis and Chronic Immune-Mediated Liver Damage”. In: *Hepatology communications* 2.6 (2018), pp. 654–665.
- [80] *Stability of Persistence Diagrams | Proceedings of the Twenty-First Annual Symposium on Computational Geometry*. URL: https://dl.acm.org/doi/abs/10.1145/1064092.1064133?casa_token=u4lpdF0AqMUAAAAA:aYEagcgVAVMN_p6SvrFbctWzCDzsLcEaiYqAVhs8wbmrV65KpZ0qKCSr2IubqX8fdZLIvSU-GAR7 (visited on 19/07/2024).
- [81] Oliver Stegle, Sarah A Teichmann and John C Marioni. “Computational and Analytical Challenges in Single-Cell Transcriptomics”. In: *Nature Reviews Genetics* 16.3 (2015), pp. 133–145.
- [82] Xianbin Su et al. “Single-Cell RNA-Seq Analysis Reveals Dynamic Trajectories during Mouse Liver Development”. In: *BMC genomics* 18.1 (2017), pp. 1–14.
- [83] Otto Sumray, Heather A. Harrington and Vidit Nanda. *Quiver Laplacians and Feature Selection*. 10th Apr. 2024. DOI: [10.48550/arXiv.2404.06993](https://doi.org/10.48550/arXiv.2404.06993). arXiv: [2404.06993](https://arxiv.org/abs/2404.06993) [cs, math, q-bio, stat]. URL: <http://arxiv.org/abs/2404.06993> (visited on 19/07/2024). Pre-published.
- [84] *The Human Protein Atlas — MDK*. URL: <https://www.proteinatlas.org/ENSG00000110492-MDK/single+cell+type/liver> (visited on 11/08/2022).

- [85] Igor Theurl et al. “Dysregulated Monocyte Iron Homeostasis and Erythropoietin Formation in Patients with Anemia of Chronic Disease”. In: *Blood* 107.10 (2006), pp. 4142–4148.
- [86] Robert Tibshirani. “Regression Shrinkage and Selection via the Lasso”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267–288.
- [87] Vincent A Traag, Ludo Waltman and Nees Jan Van Eck. “From Louvain to Leiden: Guaranteeing Well-Connected Communities”. In: *Scientific reports* 9.1 (2019), p. 5233.
- [88] Cole Trapnell et al. “The Dynamics and Regulators of Cell Fate Decisions Are Revealed by Pseudotemporal Ordering of Single Cells”. In: *Nature Biotechnology* 32.4 (Apr. 2014), pp. 381–386. ISSN: 1546-1696. DOI: [10 . 1038 / nbt . 2859](https://doi.org/10.1038/nbt.2859). URL: <https://www.nature.com/articles/nbt.2859> (visited on 18/05/2022).
- [89] Hafiz Ullah et al. “Antiviral Activity of Interferon Alpha-Inducible Protein 27 against Hepatitis B Virus Gene Expression and Replication”. In: *Frontiers in Microbiology* 12 (2021), p. 656353.
- [90] Koen Van den Berge et al. “Trajectory-Based Differential Expression Analysis for Single-Cell Sequencing Data”. In: *Nature Communications* 11.1 (5th Mar. 2020), p. 1201. ISSN: 2041-1723. DOI: [10 . 1038 / s41467 - 020 - 14766 - 3](https://doi.org/10.1038/s41467-020-14766-3). URL: <https://www.nature.com/articles/s41467-020-14766-3> (visited on 18/05/2022).
- [91] J Van Den Heuvel. “Hamilton Cycles and Eigenvalues of Graphs”. In: *Linear algebra and its applications* 226 (1995), pp. 723–730.
- [92] Robin Vandaele et al. “Stable Topological Signatures for Metric Trees through Graph Approximations”. In: *Pattern Recognition Letters* 147 (1st July 2021),

- pp. 85–92. ISSN: 0167-8655. DOI: [10.1016/j.patrec.2021.03.035](https://doi.org/10.1016/j.patrec.2021.03.035). URL: <https://www.sciencedirect.com/science/article/pii/S0167865521001306> (visited on 15/06/2022).
- [93] Aaron R Victor et al. “Epigenetic and Posttranscriptional Regulation of CD16 Expression during Human NK Cell Development”. In: *The Journal of Immunology* 200.2 (2018), pp. 565–572.
- [94] Lothar Von Collatz and Ulrich Sinogowitz. “Spektren endlicher grafen”. In: *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg* 21.1 (1st Dec. 1957), pp. 63–77. ISSN: 1865-8784. DOI: [10.1007/BF02941924](https://doi.org/10.1007/BF02941924). URL: <https://doi.org/10.1007/BF02941924> (visited on 13/09/2024).
- [95] Rui Wang, Duc Duy Nguyen and Guo-Wei Wei. “Persistent Spectral Graph”. In: *International journal for numerical methods in biomedical engineering* 36.9 (2020), e3376.
- [96] Eli Winkelman. “Optimization for Caldero ´ n Problem on Graphs”. In: (2022). URL: <https://eliwinkelman.github.io/images/calderon.pdf>.
- [97] Alex Wolf, Fidel Ramirez and Sergei Rybakov. *Scanpy Tutorials – Preprocessing and Clustering 3k PBMCs*. URL: <https://scanpy-tutorials.readthedocs.io/en/latest/pbmc3k.html> (visited on 02/08/2022).
- [98] F Alexander Wolf, Philipp Angerer and Fabian J Theis. “SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis”. In: *Genome biology* 19.1 (2018), pp. 1–5.
- [99] Baofeng Wu, Jiayu Shao and Xiyang Yuan. “Deleting Vertices and Interlacing Laplacian Eigenvalues”. In: *Chinese Annals of Mathematics, Series B* 31.2 (2010-03-01, 2010), pp. 231–236. ISSN: 1860-6261. DOI: [10.1007/s11401-008-0252-1](https://doi.org/10.1007/s11401-008-0252-1). URL: <https://doi.org/10.1007/s11401-008-0252-1> (visited on 13/06/2023).

- [100] Feng Yan et al. "From Reads to Insight: A Hitchhiker's Guide to ATAC-seq Data Analysis". In: *Genome biology* 21 (2020), pp. 1–16.
- [101] 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. DOI: [10.1002/hep.29353](https://doi.org/10.1002/hep.29353). pmid: [28681484](https://pubmed.ncbi.nlm.nih.gov/28681484/).
- [102] Ke Ye and Lek-Heng Lim. "Schubert Varieties and Distances between Subspaces of Different Dimensions". In: *SIAM Journal on Matrix Analysis and Applications* 37.3 (2016), pp. 1176–1197.
- [103] Jiating Yu et al. "Single-Cell ATAC-seq Analysis via Network Refinement with Peaks Location Information". In: *bioRxiv : the preprint server for biology* (2022), p. 2022.11.18.517159.
- [104] Dietmar MW Zaiss et al. "Emerging Functions of Amphiregulin in Orchestrating Immunity, Inflammation, and Tissue Repair". In: *Immunity* 42.2 (2015), pp. 216–226.
- [105] Abolfazl Zarjou et al. "Ferritin Light Chain Confers Protection against Sepsis-Induced Inflammation and Organ Injury". In: *Frontiers in immunology* 10 (2019), p. 131.
- [106] Yong Zhang et al. "Model-Based Analysis of ChIP-Seq (MACS)". In: *Genome biology* 9.9 (2008), pp. 1–9.