University of Nottingham

School of Mathematical Sciences

Manifold-Valued Data Analysis of Networks and Shapes

Katie Severn

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

September 2019

Dedicated to my Grandma Janet,
a female mathematician I will forever aspire to.

Abstract

This thesis is concerned with the study of manifold-valued data analysis. Manifold-valued data is a type of multivariate data that lies on a manifold as opposed to a Euclidean space. We seek to develop analogue classical multivariate analysis methods, which are appropriate for Euclidean data, for data that lie on particular manifolds. A manifold we particularly focus on is the manifold of graph Laplacians.

Graph Laplacians can represent networks and for the majority of this thesis we focus on the statistical analysis of samples of networks by identifying networks with their graph Laplacian matrices. We develop a general framework for extrinsic statistical analysis of samples of networks by this representation. For the graph Laplacians we define metrics, embeddings, tangent spaces, and a projection from Euclidean space to the space of graph Laplacians. This framework provides a way of computing means, performing principal component analysis and regression, carrying out hypothesis tests, such as for testing for equality of means between two samples of networks, and classifying networks. We will demonstrate these methods on many different network datasets, including networks derived from text and neuroimaging data.

We also briefly consider another well studied type of manifold-valued data, namely shape data, comparing three commonly used tangent coordinates used in shape analysis and explaining the difference between them and why they may not all be suitable to always use.

Acknowledgements

I would like to thank my supervisors Simon Preston and Ian Dryden for all their kindness, help and guidance. I would also like to thank them both for sparing time to discuss the scary question with me 'what is next?', I am excited to now be beginning a career in academia.

I am also very grateful to all the members of staff within the Maths department, Karthik Bharath and Andy Wood as my yearly assessors, Jane Mason for always finding time to help and all the staff who have supported me in organising Women in Maths and mental health events. I am also grateful to Michaela Mahlberg, Viola Wiegand and Anthony Hennessey for their help and discussions about the 19th century novel data. And I am grateful to the Maths for Development group for providing some really exciting research with real world applications.

Thank you to Jon and Charles for being a constant source of smiles, distraction and support in the office throughout my PhD. Maddy for being the best work travel buddy to Tanzania, this was a huge adventure for me and I am so thankful for your help. There are too many other friends in the department to mention all by name but I would like to thank them all, I have made friends for life.

Thank you to my parents you have taught me I can do whatever I put my mind to and encouraged me throughout. Sarah, Ben, Luna and Leo for always helping me when I'm stressed by putting a smile on my face. Bessy for hugs whenever I need them. My Grandparents, each one has inspired me in different ways. And thank you to the rest of my family, I am incredibly proud to be a 'Severn'!

Finally thank you to Adam for pointing out a missing 's' and for reminding me to always be brave.

Contents

Important operators

The vectorise operator vec is obtained from stacking the columns of a matrix, i.e. for a $m_1 \times m_2$ matrix $\boldsymbol{X}$ with columns $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{m_2}$, vec is defined:

$$\text{vec}(\boldsymbol{X}) = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_{m_2}^T)^T. \tag{0.0.1}$$

The vech operator is the half vectorisation of a matrix including the diagonal i.e. for a symmetric $m \times m$ matrix $\boldsymbol{X} = (x_{ij})$, vech is defined:

$$\text{vech}(\boldsymbol{X}) = (x_{11}, x_{12}, \ldots, x_{1m}, x_{22}, x_{23}, \ldots, x_{2m}, x_{33}, x_{34}, \ldots, x_{mm})^T. \tag{0.0.2}$$

The vech* operator is the half vectorisation of a matrix including the diagonal but with $\sqrt{2}$ multiplying the terms corresponding to the off-diagonal, i.e. for a symmetric $m \times m$ matrix $\boldsymbol{X} = (x_{ij})$, vech* is defined:

$$\text{vech}^*(\boldsymbol{X}) = (x_{11}, \sqrt{2}x_{12}, \ldots, \sqrt{2}x_{1m}, x_{22}, \sqrt{2}x_{23}, \ldots, \sqrt{2}x_{2m}, x_{33}, \sqrt{2}x_{34}, \ldots, x_{mm})^T. \tag{0.0.3}$$

The ϕ operator is the half vectorisation of a matrix not including the diagonal i.e. for a symmetric $m \times m$ matrix $\boldsymbol{X} = (x_{ij})$, ϕ is defined:

$$\phi(\boldsymbol{X}) = (x_{12}, \ldots, x_{1m}, x_{23}, \ldots, x_{2m}, x_{34}, \ldots, x_{m-1\,m})^T. \tag{0.0.4}$$

CHAPTER 1

Introduction

The motivating application of this work is to provide a framework for the statistical analysis of samples of networks, including principal component analysis, regression, two-sample testing and classification. A network is a mathematical structure made up of nodes and edges with corresponding weights that are present between nodes. The statistical analysis of networks dates back to at least the 1930s, however interest has increased considerably in the 21st century (Kolaczyk, 2009). Networks are able to represent many different types of data as explained in da Fontoura Costa et al. (2011), examples include social interactions, neuroimaging data and text documents. A text document is represented as a network where nodes represent words and edges are present between words that appear 'near' each other, we define these in more detail in Section 1.3.1. Whilst an extensive amount of work has been done for the analysis of individual networks, it is becoming interesting to also focus on collections of networks as well (Kolaczyk et al., 2017). As we aim to provide a framework for the statistical analysis of samples of networks we are interested in collections of networks instead of just a single one.

For a sample of networks we shall represent each network as a data structure called a *graph Laplacian* matrix, which we define later in (1.2.1). A graph Laplacian is an example of manifold-valued data. A manifold is a space that locally resembles a Euclidean space (Dryden and Mardia, 2016, page 59). Often standard Euclidean statistical methods cannot be directly applied to manifold-valued data and different methods must be developed, this is true for the manifold of graph Laplacians. Manifold-valued data has been studied frequently in different contexts, for example on a sphere in Fisher (1953) and in *shape space* in Kendall (1977), Kendall (1984) and Bookstein (1978).

1.1 Manifold-valued data

As we represent networks as graph Laplacians which lie on a manifold we will need to use ideas from manifold-valued data analysis to define statistical procedures for networks. We will define some key concepts for manifold-valued data that we will go on to use.

On a manifold a *geodesic* path between two points is the path that lies on the manifold representing, for a given distance metric, the shortest path between the two points. On a manifold there can be multiple distance metrics referred to as either an *intrinsic* or *extrinsic* distance. An intrinsic distance is the length of a shortest geodesic path in the manifold. An extrinsic distance is one induced by a Euclidean distance in an embedding of the manifold (Dryden and Mardia, 2016, p112). Formally a function $d : \mathcal{M} \times \mathcal{M} \to [0, \infty)$ is a metric for the manifold $\mathcal{M}$ if it satisfies the following four conditions,

$$\begin{aligned} &(\text{M1})\, d(\boldsymbol{x}, \boldsymbol{y}) \geq 0 \\ &(\text{M2})\, d(\boldsymbol{x}, \boldsymbol{y}) = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y} \\ &(\text{M3})\, d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x}) \\ &(\text{M4})\, d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z}), \end{aligned} \tag{1.1.1}$$

for $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{M}$ (Cullinane, 2011).

There are several different ways that one can define the mean of a sample of data that lie on a manifold. The *Fréchet* mean is a commonly used definition (Fréchet, 1948). For a random variable $Y \in \mathcal{M}$, where $\mathcal{M}$ is a manifold, the population Fréchet mean is defined as

$$\mu = \arg \inf_{\mu' \in \mathcal{M}} \mathrm{E}_Y(d^2(\mu', Y)), \tag{1.1.2}$$

where E_Y is the expectation for the random variable Y and d is a distance in $\mathcal{M}$. The sample Fréchet mean is then defined as

$$\bar{\mu} = \arg \inf_{\mu' \in \mathcal{M}} \frac{1}{n} \sum_{k=1}^{n} d^2(\mu', Y_k). \tag{1.1.3}$$

Different choices of d lead to different definitions of different means on the manifold and these means are termed either intrinsic or extrinsic (Dryden and Mardia, 2016,

Chapter 6). The Fréchet mean is an intrinsic mean if $d(.,.)$ is an intrinsic distance in $\mathcal{M}$. The sample extrinsic mean of the random variables Y_k, of dimension $m \times m$, on a manifold $\mathcal{M}$ is

$$P(\mu^*), \text{ where } \mu^* = \arg \inf_{\mu' \in \mathcal{R}^{m \times m}} \sum_{k=1}^{n} d^2(\mu', Y_k)^2, \tag{1.1.4}$$

where d is an extrinsic distance and P is a projection from the embedding space to a unique closest point in $\mathcal{M}$. Examples of extrinsic means, in the context of shape spaces, can be found in Bhattacharya and Patrangenaru (2003, 2005).

The tangent space of a manifold is a linear space, used for the statistical analysis of manifold-valued data as standard Euclidean statistical methods are often applied in the tangent space. Figure 1.1 shows a simple visualisation of a possible tangent space to a manifold, in this case a sphere. The tangent space at the pole $\boldsymbol{\nu}$ is an Euclidean approximation touching the manifold, chosen so a geodesic is mapped to a straight line preserving distance to the pole. A tangent space mapping provides a connection between the tangent space to the manifold and the inverse mapping is the map from the manifold to the tangent space (Dryden and Mardia, 2016, Chapter 5). There are often multiple ways of mapping to a tangent space however for the majority of our work on statistical analysis of networks we shall only consider one possible tangent space projection. However in Chapter 6 we shall consider a case when there are several choices of tangent space projections and compare these.

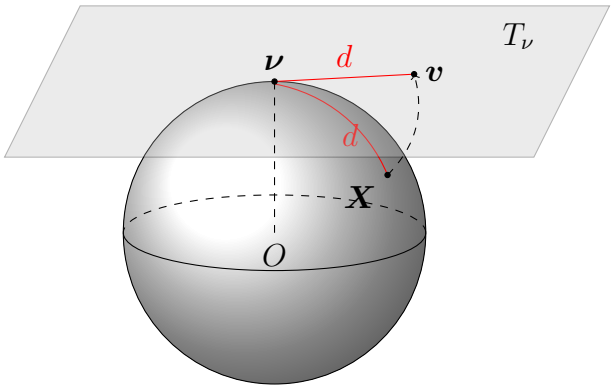


Figure 1.1: *A simple visualisation of a mapping of $\boldsymbol{X}$ onto a tangent space T_ν.*

A very well studied example of manifold-valued data is in shape analysis, and similar ideas and methodology used for this application shall also be useful in our application for the statistical analysis of networks.

1.1.1 Shape analysis

In shape analysis the definition of shape, given in Dryden and Mardia (2016, Definition 1.1), is *"all the geometrical information that remains when location, scale and rotational effects are removed from an object"*. An observation is a configuration matrix, $\boldsymbol{X}_i$ $(k \times m)$, which is the Cartesian coordinates matrix for k landmarks in m dimensions. Translation, rotation and potentially scale need to be removed from the original configurations for shape analysis to be performed.

Translation is removed by pre-multiplying by $\boldsymbol{H}$ the *Helmert sub-matrix*, first used by Kendall (1984). The Helmert sub-matrix $\boldsymbol{H}$, of dimension $k-1 \times k$, has jth row defined as

$$(\underbrace{h_j, \ldots, h_j}_{j \text{ times}}, -jh_j, \underbrace{0, \ldots, 0}_{k-j-1 \text{ times}}), \quad h_j = -(j(j+1))^{-\frac{1}{2}}, \tag{1.1.5}$$

(Dryden and Mardia, 2016, page 49). The landmark coordinates after removing translation are the Helmertized landmark coordinates,

$$\underset{k-1 \times m}{\mathbf{X}_H} = \underset{k-1 \times k}{\boldsymbol{H}} \times \underset{k \times m}{\mathbf{X}}.$$

When we apply ideas extended from shape analysis to the statistical analysis of networks we do not require the condition of objects having invariance to scale therefore work is carried out in the *size-and-shape space* (Dryden and Mardia, 2016, Chapter 5), however we do want invariance to reflection hence our size-and-shape space is defined

$$[\boldsymbol{X}]_S = \{\boldsymbol{X}_H \boldsymbol{R} : \boldsymbol{R} \in \mathcal{O}_m\}, \tag{1.1.6}$$

where $\mathcal{O}_m$ is the set of orthogonal matrices of dimension $m \times m$. The tangent coordinates, with pole $\boldsymbol{\nu}$, for this space are defined as

$$\boldsymbol{v} = \mathbf{X}_H \hat{\boldsymbol{R}} - \boldsymbol{\nu},$$

where $\hat{\boldsymbol{R}}$ is the *Procrustes rotation* of $\mathbf{X}_H$ onto $\boldsymbol{\nu}$. It is this tangent space we use in the framework for the statistical analysis of graph Laplacians. The Procrustes rotation, $\boldsymbol{R}$,

between two configurations $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is defined as

$$\hat{\boldsymbol{R}}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \arg \min_{\boldsymbol{R} \in O(m)} \|\boldsymbol{X}_1 \boldsymbol{R} - \boldsymbol{X}_2\|. \tag{1.1.7}$$

In some applications of shape analysis one may want to have invariance under scale but not reflection which leads to different spaces to consider (Dryden and Mardia, 2016, Chapter 3). Only in Chapter 6 will we consider a different space where we have invariance to scaling but not reflection, named the shape space, defined

$$[\boldsymbol{X}]_S = \{\boldsymbol{Z}\boldsymbol{R} : \boldsymbol{R} \in \mathcal{SO}_m\},$$

where

$$\boldsymbol{Z}_i = \frac{\boldsymbol{H}\boldsymbol{X}_i}{\|\boldsymbol{H}\boldsymbol{X}_i\|},$$

and $\mathcal{SO}_m$ is the set of special orthogonal matrices of dimension $m \times m$; these matrices are orthogonal but restricted to have determinant $+1$. The $\boldsymbol{Z}_i$ is a $(k-1) \times m$ matrix, on the pre-shape sphere, hence satisfying $\|\boldsymbol{Z}_i\| = 1$. The pre-shape sphere is used in much previous work, such as Le and Kendall (1993) and Mardia and Dryden (1999), and is a $(k-1)m - 1$ dimensional hypersphere where information on scaling and translation has been removed from configurations. There are several possible tangent coordinates in this case which we explore in Chapter 6.

1.1.2 Symmetric positive semi-definite matrices

Another frequently studied manifold is the space of *symmetric positive semi-definite matrices* (Moakher and Zéraï, 2011). We shall prove in Result 2.1.1 that the space of graph Laplacians is a subspace of the space of symmetric positive semi-definite matrices. Therefore it is useful to understand the space of symmetric positive semi-definite matrices, as this has been studied far more than the space of graph Laplacians. We denote the space of symmetric positive semi-definite matrices of dimension $m \times m$ by

$$\mathcal{PSD}_m = \{\boldsymbol{A}^{m \times m} : \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \geq 0 \text{ for } \boldsymbol{x} \in \mathbb{R}^m;\ \boldsymbol{A} = \boldsymbol{A}^T\}. \tag{1.1.8}$$

The space $\mathcal{PSD}_m$ is a stratified manifold, split on the rank of the matrices (Weinberger, 1994). The strata are the sets of fixed rank symmetric positive semi-definite matrices

which form a smooth manifold where, put simply, these are manifolds that we can perform calculus on (Lee, 2003). The space of $\mathcal{PSD}_m$ is a *convex cone*, which we will see in Section 2.1 is true for the space of graph Laplacians. For a space $\mathcal{C}$ to be convex any $\boldsymbol{C}_1, \boldsymbol{C}_2 \in \mathcal{C}$ must satisfy

$$c\boldsymbol{C}_1 + (1-c)\boldsymbol{C}_2 \in \mathcal{C} \text{ for any } 0 \leq c \leq 1. \tag{1.1.9}$$

For a space, $\mathcal{C}$, to be a cone any $\boldsymbol{C} \in \mathcal{C}$ must satisfy

$$c\,\boldsymbol{C} \in \mathcal{C} \text{ for any } c > 0. \tag{1.1.10}$$

Applications of positive semi-definite matrices include analysis of medical diffusion tensor data (Fletcher and Joshi, 2007) and pattern recognition (Prabhu et al., 2005). In Fiori (2009) an optimisation problem is used to calculate an intrinsic Fréchet mean of symmetric positive definite matrices and to interpolate between two matrices. An intrinsic mean of symmetric positive definite matrices is also considered in Pennec et al. (2006) using a logarithm based metric. In Arsigny et al. (2007) and Fillard et al. (2007) another logarithm based metric, named the log Euclidean metric, is used between positive definite matrices. The log Euclidean metric is also considered in Dryden et al. (2009) which compares different metrics on the space of positive definite matrices for calculating Fréchet means and interpolation. In this paper, as well as in Zhou et al. (2016), it was seen using the extrinsic metrics, such as the square root Euclidean and Procrustes size-and-shape, that embed the symmetric positive definite matrices can be beneficial, for example leading to more easily interpreted interpolations. We will use similar metrics for graph Laplacians, which we define is Section 1.2.3.

1.2 Statistical analysis of samples of networks

For a sample of networks we shall have each observation as a weighted network, denoted $G_m = (V, E)$, comprising a set of nodes, $V = \{v_1, v_2, \ldots, v_m\}$, and a set of edge weights, $E = \{w_{ij} : w_{ij} \geq 0, 1 \leq i, j \leq m\}$, indicating nodes v_i and v_j are either connected by an edge of weight $w_{ij} > 0$, or else unconnected (if $w_{ij} = 0$). An unweighted network is the special case with $w_{ij} \in \{0, 1\}$. The networks we consider in a given sample will have identical corresponding node sets to all other networks in

that sample. We assume throughout the correspondence between nodes is known as otherwise graph matching would be needed which we will not consider (Conte et al., 2004). We restrict attention to networks that are undirected and without loops, so that $w_{ij} = w_{ji}$ and $w_{ii} = 0$.

1.2.1 Properties of networks

Newman (2010) considers some of the main properties of interest of networks, examples are measures of centrality, geodesic distance between nodes and degree distribution. The degree of a node i in a network is $d_i = \sum_{j=1}^{m} w_{ij}$. Nodes with higher degrees are often seen to play an important role in a network (Newman, 2010, page 9).

There are many summary statistics available for a network, for example the average degree which is given by $\frac{1}{m}\sum_{i=1}^{m} d_i$. Another example is the algebraic connectivity of a network which is defined as the second smallest eigenvalue, λ_2, of the graph Laplacian matrix, defined in (1.2.1) (Fiedler, 1973). Newman (2010, Chapter 7 and 8) provides many more summary statistics such as the clustering coefficient in Equation (7.39) and the assortativity coefficient in Equation (7.82) of their book. We shall not use these as we are interested in the whole structure of the network data and do not want to lose information by representing networks by univariate summary statistics.

One property of a network that we shall use in Section 2.1 is the number of components a network has. A network with 1 component is called "connected" meaning there exists a path between every pair of nodes (Gross and Yellen, 2004, page 10). For any network with more than 1 component there only exists paths between pairs of nodes in the same component and this network is "disconnected". If the algebraic connectivity of a network is $\lambda_2 = 0$ then the network is disconnected (Fiedler, 1973). An example of a connected and disconnected network can be seen in Figure 1.2, for the disconnected network the nodes $\{1, 2\}$ are in one component whilst $\{3, 4, 5\}$ are in the other component.

1.2.2 Graph Laplacians

For the networks we have defined, a network can be uniquely identified by its graph Laplacian. The graph Laplacian matrix, $\boldsymbol{L} = (l_{ij})$, for the network $G_m = (V, E)$, is

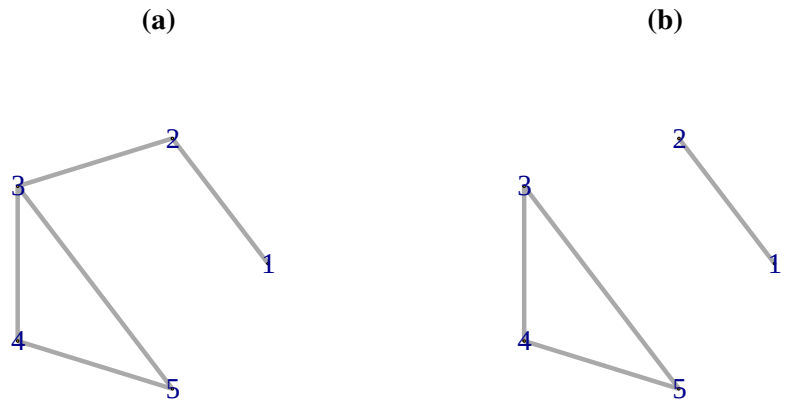


Figure 1.2: *Example of 5 node networks that are a) connected and b) disconnected with 2 components.*

defined as

$$l_{ij} = \begin{cases} -w_{ij}, & \text{if } i \neq j \\ \sum_{k \neq i} w_{ik}, & \text{if } i = j \end{cases} \tag{1.2.1}$$

for $1 \leq i, j \leq m$. It is worth noting there are other forms of graph Laplacians that we will not consider, such as the symmetric normalized Laplacian, defined in Banerjee and Jost (2008).

An in depth survey of graph Laplacians can be found in Merris (1994) which includes many results for graph Laplacian properties including its spectrum and algebraic connectivity. Graph Laplacians have been extensively studied in the field of spectral graph theory (Chung, 1997; Spielman, 2007). This topic has many applications such as in spectral clustering (von Luxburg, 2007), wavelet transforms (Hammond et al., 2011) and image segmentation (Shi and Malik, 2000). However, collections of graph Laplacians, and the space they lie on, is something studied far less.

The graph Laplacian can be written as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$, in terms of the adjacency matrix,

$$\boldsymbol{A} = (w_{ij}), \tag{1.2.2}$$

and degree matrix,

$$\boldsymbol{D} = \text{diag}(\sum_{j=1}^{m} w_{1j}, \ldots, \sum_{j=1}^{m} w_{mj}) = \text{diag}(\boldsymbol{A}\boldsymbol{1}_m), \tag{1.2.3}$$

where $\mathbf{1}_m$ is the m-vector of ones. The ith diagonal element of $\boldsymbol{D}$ equals the degree of node i. Using the graph Laplacian matrix over the degree matrix keeps information on edge weights whilst using the graph Laplacian matrix over the adjacency matrix keep information on the degree of each node. Another advantage of using the graph Laplacian matrix, $\boldsymbol{L}$, is its natural link with the algebraic connectivity of a network, defined in Section 1.2.1 as the second smallest eigenvalue, λ_2, of $\boldsymbol{L}$.

As recently seen in Ginestet et al. (2017), representing networks as graph Laplacians and defining metrics between them provides a promising method for statistical analysis of networks.

1.2.3 Metrics between networks

To perform statistical analysis of networks we must define suitable metrics that will measure distances between networks. For a function between networks to be a metric it must satisfy the conditions in (1.1.1). We will consider two general metrics between graph Laplacians, which are the:

$$\textit{Euclidean power metric:} \qquad d_\alpha(\boldsymbol{L}_1, \boldsymbol{L}_2) = \|\boldsymbol{L}_1^\alpha - \boldsymbol{L}_2^\alpha\|, \tag{1.2.4}$$

$$\textit{Procrustes power metric:} \qquad d_{\alpha,S}(\boldsymbol{L}_1, \boldsymbol{L}_2) = \inf_{\boldsymbol{R}\in\mathcal{O}(m)} \|\boldsymbol{L}_1^\alpha - \boldsymbol{L}_2^\alpha \boldsymbol{R}\|, \tag{1.2.5}$$

where $\hat{\boldsymbol{R}}$ is the ordinary Procrustes match of $\boldsymbol{L}_2^\alpha$ to $\boldsymbol{L}_1^\alpha$ (Dryden and Mardia, 2016, Chapter 7) and $\|\boldsymbol{A}\| = \{\text{trace}(\boldsymbol{A}^T\boldsymbol{A})\}^{1/2}$ is the Frobenius norm, which is also known as the Euclidean norm. Common choices of Euclidean power metrics and Procrustes power metrics are d_1, $d_{\frac{1}{2}}$ and $d_{\frac{1}{2},S}$, referred to as the *Euclidean*, *square root Euclidean* and *Procrustes size-and-shape* metrics respectively (Dryden et al., 2009).

Many metrics already exist for the comparison of networks, such as the cut distance, Hamming distance and the edit metric (Klopp and Verzelen, 2017; Shimada et al., 2016). A mass univariate approach, where each edge is considered separably, is a common comparison tool for network analysis for neuroimaging (Ginestet et al., 2014). A limitation of many of the metrics that already exist is they focus on differences between edges in networks and not the structure of a network as a whole, like the degree of nodes. For example the Hamming distance only considers differences in edges and is

defined between networks $G = (V, E)$ and $G' = (V', E')$ as

$$d_{HD}(G, G') = \sum_{i<j} \mathbb{1}\{w_{ij} \neq w'_{ij}\}.$$

Also some of the metrics are not straightforward to calculate results with, if even possible. For examples when using the distance metric the graph edit distance, defined in (1.2.6), it would not be straightforward to find the network that minimises the sum of the distances between itself and a sample of networks, which would be needed for calculating a mean. The graph edit distance is the least-cost edit operation sequence between two networks, where an edit includes node and edge insertion and deletion, it is written formally as

$$d_{GED}(G, G') = \min_{edit_1,..edit_k} \sum_{i=1}^{k} c(edit_i), \tag{1.2.6}$$

where $c(edit_i)$ is the cost of the ith edit, and the k edits transform G into G' (Gao et al., 2010).

Another issue with most existing metrics is they do not take into account node labelling; this problem can be seen by the recent pseudo-metric between networks, $NetEMD$, defined in Wegner et al. (2018). A pseudo-metric differs to a metric as it no longer satisfies condition M2 in (1.1.1) meaning $NetEMD(x, y) = 0 \nRightarrow x = y$. The $NetEMD$ metric is the mean of all modified earth movers distances of distributions of chosen features within the networks. The distributions that have performed well are the graphlet degree distributions for graphlets up to 4 or 5 nodes. A graphlet is a small connected subgraph and the graphlet degree distribution is how many nodes 'touch' each graphlet (Pržulj, 2010). This metric has been shown to perform well for comparison of certain networks when it is the network topological features that are of interest, for example when classifying Reddit communities networks of discussion based and question/answer based communities (Wegner et al., 2018). However $NetEMD$ is unsuitable for many types of networks where node labelling is important, such as text or neuroimaging networks. When the graphlet degree distribution is used the metric is unchanged by permutation of node labelling. If the same number of nodes 'touch' each possible graphlet in two networks they will have $NetEMD = 0$, even though the actual nodes doing the 'touching' are different. This means the importance of the labelling of nodes is lost, which is obviously undesirable.

Our metrics between graph Laplacians, defined in (1.2.4) and (1.2.5), do not suffer from the undesirable effect from node permutations and this is one reason it may be advantageous to use. For the networks in Figure 1.3 we calculate $NetEMD$, in Table 1.1, to illustrate the effect it only being a pseudo-metric has, and compare it with the Euclidean, square root Euclidean and Procrustes size-and-shape metric between graph Laplacians to show these do not suffer the same effect. We also include the Hamming distance. For example A the networks have had their nodes permuted, and so whilst the two networks look identical their node labelling is different. Example B and C are examples of networks representing text, for these examples an edge is present if two words appear next to each other, Example B shows networks of two sentences, '*I had my house cleaned*' and '*I had cleaned my house*', they have identical words but in a different order which changes the meaning of the sentence. In Example C the two networks represent the sentences '*Why did that researcher choose that example*' and '*I wrote this sentence for this purpose*', these sentences share no common words but have an identical structure. These examples all have $NetEMD$ of 0, if we were comparing the networks structure a distance of 0 seems reasonable, but if the nodes' meaning are of interest a larger distance would be needed. The Euclidean, square root Euclidean, Procrustes size-and-shape and Hamming distance all give distances above 0. Therefore these metrics are sensible to use for networks where the node's values are of interest. Also it is worth noting the relative distance between the pairs of networks for the Euclidean, square root Euclidean and Procrustes size-and shape metrics are similar. For each metric Example B results in the smallest distance and Example C results in the largest.

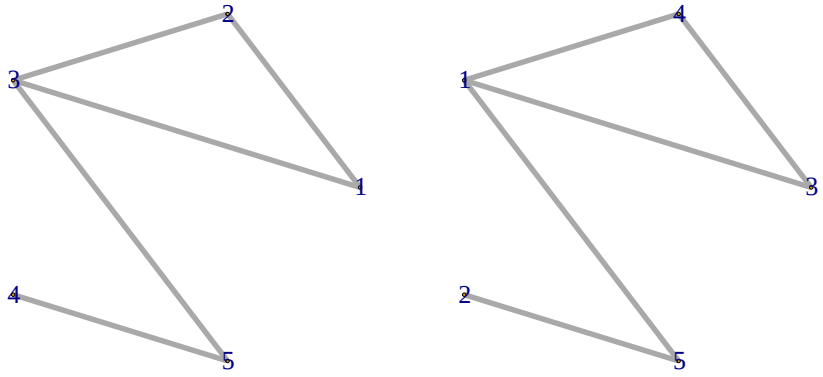
Example	NetEMD	Euclidean	Square root Euclidean	Procrustes size-and-shape	Hamming
A	0	4.47	1.64	1.62	8
B	0	3.16	1.18	1.16	4
C	0	9.17	4.90	4.90	12

Table 1.1: *Some network distance metrics between the example networks in Figure 1.3*

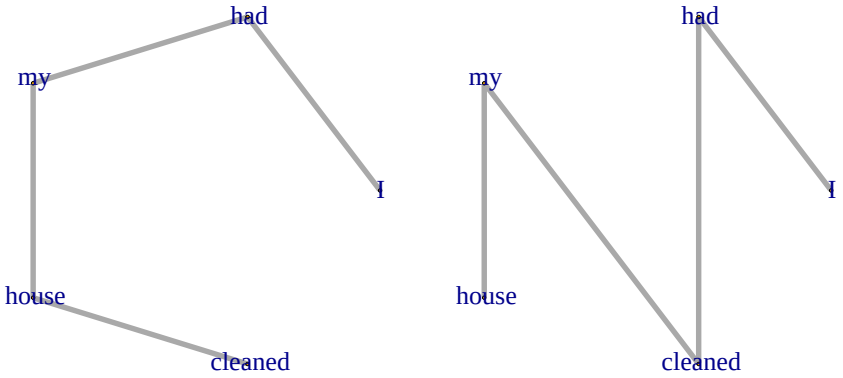
1.2.4 Statistical methods

Using the metrics defined in (1.2.4) and (1.2.5) we will develop a framework for the statistical analysis of networks represented as graph Laplacians. With this framework

Example A:



Example B:



Example C:

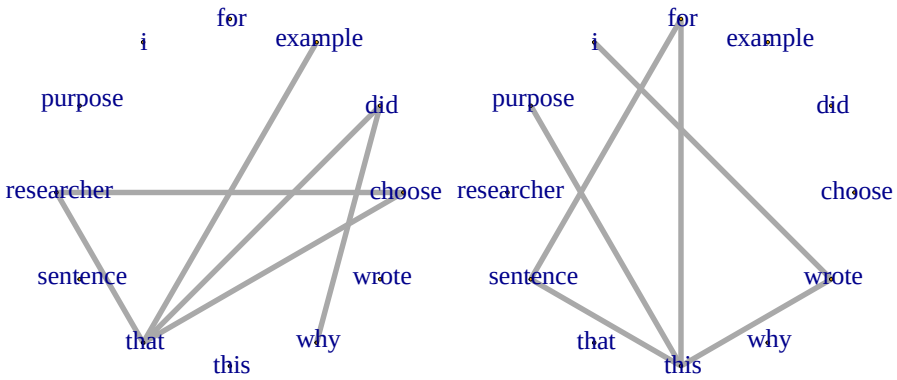


Figure 1.3: *Pairs of networks for which NetEMD=0*

we can adapt many standard statistical methods so they are suitable for samples of networks. Examples of the standard statistical methods we shall generalise include principal component analysis, linear regression and two-sample hypothesis testing. We shall now briefly describe some of the standard statistical methods we shall use.

We shall use Ward's method (Ward, 1963) for the clustering of graph Laplacians. Ward's method is an agglomerative hierarchical clustering method where to begin each graph Laplacian is assigned its own cluster and the algorithm then recursively joins the two most similar clusters, continuing until there is just one cluster left.

In the framework we define for the statistical analysis of graph Laplacians we will solve a convex optimisation problem. A general convex optimisation problem is one such that we wish to find

$$\begin{aligned} & \arg\min\{f(\boldsymbol{L})\} \\ \text{subject to:} \quad & g_i(\boldsymbol{L}) \leq 0, \quad i = 1, ..., k_g \\ & h_j(\boldsymbol{L}) = 0, \quad j = 1, ..., k_h, \end{aligned} \tag{1.2.7}$$

where f, $g_1, ...$ and $h_1, ...$ are convex functions and k_g and k_h are the number of inequality and equality constraints respectively. A convex optimisation problem has the useful characteristic that any local minimum must be the unique global minimum (Rockafellar, 1993).

We shall visualise graph Laplacians in lower dimensions using both principal component analysis (PCA) and multidimensional scaling (MDS). There are already several generalisations of PCA for manifold data, such as Geodesic PCA described in Huckemann et al. (2010) and Huckemann and Hotz (2009). The PCA we shall define in Section 2.5 for graph Laplacians is similar to Fletcher et al. (2004), where a tangent space is used to perform PCA and then results are projected back to the space of graph Laplacians. Earlier approaches of PCA in tangent spaces in shape analysis include Kent (1994) and Cootes et al. (1994).

When defining methods for regression of graph Laplacians as well as looking at parametric models for regression such as the linear model we also will use the popular non-parametric model, the Nadaraya Watson model (Watson, 1964; Nadaraya, 1965; Bierens, 1988). The Nadaraya Watson model predicts an unknown variable y with

known covariate x for the dataset $(\{y_1, x_1\}, \ldots, \{y_n, x_n\})$ as

$$y(x) = \frac{\sum_{i=1}^{n} K_h(x - x_i) y_i}{\sum_{i=1}^{n} K_H(x - x_i)}, \tag{1.2.8}$$

where K_h is a kernel function. We also consider regression with spatial covariate and for this we shall adapt Kriging, also referred to as Gaussian process prediction, a brief overview of which can be found in Chilès and Desassis (2018). Kriging is a geospatial method of estimating points on a random field. The Kriging predictor of an unknown quantity Z on a random field with known coordinates $\boldsymbol{x}$ for the dataset $(\{Z_1, \boldsymbol{x}_1\}, \ldots, \{Z_n, \boldsymbol{x}_n\})$ is

$$Z(\boldsymbol{x}) = \sum_{i=1}^{n} W(\boldsymbol{x}_i) Z_i, \tag{1.2.9}$$

where the weights, W, are chosen to reflect the spatial proximity of data points. The working to find these weights for the spatial graph Laplacians is found in Section 3.5.

We will define a two-sample test for graph Laplacians. Whilst this will rely heavily of the framework for graph Laplacians it will follow the same outline of standard two-sample tests. This outline is defining a test statistic and either finding or approximating its distribution. Common two-sample tests are the students t-tests, Hotelling's T-squared test and Chi-squared test.

In Chapter 5 we shall classify networks by representing them as graph Laplacians and one method to do so involves the use of well documented, supervised classification methods to classify the graph Laplacians. There are many possible standard classification methods we could use but the three we consider are linear discriminant analysis (LDA), random forests and support vector machines (SVM). Linear discriminant analysis is a form of classification that takes a linear combination of variables to form a rule for classification, explained in detail in Chapter 11 of Mardia et al. (1979). LDA relies on the assumption the variables are normally distributed for each class with identical covariance matrices and only the mean vector differing, however it has been seen to still work well when these assumptions are violated (Li et al., 2006). Random forests are a type of ensemble learning method that can be used in classification. Random forests create multiple decision trees which then vote for the most popular class, first described in Breiman (2001), they limit overfitting that is prone in a single decision tree. A decision tree is a classifier that partitions data in a tree like structure using decision rules

(Breiman, 2017). Support vector machines construct a hyperplane that separates the classes by as large a gap as possible. Support vector machines are particularly useful when data is not linearly separable as the kernel trick can be used to transform the data into a higher dimensional space where a hyperplane can be fitted to classify the data, more detail can be found in Chapter 12 of Hastie et al. (2005). Classification techniques are prone to over fitting and so to evaluate the success of a classification method data needs to be split into a training and test set. The training set is what the classification method will be trained on and then of course it is tested on the test set to see how the classification has performed. However the success of the classification is then dependent on how the training and test set were decided and so cross validation, where multiple training and test sets are used, is best as this reduces this dependency. A common metric to measure the success of classification is the accuracy defined as

$$accuracy = 100\frac{\textit{Number of correctly classified data points}}{\textit{Total number of data points}}. \tag{1.2.10}$$

1.2.5 Network data generating models

When applying our statistical method we will sometimes generate networks for simulation studies from different network models; the four we consider and use particularly in Chapter 4 are the *stochastic block model*, the *Erdös-Renyi random network model*, the *Watts-Strogatz small-world model* and the *normal weighted network model*.

The stochastic block model is a commonly used network model, for example it is used in block-clustering, where nodes with similar roles are clustered together (Snijders and Nowicki, 1997; McDaid et al., 2013). A stochastic block model for an m node network, partitions the node set into k subsets $C_1, \ldots, C_k$. The probability of an edge between nodes i and j is then given by p_{uv} where $i \in C_u$ and $j \in C_v$. A stochastic block model can be represented by a probability matrix $P = (p_{ij})$, where p_{ij} is the probability of an edge being present between node i and j. We do not allow mixed membership in the stochastic block model, as described in Airoldi et al. (2008). When there is only one block the network is an Erdös-Renyi random network and so the probability of any edge being present, p_{ij}, is constant for all nodes.

The Watts-Strogatz small-world model is described in Watts and Strogatz (1998). For the Watts-Strogatz model we set the size of the lattice along each dimension as 1, nei is the neighbourhood sizes each node is originally connected with respectively and p is

the respective rewiring probability.

Another model we shall use we call the normal weighted network model which produces networks with weights that are modelled normally, $w_{ij} \sim \mathcal{N}(p, \sigma^2)$, for $1 \leq i, j \leq m$. To prevent negative weights occurring p and σ must be chosen so the chance of the weight being negative is negligible.

However for the majority of our work we shall apply our methods to real network data examples which we shall now describe.

1.3 Datasets to be used

1.3.1 19th Century novels

As stated in Moisl (2015) *'Linguistics is a science, and should therefore use scientific methodology'*, and we will use our statistical methods for corpus linguistic networks. In corpus linguistics, networks are used to model documents comprising a text corpus (Phillips, 1983). By representing text as networks and then graph Laplacians we provide a way of answering questions such as, what is the mean of a sample of texts, how does writing style change with time and how can we classify the author of a text given samples of their previous texts. A recent famous example of classifying texts is for the analysis of the novel 'The Cuckoo's Calling' written under the pen name 'Robert Galbraith'; this was found to actually be written by the famous J.K. Rowling (Juola, 2015). An example of studying differences between authors using text networks is seen in Antiqueira et al. (2007), however this just uses network summary statistics described in Section 1.2.1, for example the average degree of the nodes. Our approach shall compare networks as whole data objects.

To represent a text document as a network each node represents a word, and edges indicate words that co-occur within some span–typically 5 words, which we use henceforth–of each other (Evert, 2008). The span of 5 is justified in corpus linguistics due to the idea from Miller (1956) that the number of objects an average person can hold in working memory is between 5 and 9 and so this is true for words also. This representation conserves information on the co-occurrence of words, and these co-occurrences can be distinctive of different texts, be it authors or genre. Representing texts using collocation is perhaps a more intelligent way to analyse texts then representing them by the

commonly used bag of words model, where only word frequency is considered and the order of words is ignored (Wallach, 2006). The R package `CorporaCoCo` by Hennessey et al. (2017) can be used to convert text into its co-occurrences.

Author	Novel name	Abbreviation	Year written
Austen	Lady Susan	LS	1794
Austen	Sense and Sensibility	SE	1795
Austen	Pride and Prejudice	PR	1796
Austen	Northanger Abbey	NO	1798
Austen	Mansfield Park	MA	1811
Austen	Emma	EM	1814
Austen	Persuasion	PE	1815
Dickens	The Pickwick Papers	PP	1836
Dickens	Oliver Twist	OT	1837
Dickens	Nicholas Nickleby	NN	1838
Dickens	The Old Curiosity Shop	OCS	1840
Dickens	Barnaby Rudge	BR	1841
Dickens	Martin Chuzzlewit	MC	1843
Dickens	A Christmas Carol	C	1843
Dickens	Dombey and Son	DS	1846
Dickens	David Copperfield	DC	1849
Dickens	Bleak House	BH	1852
Dickens	Hard Times	HT	1854
Dickens	Little Dorrit	LD	1855
Dickens	A Tale of Two Cities	TTC	1859
Dickens	Great Expectations	GE	1860
Dickens	Our Mutual Friend	OMF	1864
Dickens	The Mystery of Edwin Drood	ED	1870

Table 1.2: *The Jane Austen and Charles Dickens novels from the CLiC database (Mahlberg et al., 2016).*

The text networks we focus on are for the full text in novels written by Jane Austen and Charles Dickens dataset [1] as listed in Table 1.2, obtained from CLiC (Mahlberg et al., 2016). For each of the 7 Austen and 16 Dickens novels, the "year written" refers to the year in which the author started writing the novel; see The Jane Austen Society of North America (2018) and Charles Dickens Info (2018).

We choose to study Dickens novels as they are frequently studied in corpus linguistics, for example in Mahlberg et al. (2016, 2013). Austen novels are a good set of novels to use alongside Dickens novels as they were written in a similar time period and too

[1] *Christmas Carol* and *Lady Susan* are short novellas rather than novels, but we shall use the term "novel" for each of the works for ease of explanation.

Author	Novel name	Abbreviation
Earl of Beaconsfield Benjamin Disraeli	Sybil, or the two nations	sy
Earl of Beaconsfield Benjamin Disraeli	Vivian Grey	vi
Mary Braddon	Lady Audley's Secret	la
Anne Brontë	Agnes Grey	ag
Charlotte Brontë	Jane Eyre	ja
Charlotte Brontë	The Professor	pr
Emily Brontë	Wuthering Heights	wh
Baron Edward Bulwer Lytton	The Last Days of Pompeii	po
Elizabeth Gaskell	Cranford	cr
Elizabeth Gaskell	Mary Barton	ma
Elizabeth Gaskell	North and South	no
Wilkie Collins	Antonina, or the Fall of Rome	an
Wilkie Collins	Armadale	ar
Wilkie Collins	The Woman in White	ww
Arthur Conan Doyle	The Hound of the Baskervilles	ba
George Eliot	Daniel Deronda	de
George Eliot	The Mill on the Floss	mi
Thomas Hardy	Jude the Obscure	ju
Thomas Hardy	The Return of the Native	na
Thomas Hardy	Tess of the D'Urbervilles	te
William Makepeace Thackeray	Vanity Fair	va
Robert Louis Stevenson	The Strange Case of Dr Jekyll and Mr Hyde	je
Bram Stoker	Dracula	dr
Mary Shelley	Frankenstein	fr
Anthony Trollope	The Small House at Allington	al
Oscar Wilde	The Picture of Dorian Gray	do

Table 1.3: *More novels from the CLiC database (Mahlberg et al., 2016).*

have been studied extensively, for example in Mahlberg (2010); Burrows (1987). We will also briefly look at a larger set of novels of all 19th century authors available from CLiC (Mahlberg et al., 2016), this includes the Austen and Dickens novels as well as the novels found in Table 1.3.

Word	**Rank in all Austen and Dickens novels**	**Rank in Dickens novels**	**Rank in Austen novels**
the	1	1	1
and	2	2	3
to	3	3	2
of	4	4	4
a	5	5	5
i	6	6	7
in	7	7	8
that	8	8	13
it	9	11	10
he	10	10	16
his	11	9	20
was	12	13	9
you	13	12	15
with	14	14	21
her	15	16	6
as	16	15	18
had	17	17	17
for	18	20	19
at	19	21	25
mr	20	18	38
not	21	26	12
be	22	28	14
she	23	31	11
said	24	19	58
have	25	25	23

Table 1.4: *The most common 25 words in the Austen and Dickens novels.*

For each novel we produce a network representing pairwise word co-occurrence. A choice that needs to be made is if we allow co-occurrences over sentence boundaries and chapter boundaries, (Evert, 2008, Section 3) for this data we allow it. If the node set V corresponded to every word in all the novels it would be very large, for the Austen and Dickens subset this would give $m = 48285$, but a relatively small number of words are used far more than others. In the Austen and Dickens subset the top $m = 50$ words cover 45.6% of the total word frequency, $m = 1000$ cover 79.6%, and $m = 10000$ cover

96.7%. We focus on a truncated set of the m most frequent words and the w_{ij}'s are the pairwise co-occurrence counts between these words. In our analysis we choose $m = 1000$ as a sensible trade-off between having very large, very sparse graph Laplacians versus small graph Laplacians of just the most common words. Truncating a novel's word set has been shown to be effective before, for example Burrows (1987) considers just the high frequency words in Austen novels to get insightful results. For each novel and the truncated node set, the network produced is converted to a graph Laplacian. A pre-processing step for the novels is to normalise each graph Laplacian, in order to remove the gross effects of different lengths of the novels, by dividing each graph Laplacian by its own trace, resulting in a trace of 1 for each novel.

As an indication of the broad similarity of the most common words we list the top 25 words for the Austen and Dickens subset in Table 1.4, these words are almost identical to the top 25 words for the full 19th century novel set. Of the top 25 words across all novels 22 appear in the most frequent 25 words for the Dickens novels and 23 for the Austen novels. The words *not*, *be*, *she* do not appear in Dickens' top 25 and the words *mr* and *said* do not appear in Austen's top 25. Some differences in relative rank are immediately apparent: *her*, *she*, *not* having higher relative rank in Austen and *he*, *his*, *mr*, *said* having higher relative rank in Dickens.

Our key statistical goals for the novel data are to investigate the authors' evolving writing styles, by regressing the networks on "year written"; to explore dominant modes of variability, by developing principal component analysis for samples of networks; and to test for significance of differences in Austen's and Dickens' writing styles, via a two-sample test of equality of mean networks.

1.3.2 M-money transaction networks

Another network dataset we shall use throughout is the M-money transaction network dataset, which corresponds to the movement of M-money in Tanzania. M-money transactions include sending and receiving money, making savings deposits, bill payments, making non-cash payments and transferring money from ones mobile phone account to bank accounts and vice versa as described in Mpogole et al. (2016, page 4).

We convert the M-money transactions for the year 2014 into daily networks, giving 365 networks, made of ($m =$)130 nodes representing the districts of Tanzania found in Table 1.5. An edge is present if a transaction occurred between the two districts on

the day. This creates an unweighted network as an edge is either present, if there is a transaction or not. As the nodes for these networks correspond with a spatial location we can plot these networks on a map of Tanzania which we shall do in Example 2.3.2.

Again like with the networks representing Austen and Dickens novels, we pre-process these daily networks to standardise by dividing by the trace, so the graph Laplacians for each day have trace=1.

The uptake of M-money in east African countries like Tanzania has been extremely high, from zero to 5.5 million users in its first 4 years (Mpogole et al., 2016). Tanzania like many emerging economies is struggling to keep key demographic data, such as socio-demographic status, up to date. Engelmann et al. (2018) explains how studying M-money transactions can fill in some of the gaps in the demographic data. As the M-money data can give insightful demographic results it is an interesting dataset to study, especially as very little research exists on it currently. We shall use the M-money networks throughout to demonstrate our methods on, particularly focussing on identifying differences between transactions on weekdays and weekends.

1.3.3 Neuroimaging- fMRI data

Another motivating application for the statistical analysis of network data is from neuroimaging. Using functional MRI images of brains, correlations emerge between functionally related areas of the brain. These are referred to as functional connectivity and give detailed maps of complex neural systems (Biswal et al., 2010).

We use the same dataset as Ginestet et al. (2017) kindly provided by Dr Cedric Ginestet from the the 1000 Functional Connectomes Project launched by Biswal et al. (2010). The data we use from Ginestet et al. (2017) parcellates the scan for a participant into a set of 50 cortical and subcortical regions using the Automated Anatomical Labelling (AAL) template (Tzourio-Mazoyer et al., 2002). As pointed out in Ginestet et al. (2017) the resulting networks are sensitive to the choice of parcellation and just as in their work, our own work generalizes to other parcellations.

For the data there are a total of 1017 participants, with 462 males and 555 females. For each participant a correlation matrix $\boldsymbol{S}_k = (s_{ij})$ is created between each area of the brain from the scans. The correlation matrix is converted to a network with areas of the brain as nodes and edges present between correlated areas (Biswal et al., 2010; Ginestet et al., 2017). The network and hence graph Laplacian is created by thresholding, giving

District names			
Arumeru	Arusha	Babati	Bagamoyo
Bariadi	Biharamulo	Bukoba Rural	Bukoba Urban
Bukombe	Bunda	Chake	Chunya
Dodoma Rural	Dodoma Urban	Geita	Hai
Hanang	Handeni	Igunga	Ilala
Ileje	Ilemela	Iramba	Iringa Rural
Iringa Urban	Kahama	Karagwe	Karatu
Kasulu	Kibaha	Kibondo	Kigoma Rural
Kigoma Urban	Kilindi	Kilolo	Kilombero
Kilosa	Kilwa	Kinondoni	Kisarawe
Kishapu	Kiteto	Kondoa	Kongwa
Korogwe	Kwimba	Kyela	Lindi Rural
Lindi Urban	Liwale	Ludewa	Lushoto
Mafia	Magu	Makete	Manyoni
Masasi	Maswa	Mbarali	Mbeya Rural
Mbeya Urban	Mbinga	Mbozi	Mbulu
Meatu	Micheweni	Misungwi	Mkinga
Mkoani	Mkuranga	Monduli	Morogoro Rural
Morogoro Urban	Moshi Rural	Moshi Urban	Mpanda
Mpwapwa	Mtwara Rural	Mtwara Urban	Mufindi
Muheza	Muleba	Musoma Rural	Musoma Urban
Mvomero	Mwanga	Nachingwea	Namtumbo
Newala	Ngara	Ngorongoro	Njombe
Nkasi	Nyamagana	Nzega	Pangani
Rombo	Ruangwa	Rufiji	Rungwe
Same	Sengerema	Serengeti	Shinyanga Rural
Shinyanga Urban	Sikonge	Simanjiro	Singida Rural
Singida Urban	Songea Rural	Songea Urban	Sumbawanga Rural
Sumbawanga Urban	Tabora Urban	Tandahimba	Tanga
Tarime	Temeke	Tunduru	Ukerewe
Ulanga	Urambo	Urban	Uyui
West	Wete	Zanzibar Central	Zanzibar North A
Zanzibar North B	Zanzibar South		

Table 1.5: *District names of Tanzania*

the weights of each edge as

$$w_{i,j} = \begin{cases} 0, & \text{if } s_{ij} < c \\ 1, & \text{if } s_{ij} \geq c \end{cases} \tag{1.3.1}$$

for some thresholding value c. This c could be a constant for all the networks, but we choose c being a quantile of the correlation values for each correlation matrix, we denote this by Q. There is some debate on how to threshold a correlation matrix into a network, briefly explained in Ginestet et al. (2014), however for the methods we shall use on this data we feel our thresholding is sensible.

Again by representing these networks as graph Laplacians, two-sample tests can be performed to study significant differences in brain activity network means between different demographic groups. We shall consider differences in brain activity network means for gender.

1.3.4 Enron email corpus

The Enron dataset consists of daily networks representing the email interaction between employees at the Enron company. Enron was an American energy company that was hit by an accountancy scandal which resulted in its ultimate closure, more detail on the scandal can be found in Healy and Palepu (2003). During the investigation of Enron the Enron corpus was collected consisting of a large set of emails between Enron employees. This data was made public during the legal investigation of Enron by the Federal Energy Regulatory Commission (Klimt and Yang, 2004). An overview of this dataset can be found in Diesner et al. (2005).

Similar to Shetty and Adibi (2004) we use this data to form social networks between the (m =)151 employees we have present. For each month we create a network with employees as nodes and edges with weights that are the number of emails exchanged between the two employees in the given month. We can then represent these as graph Laplacians for each month. The networks we have are for the months inclusive of June 1999 to April 2002. Just like with the networks representing Austen and Dickens novels, we pre-process these to standardise by dividing by the trace, so the graph Laplacians for each month have trace=1.

The Enron dataset has been studied extensively as social networks due to the unique-

ness of the dataset. Previous work on the dataset includes studying the hierarchy, clustering and importance of the employers within Enron (Agarwal et al., 2012; Wilson and Banzhaf, 2009). A lot of work has also explored the time structure of the Enron networks and how the email interactions change with time (Diesner and Carley, 2005). We will focus on the time structure of the Enron networks, especially trying to identify changes within the network that correspond to the scandal.

1.3.5 Shape data

In Chapter 6 we no longer study network data but instead use shape data as we defined in Section 1.1.1. The motivating data we consider is an enzyme dataset containing the configurations of enzymes with $k = 88$ biological landmarks in $m = 3$ dimensions at $n = 4216$ different times, which we use in Example 6.3.1. Example configurations for this data are shown in Figure 1.4.

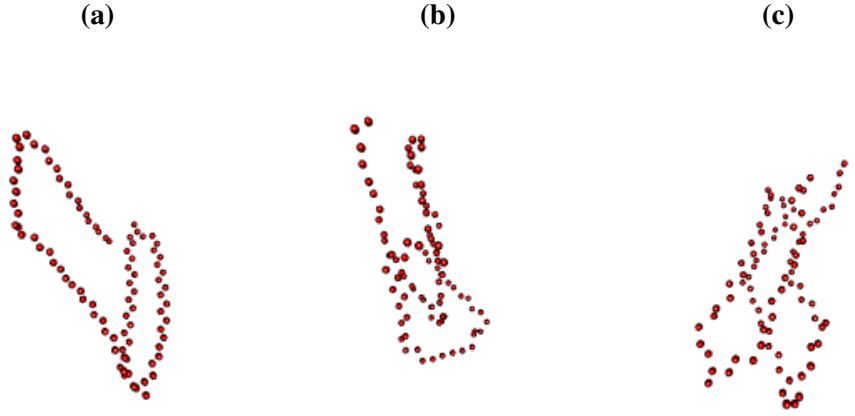


Figure 1.4: *Landmark configurations of the enzyme data at time (i) 1, (ii) 2000 and (iii) 4000.*

We also use briefly in Chapter 6 use three other shape datasets. The first is the dataset of ape skull landmarks which contain (k=)8 landmarks in (m=)2 dimensions for (n=)167 individuals, including gorillas, chimpanzees and orangutans. We also use the landmarks of a DNA molecule that moves in time, with (k=)22 atoms/landmarks in (m=)3 dimensions for (n=)30 time points. The final dataset is the landmarks for sand grain profiles from the Baltic sea and Caucasian River Selenchuk for (k=)50 landmarks in (m=)2 dimensions for (n=)49 grains, the original data for this is from Stoyan (1997).

1.4 Thesis outline

In Chapter 2 we shall provide a general framework for the statistical analysis of networks by representing them as graph Laplacians. This framework will be used throughout Chapters 2, 3, 4 and 5. The framework will involve defining an embedding space, tangent space and metric for the graph Laplacians space. We will use this framework to apply some basic statistical produces such as calculating the mean and performing principal component analysis.

As the novel dataset and Enron dataset have a time structure in Chapter 3 we define parametric and non-parametric methods of regression for graph Laplacians. For the Enron data we shall see an unwanted phenomena named the horseshoe effect, due to its time structure, and so we define a method of removing this effect.

In Chapter 4 we define a two-sample test between graph Laplacians to test samples for a difference in population mean of samples. When we apply our two-sample test to the Austen and Dickens data we see there is significant evidence to suggest a difference in population mean for the authors, hence we provide a method of investigating what the specific differences between these authors are.

Samples of graph Laplacians can belong in different classes for example the novels being in a class of novels written by Dickens vs those not written by him. In Chapter 5 we define a method of classifying graph Laplacians into different classes and also provide a method of detecting anomalies in a sample of graph Laplacians.

In Chapter 6 we consider a different type of manifold-valued data, shape data on the shape space. For this manifold the choice of tangent coordinates is important and we investigate why this is. We provide advice on which tangent coordinates to use under certain cases.

Finally we summarise our findings in Chapter 7. We give future work, including explaining how the framework we have developed could be generalised to other metrics between graph Laplacians.

CHAPTER 2

Population network estimation using graph Laplacians

2.1 Space of graph Laplacians

In this chapter we will define the space of Graph Laplacians and use the fact it is a subspace of the space of positive semi-definite matrices. We then shall define a framework to perform statistical analysis on samples of networks that are represented as graph Laplacians. The framework presented in this chapter can be found in Severn et al. (2019). This framework involves the embedding of graph Laplacians. This embedding of whole networks represented as graph Laplacians is completely different to network embeddings which focus on the embedding of nodes of a single network, for example in Chen et al. (2018). Once the framework has been introduced we shall use it to define methods of calculating means, interpolating and performing PCA on graph Laplacians.

From the definition of a graph Laplacian in (1.2.1) it is clear the space of all graph Laplacians of dimension $m \times m$ can be written as

$$\begin{aligned} &\mathcal{L}_m = \{\boldsymbol{L} = (l_{ij})\} \text{ such that:} \\ &\quad \boldsymbol{L} = \boldsymbol{L}^T \qquad \textit{(symmetric)}, \\ &\quad l_{ij} \leq 0 \, \forall i \neq j \qquad \textit{(non-positive off-diagonal elements)}, \\ &\quad \boldsymbol{L}\mathbf{1}_m = \mathbf{0}_m \qquad \textit{(zero row sum)}, \end{aligned} \tag{2.1.1}$$

where $\mathbf{1}_m$ and $\mathbf{0}_m$ are the m-vector of ones and zeroes respectively. We note due to the rows summing to zero that the diagonal elements must be non-negative as the off-

diagonals are non-positive. Also due to the symmetry, $\boldsymbol{L}^T\mathbf{1}_m = \mathbf{0}_m$, hence the columns also sum to zero.

The space of $\mathcal{L}_m$ is a manifold, in particular it is a sub-manifold of $\mathbb{R}^{m\times m}$ with corners (Ginestet et al., 2017). A d dimensional manifold with corners can be locally modelled by $[0,\infty)^k \times \mathbb{R}^{d-k}$; for full details see Joyce (2009). Many of the methods we will define are adapted from the space of symmetric positive semi-definite matrices, $\mathcal{PSD}_m$ (defined in (1.1.8)), which we shall now prove $\mathcal{L}_m$ is a subset of, and also prove some similar properties these spaces share.

Result 2.1.1. *$\mathcal{L}_m \subset \mathcal{PSD}_m$, where $\mathcal{PSD}_m$ is the space of symmetric positive semi-definite matrices of dimension $m \times m$.*

Proof. For $\boldsymbol{L} \subset \mathcal{PSD}_m$ we must have $\boldsymbol{L} \in \mathcal{L}_m \Rightarrow \boldsymbol{L} \in \mathcal{PSD}_m$. A sufficient condition for $\boldsymbol{L} \in \mathcal{PSD}_m$ is for $\boldsymbol{L}$ to have real positive diagonal elements and to be diagonally dominant (De Klerk, 2006, page 232). A matrix, $\boldsymbol{A} = (a_{ij})$ is diagonally dominant if $|a_{ii}| \geq \sum_{i\neq j} |a_{ij}|$ for all i. For a $\boldsymbol{L} \in \mathcal{L}_m$ it is clear it has positive diagonal elements and that $|l_{ii}| = \sum_{i\neq j} |l_{ij}|$, hence any $\boldsymbol{L} \in \mathcal{L}_m$ is diagonally dominant and so $\boldsymbol{L} \in \mathcal{PSD}_m$. □

Just like the space $\mathcal{PSD}$, the space for graph Laplacians is a convex cone. For definitions of "convex" and "cone" see (1.1.9) and (1.1.10) respectively.

Result 2.1.2. *The space $\mathcal{L}_m$ is a convex space.*

Proof. It is sufficient to show for any $\boldsymbol{L}_1 = (l^1_{ij}), \boldsymbol{L}_2 = (l^2_{ij}) \in \mathcal{L}_m$ that $\boldsymbol{L} = (l_{ij}) = c\boldsymbol{L}_1 + (1-c)\boldsymbol{L}_2 \in \mathcal{L}_m$ for any $0 \leq c \leq 1$. We can see

$$\begin{aligned}
&\boldsymbol{L}^T = c\boldsymbol{L}_1^T + (1-c)\boldsymbol{L}_2^T = c\boldsymbol{L}_1 + (1-c)\boldsymbol{L}_2 = \boldsymbol{L}, && \textit{(symmetry)}\\
&l_{ij} = cl^1_{ij} + (1-c)l^2_{ij} \leq 0 \text{ for } i \neq j && \textit{(non-positive off-diagonal elements)}\\
&\boldsymbol{L}\mathbf{1} = c\boldsymbol{L}_1\mathbf{1} + (1-c)\boldsymbol{L}_2\mathbf{1} = \mathbf{0}, && \textit{(zero row sum).}
\end{aligned}$$

Clearly then $\boldsymbol{L} \in \mathcal{L}_m$ as it satisfies all the graph Laplacian conditions in (2.1.1) we have convexity. □

Result 2.1.3. *The space $\mathcal{L}_m$ is a cone.*

Proof. It is sufficient to show that for any $\boldsymbol{L}_1 = (l^1_{ij}) \in \mathcal{L}_m$ we must have $\boldsymbol{L} = (l_{ij}) =$

$c\boldsymbol{L}_1 \in \mathcal{L}_m$, for any $c > 0$. We see

$$\begin{aligned} \boldsymbol{L}^T &= c\boldsymbol{L}_1^T = c\boldsymbol{L}_1 = \boldsymbol{L}, && \textit{(symmetry)} \\ l_{ij} &= cl_{ij}^1 \leq 0 \text{ for } i \neq j, && \textit{(non-positive off-diagonal elements)} \\ \boldsymbol{L}\mathbf{1} &= c\boldsymbol{L}_1\mathbf{1} = \mathbf{0} && \textit{(zero row sum).} \end{aligned}$$

Clearly then $\boldsymbol{L} \in \mathcal{L}_m$ as it satisfies all the graph Laplacian conditions in (2.1.1) so $\mathcal{L}_m$ is a cone. □

Just like the space $\mathcal{PSD}_m$, our space of interest $\mathcal{L}_m$ is also a stratified manifold, it can be written as

$$\mathcal{L}_m = \mathcal{L}_m^{(1)} \cup \mathcal{L}_m^{(2)} \cup ... \cup \mathcal{L}_m^{(m-1)},$$

where $\mathcal{L}_m^{(r)}$ are the strata defined as

$$\mathcal{L}_m^{(r)} = \{\boldsymbol{L} \in \mathcal{L}_m : \text{rank}(\boldsymbol{L}) = r\},$$

which is the space of graph Laplacians of rank r. For an m node network the rank of its graph Laplacian corresponds to the number of components of the network, defined in Section 1.2.1. A graph Laplacian representing a network with $m - r$ components has rank r. Each stratum $\mathcal{L}_m^{(r)}$, is a sub-manifold of $\mathbb{R}^{m\times m}$ with dimension $mr - \frac{r(r-1)}{2}$. Previous work in Ginestet et al. (2017) focussed on the space of graph Laplacians representing fully connected networks, $\mathcal{L}_m^{(m-1)}$ and only briefly considered disconnected networks having precisely $m - r$ components. We however will work with the much more general space $\mathcal{L}_m$.

2.2 Framework

The general framework we will define in this section for the statistical analysis of graph Laplacians is shown schematically in Figure 2.1. This framework involves embedding a graph Laplacian (F_α) and then mapping this into a tangent space, using the inverse exponential map ($\exp_\nu^{-1}$), where statistical analysis can be performed. The results from the analysis are then mapped from the tangent space to the embedding space ($\exp_\nu$) where an inverse embedding is applied (F_α^{-1}), the result is still not guaranteed to be a

graph Laplacian and so another projection ($\mathrm{P}_{\mathcal{L}}$) is needed to project the result into $\mathcal{L}_m$. The identity projection, Id, maps an object to itself and the mapping illustrates that $\mathcal{L}_m \subset \mathcal{PSD}_m$.

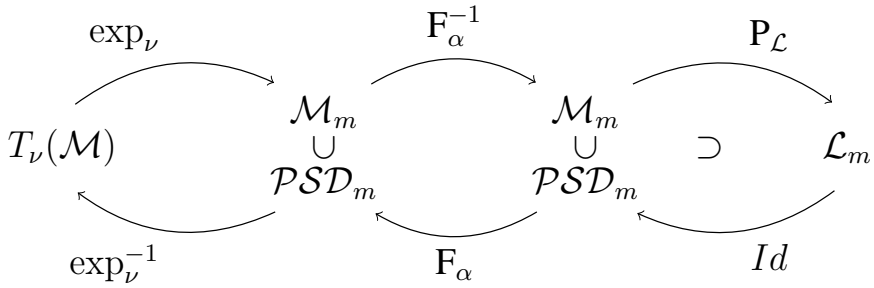


Figure 2.1: *Schematic for the general framework for the statistical analysis of graph Laplacians. The embedding, F_α, inverse embedding, F_α^{-1}, and embedding space, $\mathcal{M}_m$, are defined in Section 2.2.1 and 2.2.3 respectively. The tangent space, $T_\nu(\mathcal{M})$, and associated projections, $\exp_\nu^{-1}$ and $\exp_\nu$, are defined in Section 2.2.4. The projection, $P_{\mathcal{L}}$, is defined in Section 2.2.5.*

2.2.1 Embedding

To embed a graph Laplacian we first write $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ by the spectral decomposition theorem, with $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ and $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m)$, where $\{\lambda_i\}_{i=1,\ldots,m}$ and $\{\boldsymbol{u}_i\}_{i=1,\ldots,m}$ are the eigenvalues and corresponding eigenvectors of $\boldsymbol{L}$. Since $\mathcal{L}_m \subset \mathcal{PSD}_m$ then $\lambda_i \geq 0$, hence for any $\alpha > 0$

$$\mathrm{F}_\alpha(\boldsymbol{L}) = \boldsymbol{L}^\alpha = \boldsymbol{U}\boldsymbol{\Lambda}^\alpha\boldsymbol{U}^T : \mathcal{PSD}_m \to \mathcal{M}_m, \tag{2.2.1}$$

embeds $\mathcal{PSD}_m$ into $\mathcal{M}_m$, where $\mathcal{M}_m$ is an embedding space, dependent on the choice of metric, and defined for specific metrics in Section 2.2.3. A common choice for α for us will be either $\alpha = 1$ or $\alpha = \frac{1}{2}$.

We observe the following, which is useful in later proofs.

Result 2.2.1. *For $\boldsymbol{L} \in \mathcal{L}_m$ then $F_\alpha(\boldsymbol{L})$ is centred, meaning $F_\alpha(\boldsymbol{L})\mathbf{1}_m = \mathbf{0}_m$.*

Proof. $\boldsymbol{L}$ is centred as $\boldsymbol{L}\mathbf{1}_m = \mathbf{0}_m$, this means $\boldsymbol{L}$ has an eigenvalue $\lambda_i = 0$ corresponding to the eigenvector $\boldsymbol{u}_i = \mathbf{1}_m$. As $\mathrm{F}_\alpha(\boldsymbol{L}) = \boldsymbol{U}\boldsymbol{\Lambda}^\alpha\boldsymbol{U}^T$, the eigenvectors of $\mathrm{F}_\alpha(\boldsymbol{L})$ are the columns of $\boldsymbol{U}$ hence $\boldsymbol{u}_i = \mathbf{1}_m$ is also an eigenvector of $\mathrm{F}_\alpha(\boldsymbol{L})$ and its corresponding eigenvalue is $\lambda_i^\alpha = 0^\alpha = 0$. Therefore $\mathrm{F}_\alpha(\boldsymbol{L})\mathbf{1}_m = \mathbf{0}_m$ hence $\mathrm{F}_\alpha(\boldsymbol{L})$ is also centred. □

2.2.2 Metrics

As explained in Section 1.1 embeddings can be used to define metrics and this is the case for our space $\mathcal{L}_m$. The Euclidean power metric (1.2.4) and Procrustes power metric (1.2.5) we introduced between graph Laplacians can now be written in terms of the embedding, F_α, for $\boldsymbol{L}_1, \boldsymbol{L}_2 \in \mathcal{L}_m$, as

$$d_\alpha(\boldsymbol{L}_1, \boldsymbol{L}_2) = \|\mathrm{F}_\alpha(\boldsymbol{L}_1) - \mathrm{F}_\alpha(\boldsymbol{L}_2)\| \tag{2.2.2}$$

$$d_{\alpha,S}(\boldsymbol{L}_1, \boldsymbol{L}_2) = \inf_{\boldsymbol{R}\in\mathcal{O}(m)} \|\mathrm{F}_\alpha(\boldsymbol{L}_1) - \mathrm{F}_\alpha(\boldsymbol{L}_2)\boldsymbol{R}\|. \tag{2.2.3}$$

These metrics in fact hold more generally for $\boldsymbol{L}_1, \boldsymbol{L}_2 \in \mathcal{PSD}_m$. Using definitions from Section 1.1 we can see on $\mathcal{L}_m$ the Euclidean distance d_1 is intrinsic as F_α is just the identity map, but in general d_α and $d_{\alpha,S}$ are extrinsic with respect to the manifold, as they are Euclidean distances in the embedding space. As explained in Section 1.2.3, common choices of metrics are d_1, $d_{\frac{1}{2}}$ and $d_{\frac{1}{2},S}$, referred to as the Euclidean, square root Euclidean and Procrustes size-and-shape metrics respectively.

Example 2.2.1: Clustering of the Austen and Dickens novel data

We initially compare some choices of distance metrics on the Austen and Dickens data after constructing the graph Laplacians from the $m = 1000$ most frequent words across all 23 novels. Figure 2.2 (left column) shows the results of a hierarchical cluster analysis using Ward's method (Ward, 1963), described in Section 1.2.4, based on pairwise distances between novels using the metrics d_1, $d_{\frac{1}{2}}$ and $d_{\frac{1}{2},S}$. The dendrograms in Figure 2.2 are a graphical way of representing how the clusters are formed at each stage in the algorithm. The dendrograms when using $d_{\frac{1}{2}}$ and $d_{\frac{1}{2},S}$ separate the authors into two very distinct clusters, shown by the dashed line, whereas when using d_1, Dickens' *David Copperfield* and *Great Expectations* are clustered with Austen's *Lady Susan* which although seems unsatisfactory, actually these three novels all contain more first person narration which could explain them clustering together. The next sub-division of the Dickens cluster using $d_{\frac{1}{2}}$ or $d_{\frac{1}{2},S}$ splits the novels into groups of the earlier novels versus later novels, with the exception being the historical novel *A Tale of Two Cities* which is clustered with the earlier novels. There is not such a clear sub-division for Dickens when using d_1. In the Austen cluster when using $d_{\frac{1}{2}}$ or $d_{\frac{1}{2},S}$ there is clearly a large distance between *Lady Susan* and the rest. *Lady Susan* is Austen's earliest work, a short novella published 54 years after Austen's death.

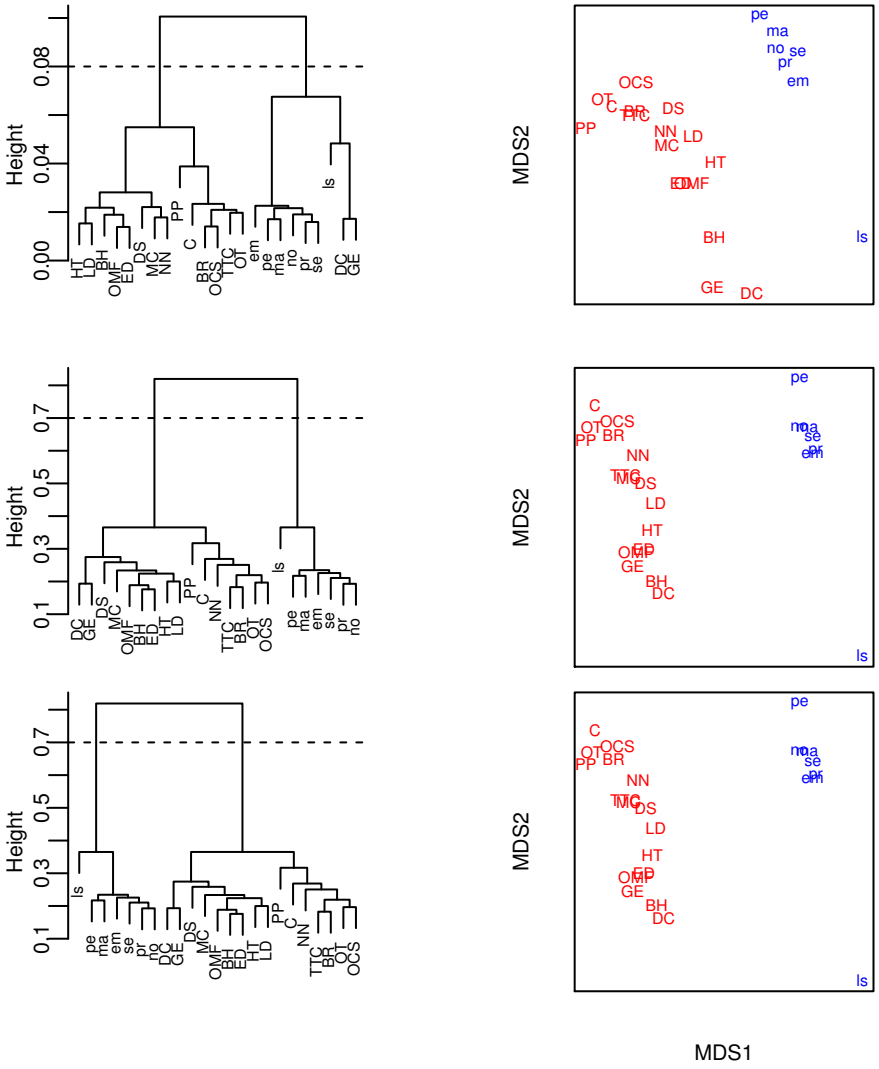


Figure 2.2: *Cluster analysis and MDS plots based on (from top to bottom) the Euclidean distance, d_1, square root Euclidean distance, $d_{\frac{1}{2}}$, and Procrustes size and shape distance, $d_{\frac{1}{2},S}$ each with $m = 1000$. The dashed horizontal line on the dendrogram indicates the cut to form two distinct clusters. The plots display Austen's novels in blue and lower case, and Dickens' novels in red and upper case, the abbreviations are found in Table 1.2.*

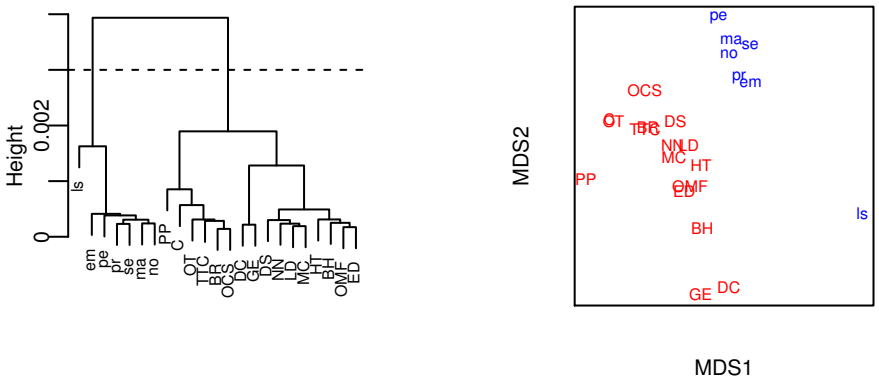


Figure 2.3: *Cluster analysis and MDS plots based on the Euclidean distance, d_1 using full word set of novels. The dashed horizontal line on the dendrogram indicates the cut to form two distinct clusters. The plot displays Austen's novels in blue and lower case, and Dickens' novels in red and upper case, the abbreviations are found in Table 1.2.*

Figure 2.2 (right column) shows corresponding plots for the novels of the first two multi-dimensional scaling (MDS) variables from a classical multi-dimensional scaling analysis, also referred to as Principal coordinate analysis (PCoA) in Gower (1966). The $d_{\frac{1}{2}}$ and $d_{\frac{1}{2},S}$ MDS plots are visibly identical, although they are slightly different numerically. We see that there is a clear separation in MDS space between Austen's and Dickens' works with a very strong separation in MDS1 using $d_{\frac{1}{2}}$ and $d_{\frac{1}{2},S}$, and less so d_1. An alternative method of clustering to Ward's method is k-means clustering applied to the first two MDS coordinates. For the k-means clustering k was chosen to be 2 and for each metric k-means clustered the novels by author exactly.

When using d_1 it is computationally possible to calculate the metric between the novels using the entire set of 48285 words, which leads to graph Laplacians of dimension $m = 48285$. Results based on the entire set are shown in Figure 2.3, and appear similar to using the Euclidean metric on the truncated words set, except now in the dendrogram Austen and Dickens are separated completely.

2.2.3 Reverse embeddings

The framework requires an inverse to the embeddings. We consider three choices of F_α^{-1} for the reverse mapping back from the embedding space, which are suitable for different scenarios. The choice of F_α^{-1} is dependent on whether we want to project to

$\mathcal{PSD}$ before reversing the powering of α.

When using the Euclidean power metric, the space $\mathcal{M}_m$ is the space of real symmetric $m \times m$ matrices with centred rows and columns, and we use

$$\mathrm{F}_\alpha^{-1}(\boldsymbol{Q}) = \begin{cases} (\boldsymbol{Q})^{\frac{1}{\alpha}}, \text{ when } \frac{1}{\alpha} \text{ is an odd integer } : \mathcal{M}_m \to \mathcal{M}_m \\ \left(\dfrac{\boldsymbol{Q} + \boldsymbol{Q}^T + \{(\boldsymbol{Q} + \boldsymbol{Q}^T)^T(\boldsymbol{Q} + \boldsymbol{Q}^T)\}^{\frac{1}{2}}}{4} \right)^{\frac{1}{\alpha}}, \text{ otherwise } : \mathcal{M}_m \to \mathcal{PSD}_m \end{cases}$$

The second expression before taking the power $\frac{1}{\alpha}$ is the closest symmetric positive semi-definite matrix to $\boldsymbol{Q}$ in terms of Frobenius distance (Higham, 1988). If $\boldsymbol{Q} \in \mathcal{PSD}_m$ then this projection has no effect. But for $\boldsymbol{Q} \notin \mathcal{PSD}_m$ with eigenvalues $\varrho_1, \ldots \varrho_m$ it has at least one eigenvalue $\varrho_i < 0$. Therefore for $\frac{1}{\alpha} \notin \mathbb{Z}$ we project to the closest symmetric positive semi-definite as in this case raising $\boldsymbol{Q}$ to the power $\frac{1}{\alpha} \notin \mathbb{Z}$ is only real if $\boldsymbol{Q} \in \mathcal{PSD}$. When $\frac{1}{\alpha} \in \mathbb{Z}$ then in $\boldsymbol{Q}^{\frac{1}{\alpha}}$ a negative eigenvalue, ϱ_i, becomes

$$\varrho_i^{\frac{1}{\alpha}} = \begin{cases} < 0, & \text{if } \frac{1}{\alpha} \text{ is odd} \\ > 0, & \text{if } \frac{1}{\alpha} \text{ is even} \end{cases}$$

as $\varrho_i < 0$ we would want the corresponding eigenvalue in $\boldsymbol{Q}^{\frac{1}{\alpha}}$ to be negative or close to 0, and this is only true when $\frac{1}{\alpha}$ is odd, and therefore when $\frac{1}{\alpha}$ is even we project first into $\mathcal{PSD}$ before raising the power.

For the Procrustes power metric, the space $\mathcal{M}_m$ is the reflection size-and-shape space, denoted $RS\Sigma_{m-1}^m$ (Dryden et al., 2009; Dryden and Mardia, 2016, p67), and in this case we use

$$\mathrm{F}_\alpha^{-1}(\boldsymbol{Q}) = (\boldsymbol{Q}\boldsymbol{Q}^T)^{\frac{1}{2\alpha}} : \mathcal{M}_m \to \mathcal{PSD}_m.$$

We choose this reverse map as it removes the orthogonal matrices from the Procrustes fits, which we will see in the next section is introduced from the exponential map.

2.2.4 Tangent space

To perform further statistical analysis, such as interpolation, extrapolation and PCA, the inverse exponential map, $\exp_\nu^{-1}$, is used to project into a tangent space from $\mathcal{M}_m$, in

which standard statistical methods can be applied, where $\boldsymbol{\nu} \in \mathcal{M}_m$ denotes the pole of the projection. The inverse exponential map is commonly used in tangent projections, for example in Dryden and Mardia (2016, Section 3) and Schmidt et al. (2006), as well as using this mapping in our graph Laplacian framework, in Section 6.2.5 we study this mapping for a different space, the shape space.

We have seen that in Result 2.2.1 the centering constraints on graph Laplacians are preserved for our choice of embedding F_α in $\mathcal{M}_m$. We can remove the centering constraints and reduce the dimensions when projecting to a tangent space by pre and post multiplying by the $m-1 \times m$ Helmert sub-matrix $\boldsymbol{H}$ and its transpose, defined in Section 1.1.1, as a component of the projection.

For the Euclidean power metric, defined in (2.2.2), we define the inverse exponential map $\exp_\nu^{-1}$ to the tangent space $T_\nu(\mathcal{M}_m) = \mathbb{R}^{\frac{m(m-1)}{2}}$ as

$$\begin{aligned} \exp_\nu^{-1}(\boldsymbol{Q}) &= \text{vech}^*\{\boldsymbol{H}(\boldsymbol{Q}-\boldsymbol{\nu})\boldsymbol{H}^T\} \\ \exp_\nu(\boldsymbol{v}) &= \boldsymbol{\nu} + \boldsymbol{H}^T(\text{vech}^*)^{-1}(\boldsymbol{v})\boldsymbol{H}, \end{aligned} \tag{2.2.4}$$

where vech* is defined in (0.0.3). For this definition of tangent space for the Euclidean power metric $\mathcal{M}_m$ is actually Euclidean, with zero curvature, and the results from statistical procedures are often unaffected by the choice of $\boldsymbol{\nu}$ so we often take $\boldsymbol{\nu} = \boldsymbol{0}$.

For the Procrustes power metric, defined in (2.2.3), we define the map $\exp_\nu^{-1}$ to the tangent space $T_\nu(\mathcal{M}_m) = \mathbb{R}^{m-1 \times m-1}$ as

$$\begin{aligned} \exp_\nu^{-1}(\boldsymbol{Q}) &= \text{vec}\{\boldsymbol{H}(\boldsymbol{Q}\hat{\boldsymbol{R}}-\boldsymbol{\nu})\boldsymbol{H}^T\} \\ \exp_\nu(\boldsymbol{v}) &= (\boldsymbol{\nu} + \boldsymbol{H}^T\text{vec}^{-1}(\boldsymbol{v})\boldsymbol{H})\tilde{\boldsymbol{R}} \end{aligned} \tag{2.2.5}$$

where vec is defined in (0.0.1). $\hat{\boldsymbol{R}}$ is the ordinary Procrustes match of $\boldsymbol{Q}$ to $\boldsymbol{\nu}$ defined in (1.1.7) and $\tilde{\boldsymbol{R}}$ is the ordinary Procrustes match from $(\boldsymbol{\nu} + \boldsymbol{H}^T\text{vec}^{-1}(\boldsymbol{v})\boldsymbol{H})$ to $\boldsymbol{\nu}$ (Dryden and Mardia, 2016, chapter 7). The reflection size-and-shape space is a space with positive curvature (Kendall et al., 1999) and the choice of $\boldsymbol{\nu}$ depends on what statistical analysis is being performed. A sensible choice for $\boldsymbol{\nu}$ is often the unprojected sample Fréchet mean, defined later in Section 2.3.

For the Euclidean power metric the Euclidean distance in the embedding space is conserved in the tangent space, the three following results prove this. This is useful as it can sometimes simplify calculations.

Result 2.2.2. *For $\boldsymbol{Q} = F_\alpha(\boldsymbol{L})$ where $\boldsymbol{L} \in \mathcal{L}_m$ then*

$$\begin{aligned}\|\boldsymbol{H}\boldsymbol{Q}\boldsymbol{H}^T\| &= \|\boldsymbol{Q}\| \\ &= \sqrt{trace(\boldsymbol{Q}^T\boldsymbol{Q})}.\end{aligned}$$

Proof. $\boldsymbol{Q}$ is centered from Result 2.2.1, and clearly also $\boldsymbol{Q} = \boldsymbol{Q}^T$.

$$\begin{aligned}\|\boldsymbol{H}\boldsymbol{Q}\boldsymbol{H}^T\| &= \sqrt{\text{trace}(\boldsymbol{H}\boldsymbol{Q}^T\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{Q}\boldsymbol{H}^T)} \\ &= \sqrt{\text{trace}(\boldsymbol{H}\boldsymbol{Q}^T\boldsymbol{C}\boldsymbol{Q}\boldsymbol{H}^T)} \\ &= \sqrt{\text{trace}(\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{Q}^T\boldsymbol{C}\boldsymbol{Q})} \qquad \text{as trace is invariant under cyclic permutation} \\ &= \sqrt{\text{trace}(\boldsymbol{C}^T\boldsymbol{Q}^T\boldsymbol{Q})} \\ &= \sqrt{\text{trace}(\boldsymbol{Q}^T\boldsymbol{Q})} \\ &= \|\boldsymbol{Q}\|,\end{aligned}$$

we have used that $\boldsymbol{H}^T\boldsymbol{H} = \boldsymbol{C}$ (Dryden and Mardia, 2016, page 63) where $\boldsymbol{C}$ is the centering matrix (Dryden and Mardia, 2016, Equation 2.3) which has no effect on $\boldsymbol{Q}$ as it is centered, meaning $\boldsymbol{C}\boldsymbol{Q} = \boldsymbol{Q}$. □

Result 2.2.3. *For an $m \times m$ symmetric matrix $\boldsymbol{S} = \boldsymbol{S}^T$ then*

$$\|vech^*(\boldsymbol{S})\| = \|\boldsymbol{S}\|$$

Proof.

$$\begin{aligned}\|\text{vech}^*(\boldsymbol{S})\| &= (\sum_{i=1}^{m}\sum_{j<i}(\sqrt{2}s_{ij})^2 + \sum_{i=1}^{m} s_{ii}^2)^{\frac{1}{2}} \\ &= (\sum_{i=1}^{m}\sum_{j<i} 2s_{ij}^2 + \sum_{i=1}^{m} s_{ii}^2)^{\frac{1}{2}} \\ &= (\sum_{i=1}^{m}\sum_{j=1}^{m} s_{ij}^2)^{\frac{1}{2}} \\ &= \|\boldsymbol{S}\|.\end{aligned}$$

□

Result 2.2.4. *For the Euclidean power metric the Euclidean distance in the embedding space is conserved in the tangent space, meaning that for $\boldsymbol{Q}_1 = F_\alpha(\boldsymbol{L}_1)$ and $\boldsymbol{Q}_2 =$*

$F_\alpha(\boldsymbol{L}_2)$ where $\boldsymbol{L}_1, \boldsymbol{L}_2 \in \mathcal{L}_m$ and $\boldsymbol{\nu}$ is the chosen pole, then

$$\|\exp_\nu^{-1}(\boldsymbol{Q}_1) - \exp_\nu^{-1}(\boldsymbol{Q}_2)\| = \|\boldsymbol{Q}_1 - \boldsymbol{Q}_2\|.$$

Proof. The proof for this result relies on Result 2.2.2 and 2.2.3,

$$\begin{aligned}
\|\exp_\nu^{-1}(\boldsymbol{Q}_1) - \exp_\nu^{-1}(\boldsymbol{Q}_2)\| &= \|\text{vech}^*(\boldsymbol{H}(\boldsymbol{Q}_1 - \boldsymbol{\nu})\boldsymbol{H}^T) - \text{vech}^*(\boldsymbol{H}(\boldsymbol{Q}_2 - \boldsymbol{\nu})\boldsymbol{H}^T)\| \\
&= \|\text{vech}^*(\boldsymbol{H}(\boldsymbol{Q}_1 - \boldsymbol{\nu})\boldsymbol{H}^T - \boldsymbol{H}(\boldsymbol{Q}_2 - \boldsymbol{\nu})\boldsymbol{H}^T)\| \\
&= \|(\boldsymbol{H}(\boldsymbol{Q}_1 - \boldsymbol{\nu})\boldsymbol{H}^T) - (\boldsymbol{H}(\boldsymbol{Q}_2 - \boldsymbol{\nu})\boldsymbol{H}^T)\| \\
&= \|(\boldsymbol{H}(\boldsymbol{Q}_1 - \boldsymbol{\nu} - \boldsymbol{Q}_2 + \boldsymbol{\nu})\boldsymbol{H}^T)\| \\
&= \|(\boldsymbol{H}(\boldsymbol{Q}_1 - \boldsymbol{Q}_2)\boldsymbol{H}^T)\| \\
&= \|\boldsymbol{Q}_1 - \boldsymbol{Q}_2\|.
\end{aligned}$$

□

2.2.5 Projection

We carry out analysis in the tangent space, e.g. computing a sample mean, and so results are found in this space. After inverting the tangent space projection and inverting the embedding for results the results still may not lie in the graph Laplacian space, $\mathcal{L}_m$. So we are interested in projecting from the matrix space $\mathcal{M}_m$, defined for different metrics in Section 2.2.3, to the space of graph Laplacians, $\mathcal{L}_m$, as results can only be interpreted in this space. We seek a P that maps $\boldsymbol{Y} = (y_{ij}) \in \mathcal{M}_m$ to the "closest point" in $\mathcal{L}_m$. For the Euclidean and Procrustes power metric such projections are

$$\begin{aligned}
\mathrm{P}_\alpha(\boldsymbol{Y}) &= \arg\inf_{\boldsymbol{L}\in\mathcal{L}_m} d_\alpha(\boldsymbol{Y}, \boldsymbol{L}) \\
\mathrm{P}_{\alpha,S}(\boldsymbol{Y}) &= \arg\inf_{\boldsymbol{L}\in\mathcal{L}_m} d_{\alpha,S}(\boldsymbol{Y}, \boldsymbol{L}).
\end{aligned} \tag{2.2.6}$$

For certain $\alpha \neq 1$ when $\boldsymbol{Y} \notin \mathcal{PSD}_m$ the distances may not be defined, however for these α values the reverse embeddings, defined in Section 2.2.3, ensures $\boldsymbol{Y} \in \mathcal{PSD}_m$ so the distances will be defined. It is desirable that when computing the projection we have a convex optimisation problem, defined in (1.2.7), so the local minimum will be the unique global minimum (Rockafellar, 1993).

Result 2.2.5. *For P_α with $\alpha = 1$ the projection can be found by solving a convex*

optimisation problem guaranteeing a unique solution.

Proof. The projection can be written as minimising

$$\begin{aligned} f(\boldsymbol{Y}) &= d_1^2(\boldsymbol{L}, \boldsymbol{Y}) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m}(l_{ij} - y_{ij})^2 \\ \text{subject to:} \quad l_{ij} - l_{ji} &= 0, \quad 0 \leq i, j \leq m \\ \sum_{j=1}^{m} l_{ij} &= 0, \quad 0 \leq i \leq m \\ l_{ij} &\leq 0, \quad 0 \leq i, j \leq m \text{ and } i \neq j. \end{aligned} \tag{2.2.7}$$

To prove this optimisation is convex we first note the constraints are all convex, as they are all linear functions. We then need to prove the function we are minimising is convex, which is

$$\sum_{i=1}^{m}\sum_{j=1}^{m}(l_{ij} - y_{ij})^2 = (\boldsymbol{l} - \boldsymbol{y})^T(\boldsymbol{l} - \boldsymbol{y}),$$

where $\boldsymbol{l} = \text{vec}(\boldsymbol{L})$ and $\boldsymbol{y} = \text{vec}(\boldsymbol{Y})$, where vec is defined in (0.0.1). To prove this is convex we must calculate the Hessian by differentiating the function twice,

$$\begin{aligned} \frac{\partial(\boldsymbol{l} - \boldsymbol{y})^T(\boldsymbol{l} - \boldsymbol{y})}{\partial \boldsymbol{l}} &= \frac{\partial(\boldsymbol{l}^T\boldsymbol{l} - \boldsymbol{l}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{l} + \boldsymbol{y}^T\boldsymbol{y})}{\partial \boldsymbol{l}} \\ &= 2\boldsymbol{l}^T - 2\boldsymbol{y}^T \\ \frac{\partial^2(\boldsymbol{l} - \boldsymbol{y})^T(\boldsymbol{l} - \boldsymbol{y})}{\partial \boldsymbol{l}^T \partial \boldsymbol{l}} &= 2\boldsymbol{I}_{m^2}. \end{aligned}$$

The Hessian is thus $2\boldsymbol{I}_{m^2}$ which is positive definite meaning this function is strictly convex. □

The natural projection for each metric would minimise their respective distance to $\mathcal{L}_m$, as in (2.2.6). However for $\alpha \neq 1$ the optimization is not in general convex. Therefore as the projection for the Euclidean power metric with $\alpha = 1$, defined in (2.2.7), involves convex optimisation we will use the projection P_1 from now on for all our metrics and we will refer to this projection as $\text{P}_{\mathcal{L}}$. To implement the projection, P_1, we can, for example, use either the `CVXR` (Fu et al., 2018) or `rosqp` (Anderson, 2018) packages in R (R Core Team, 2018) to solve the optimisation. `rosqp` is particularly fast even

for large dimensions, such as $m = 1000$. As the unique global solution can be found computationally then for $\boldsymbol{Y}_1, \boldsymbol{Y}_2 \in \mathcal{M}_m$ if $\boldsymbol{Y}_1 = \boldsymbol{Y}_2$ then $\mathrm{P}_{\mathcal{L}}(\boldsymbol{Y}_1) = \mathrm{P}_{\mathcal{L}}(\boldsymbol{Y}_2)$ so the projection is unique, clearly this implication only holds one way as the projection is many to one.

2.3 Means

Now our framework is defined we can define the mean of a set of graph Laplacians.

We define the population mean for graph Laplacians as

$$\begin{aligned} \mu &= \mathrm{P}_{\mathcal{L}}(\eta), \\ \text{where } \eta &= \arg\inf_{u \in \mathcal{PSD}_m} \boldsymbol{E}[d^2(\boldsymbol{L}, u)], \end{aligned} \tag{2.3.1}$$

assuming μ exists. The sample mean for a set of graph Laplacians is then defined as

$$\begin{aligned} \hat{\mu} &= \mathrm{P}_{\mathcal{L}}(\hat{\eta}), \\ \text{where } \hat{\eta} &= \arg\inf_{u \in \mathcal{PSD}_m} \frac{1}{n}\sum_{k=1}^{n} d^2(\boldsymbol{L}_k, u). \end{aligned} \tag{2.3.2}$$

The sample mean is the sample Fréchet mean in the embedding space, defined in (1.1.3), that has had the inverse embedding applied. However $\hat{\eta}$ is not guaranteed to be a graph Laplacian and so the projection is then used to guarantee the result will lie in $\mathcal{L}_m$.

For the Euclidean power distance we have

$$\begin{aligned} \eta &= \mathrm{F}_{\alpha}^{-1}\left(\mathbb{E}[\mathrm{F}_{\alpha}(\boldsymbol{L})]\right) \\ \hat{\eta} &= \mathrm{F}_{\alpha}^{-1}\left(\frac{1}{n}\sum_{k=1}^{n}\mathrm{F}_{\alpha}(\boldsymbol{L}_k)\right) \\ &= \left(\frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{L}_k)^{\alpha}\right)^{\frac{1}{\alpha}}, \end{aligned}$$

and μ and $\hat{\mu}$ are unique in this case. For the Procrustes power distance μ and $\hat{\mu}$ may be sets, and the conditions for uniqueness rely on the support and curvature of the space (Le, 1995). Result 13.1 of Dryden and Mardia (2016) proven by Kendall (1990) states if the support of the distribution is a geodesic ball B_r then there exists a unique mean in B_r, this holds for our data and so we can assume uniqueness. A stronger condition

for global uniqueness is if the support of the distribution is a geodesic ball B_r such that B_{2r} is regular then the mean is unique even outside of B_r.

A regular geodesic ball, $B_r(p)$, of radius r centred at p has the cut locus of p not meet the ball $B_r(p)$ and the supremum of the sectional curvature of the ball must be less than $(\frac{\pi}{2r})^2$ (Dryden and Mardia, 2016, definition 13.2). Example 13.1 in Dryden and Mardia (2016) gives a method for checking this assumption for size-and-shape space. The guarantee of a unique mean within the support of the data is all we require as we are not interested in values outside the support of the data, and therefore we do not consider proving a global mean.

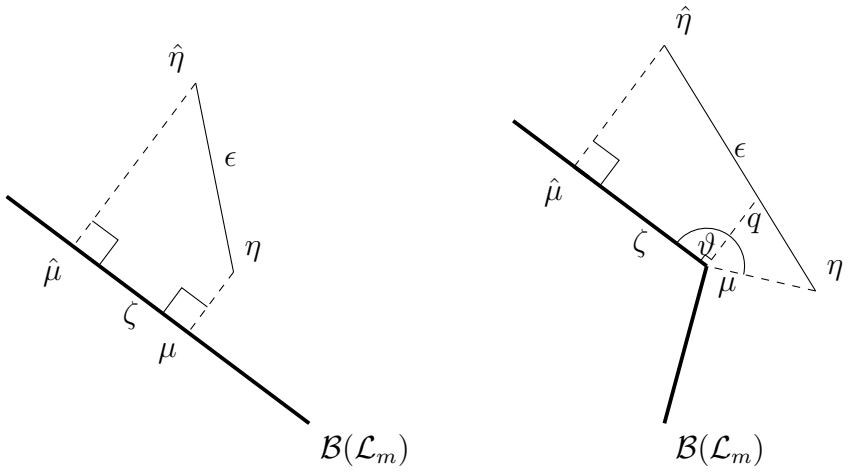
We have seen in Section 1.1 that there are two classes of means on a manifold, the intrinsic mean and extrinsic mean defined in (1.1.4). Our mean in the graph Laplacian space is an extrinsic mean in general. Although for the Euclidean power metric when $\alpha = 1$, we have $\hat{\mu} = \hat{\eta}$ and the mean is a Fréchet intrinsic mean (Fréchet, 1948; Ginestet et al., 2017) in this case.

Result 2.3.1. *Let $\boldsymbol{L}_k$, $k = 1, \ldots, n$, be a random sample of i.i.d. observations from a distribution with population mean μ in (2.3.1). For the power Euclidean distance d_α the estimator $\hat{\mu}$, in (2.3.2), is a consistent estimator of μ.*

Proof. For an estimator $\hat{\mu}$ to be consistent for a population mean μ, it must converge in probability to μ as $n \to \infty$. Let $\{\hat{\mu}_n\}$ be a sequence of estimates from a sample set $\{\boldsymbol{L}_1, \ldots, \boldsymbol{L}_n\}$, for this to converge in probability to μ then for any $\epsilon > 0$ and any $\delta > 0$ there exists a number N such that for all $n \geq N$ $P_n < \delta$, where $P_n = P(|\hat{\mu}_n - \mu| > \epsilon)$.

From the law of large numbers $\frac{1}{n}\sum_{k=1}^{n}(\mathrm{F}_\alpha(\boldsymbol{L}_k)$ converges in probability to $\mathbb{E}[\mathrm{F}_\alpha(\boldsymbol{L})]$ and hence $\hat{\eta} = \left(\frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{L}_k)^\alpha\right)^{\frac{1}{\alpha}}$ converges in probability to $\eta = (\mathbb{E}[\mathrm{F}_\alpha(\boldsymbol{L})])^{\frac{1}{\alpha}}$, by the continuous mapping theorem as long as η exists and is unique.

We now need to show the convergence in probability holds when we project $\hat{\eta}$ and η to $\mathcal{L}_m$. If the projection is not needed for η then clearly the convergence will hold. As $\mathcal{L}_m \subset \mathcal{M}_m$ then when the projection is needed it will always project to the boundary of $\mathcal{L}_m$ denoted $\mathcal{B}(\mathcal{L}_m)$. To have convergence in probability of $\hat{\eta}$, for any $\epsilon > 0$ and $\delta > 0$ there must exists an N_1 such that for $n \geq N_1$ then $P(|\hat{\mu} - \mu| > \epsilon) < \delta$. We know from the convergence in probability of $\hat{\eta}$, that for any $\epsilon > 0$ and $\delta > 0$ there exists an N_2 such that for $n \geq N_2$, $P(|\hat{\eta} - \eta| > \epsilon) < \delta$. We choose an ϵ small enough so that the boundary of the graph Laplacian space can be thought to have 0 curvature. From Ginestet et al. (2017) we know $\mathcal{L}_m$ is a manifold with corners, and stated briefly a d



(a) Case 1 **(b)** Case 2

Figure 2.4: *Schematic to the two cases used to prove if there is convergence between the sample and population unprojected means then there is convergence between the sample and population projected means.*

dimensional manifold with corners can be locally modelled by $[0, \infty)^k \times \mathbb{R}^{d-k}$, for full details see Joyce (2009). Let $|\hat{\eta} - \eta| = \epsilon$ and $|\hat{\mu} - \mu| = \zeta$. This leads to two cases, shown in Figure 2.4:

- Case 1: μ is not on a corner of $\mathcal{B}(\mathcal{L}_m)$. In this case the estimator behaves as in Figure 2.4a. The estimator $\hat{\eta}$ is orthogonally projected onto $\hat{\mu}$, hence due to Pythagoras' theorem it is clear $\zeta \leq \epsilon$.
- Case 2: μ is on a corner of $\mathcal{B}(\mathcal{L}_m)$. In this case the estimator behaves as in Figure 2.4b. Clearly $\frac{\pi}{2} \leq \vartheta \leq \pi$. We consider a point q along the line between $\hat{\eta}$ and η such that the angle between $\hat{\mu}$, μ and q is $\frac{\pi}{2}$. Note $\zeta \leq |\hat{\eta} - q|$ following identical arguments as in case 1, and clearly $|\hat{\eta} - q| \leq \epsilon$. Hence $\zeta \leq \epsilon$.

We do not consider when $\hat{\mu}$ is on a corner when μ is not on a corner as for small enough ϵ this will not occur. We now have for $n \geq N_2$ that $\zeta \leq \epsilon$, hence

$$\delta > P(|\hat{\eta} - \eta| > \epsilon) = P(|\hat{\mu} - \mu| > \zeta) \geq P(|\hat{\mu} - \mu| > \epsilon).$$

Therefore when $n \geq \max(N_1, N_2)$ then $P(|\hat{\mu} - \mu| > \epsilon) < \delta$ and so $\{\hat{\mu}\}$ converges in probability to μ as $n \to \infty$, i.e. $\hat{\mu}$ is a consistent estimator. □

A similar result to Result 2.3.1 holds for $d_{\alpha,S}$ where stronger conditions for consistency of $\hat{\eta}$ are given in Bhattacharya and Patrangenaru (2003), but an identical projection

argument used in the proof for d_α holds.

(a) Austen Euclidean mean

(b) Dickens Euclidean mean

Example 2.3.1: Means of the Austen and Dickens novel data

Figure 2.5 shows an illustration of the sample means for Austen and Dickens novels using d_1, $d_{\frac{1}{2}}$ and $d_{\frac{1}{2},S}$ with the 1000 words arranged in a grid and edges drawn between words which co-occur with weight $w_{ij} \geq 10^{-5} \sum_{k=1}^{m} w_{kk}$. The means for the different metrics all look very similar for both authors, perhaps unsurprisingly as approximately half of the words in each novel are represented by the first 50 words. These plots demonstrate how for large networks it is hard to visually compare the means of samples of networks thus motivating our two sample test for equality of means in Section 4 and

(c) Austen square root Euclidean mean

(d) Dickens square root Euclidean mean

(e) Austen Procrustes size-and-shape mean

(f) Dickens Procrustes size-and-shape mean

Figure 2.5: *The mean for both Austen's novels and Dickens novels using* d_1, $d_{\frac{1}{2}}$ *and* $d_{\frac{1}{2},S}$ *based on the top m=1000 word pairs. Zoom in for more detail.*

more specifically the method in Section 4.6.1 to explore specific differences in mean co-occurrences for the authors. These plots are drawn using the program `Cytoscape` (Shannon et al., 2003) and more detail can be seen by magnifying the view to a large extent, for example there are more co-occurrences of *she*, *her* by Austen and *the*, *his*, *don't* by Dickens.

Example 2.3.2: Means of the M-money transaction data
For the M-money transaction networks, described in Section 1.3.2, we use our framework to calculate the mean daily network for d_1, $d_{\frac{1}{2}}$ and $d_{\frac{1}{2},S}$, found in Figure 2.6. The mean networks are plotted on the map of Tanzania, with a node plotted at the centre of the corresponding district it represents. The nodes are sized proportionally to their degree and edges are only drawn between words which co-occur with weight $w_{ij} \geq 10^{-4} \sum_{k=1}^{m} w_{kk}$. The mean networks look very similar for each metric, with the majority of transactions involving districts on the west around Dar es Salaam, a major Tanzanian city. This is expected as this is the most populated area. Another node contributing to a large proportion of transaction is Dodoma Urban which is the capital of Tanzania.

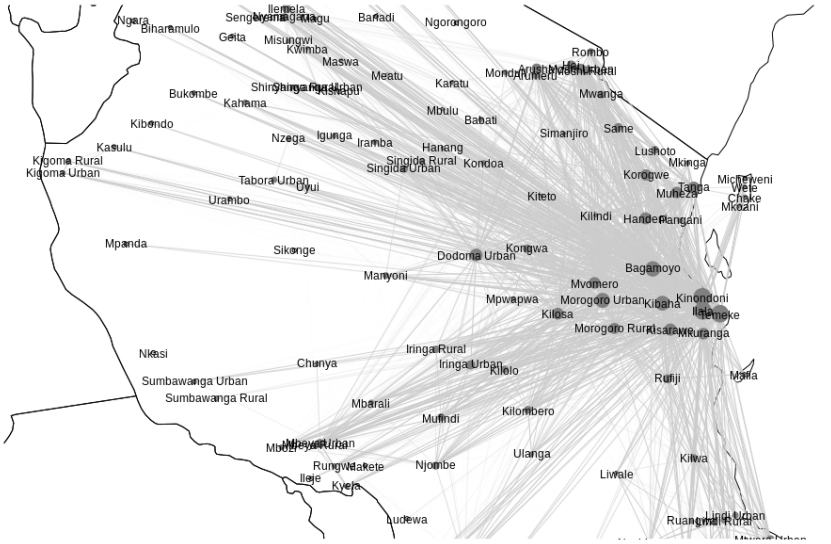
2.4 Geodesics and interpolation

We now consider an interpolation path, $\boldsymbol{L}(c)$, where c is the position along the path, $0 \leq c \leq 1$, between the graph Laplacians at $\boldsymbol{L}(0)$ and $\boldsymbol{L}(1)$. For $c < 0$ and $c > 1$ the path $\boldsymbol{L}(c)$ is extrapolating from the graph Laplacians at $\boldsymbol{L}(0)$ and $\boldsymbol{L}(1)$. The interpolation and extrapolation path between graph Laplacians for each metric is defined by first finding the geodesic path in the tangent space between the embedded graph Laplacians, which is then projected to $\mathcal{L}_m$. This is given by

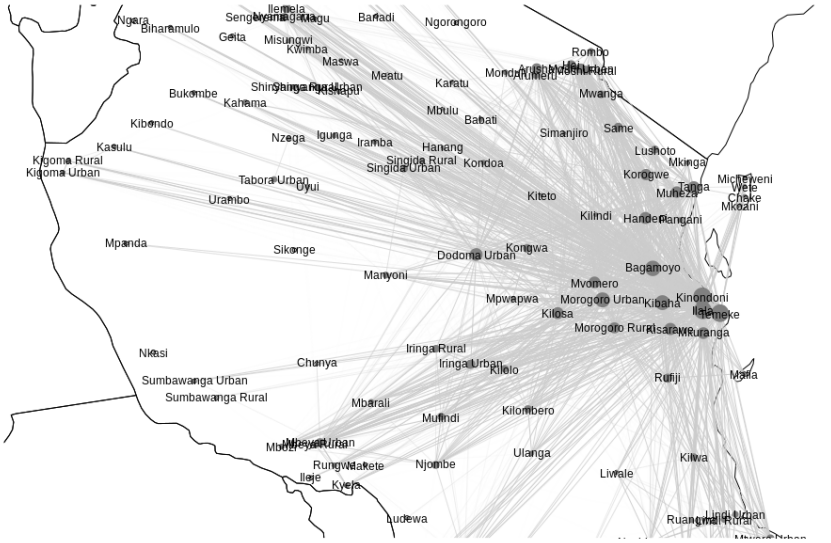
$$\boldsymbol{L}(c) = \mathrm{P}_{\mathcal{L}}(\mathrm{F}_{\alpha}^{-1}(\exp_{\nu}\{c \exp_{\nu}^{-1}(\mathrm{F}_{\alpha}(\boldsymbol{L}_2))\})), \tag{2.4.1}$$

where $\boldsymbol{L}_1 = \mathrm{P}_{\mathcal{L}}(\mathrm{F}_{\alpha}^{-1}(\boldsymbol{\nu}))$.

(a) Euclidean mean



(b) Square root Euclidean mean



(c) Procrustes size-and-shape mean

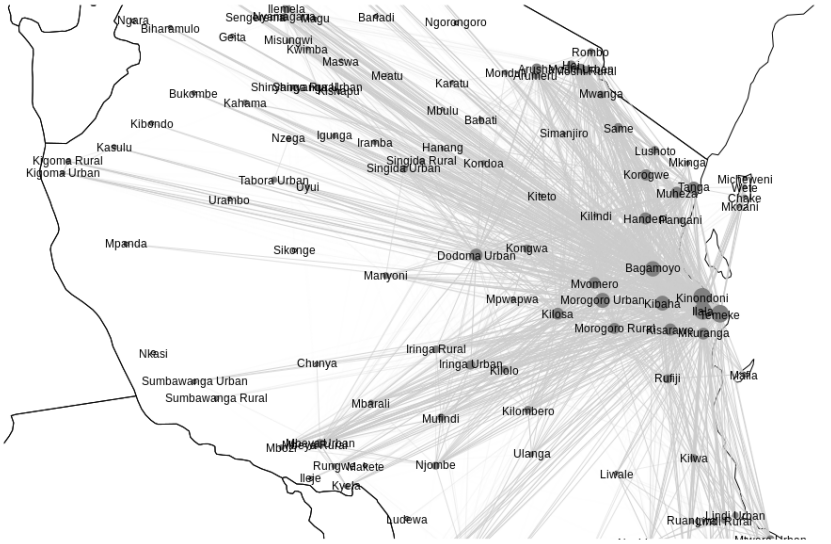


Figure 2.6: *The mean for the M-money networks using* d_1, $d_{\frac{1}{2}}$ *and* $d_{\frac{1}{2},S}$.

For the Euclidean power metrics this can be simplified

$$\begin{aligned}\boldsymbol{L}(c) &= \mathrm{P}_{\mathcal{L}}(\mathrm{F}_{\alpha}^{-1}(\exp_{\nu}\{c\exp_{\nu}^{-1}(\mathrm{F}_{\alpha}(\boldsymbol{L}_2))\})),\\ &= \mathrm{P}_{\mathcal{L}}(\mathrm{F}_{\alpha}^{-1}(\mathrm{F}_{\alpha}(\boldsymbol{L}_1) + c\mathrm{F}_{\alpha}(\boldsymbol{L}_2) - c\mathrm{F}_{\alpha}(\boldsymbol{L}_1)))\\ &= \mathrm{P}_{\mathcal{L}}(\mathrm{F}_{\alpha}^{-1}(c\mathrm{F}_{\alpha}(\boldsymbol{L}_2) - (1-c)\mathrm{F}_{\alpha}(\boldsymbol{L}_1))).\end{aligned} \tag{2.4.2}$$

Therefore the interpolation path for the Euclidean power metric is just the geodesic in the embedding space, given in (2.4.3), projected back into $\mathcal{L}_m$.

$$c\mathrm{F}_{\alpha}(\boldsymbol{L}_2) + (1-c)\mathrm{F}_{\alpha}(\boldsymbol{L}_1). \tag{2.4.3}$$

Note when $\alpha = 1$ and $0 \leq c \leq 1$ the projection is not required, as the geodesic path actually lies in $\mathcal{L}_m$, and is

$$c\boldsymbol{L}_2 + (1-c)\boldsymbol{L}_1.$$

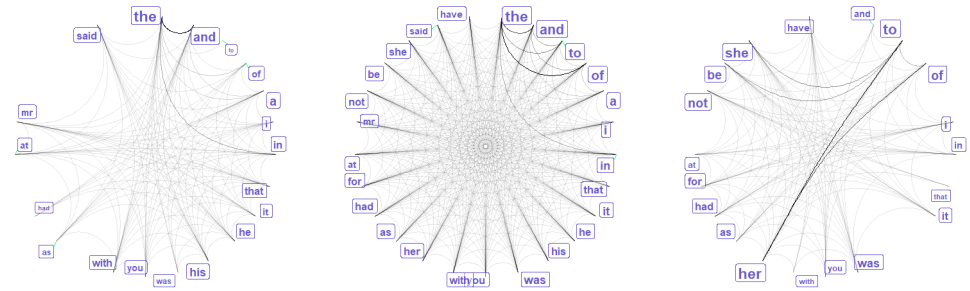
Example 2.4.1: Interpolation and extrapolation for the Austen and Dickens novel data

Figure 2.7 shows the interpolation and extrapolation paths between the mean Austen and mean Dickens novels, when using d_1, $d_{\frac{1}{2}}$ and $d_{\frac{1}{2},S}$. The plots only include the 25 nodes corresponding to the most frequent words out of $m = 1000$ nodes. The size of a node in the networks is proportional to its degree and the thickness of edges proportional to their weight. For each metric the paths look very similar. For $c = 0.5$ the network shown is the mean network between the mean of the Dickens and Austen mean networks. At $c = 6$ we are extrapolating past Austen's mean network and the feminine words have larger degrees and their edges have larger weights, for example '*her*' to '*to*' and '*of*', and '*she*' to '*to*'. For $c = -5$ we are extrapolating past Dickens mean and the nodes for '*she*' and '*her*' are actually removed indicating they have degree 0, which is further evidence of the fact Austen used feminine words more than Dickens.

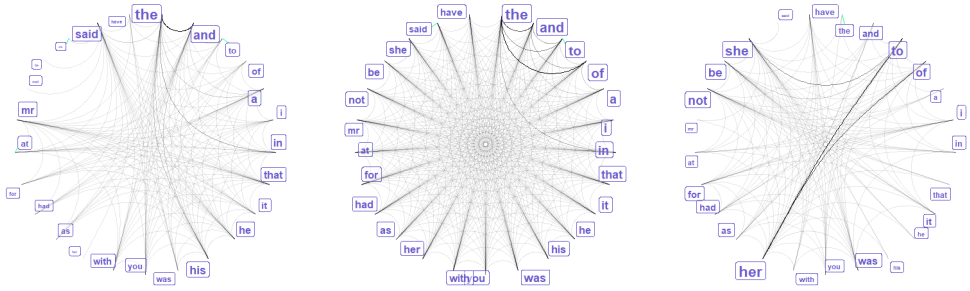
2.5 Principal component analysis

Principal component analysis (PCA) is a useful statistical method for describing dominant modes of variability within a dataset. However the method relies on the data lying in a Euclidean space. We will use the tangent space of graph Laplacians to perform

(a) $c = -5$ **(b)** $c = 0.5$ **(c)** $c = 6$



(d) $c = -5$ **(e)** $c = 0.5$ **(f)** $c = 6$



(g) $c = -5$ **(h)** $c = 0.5$ **(i)** $c = 6$

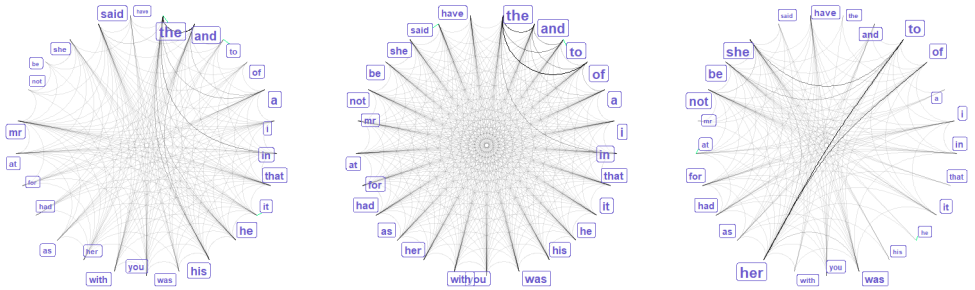


Figure 2.7: *Interpolation ($c = 0.5$) and extrapolation ($c = -5, c = 6$) networks between Dickens' and Austen's mean novels using d_1 in a), b) and c), and $d_{\frac{1}{2}}$ in d), e) and f) and $d_{\frac{1}{2},S}$ in g), h) and i) . The top 25 words are displayed where the mean novels for the authors are estimated using the respective metric and $m = 1000$.*

PCA in and then project results back to the space of graph Laplacians.

To perform PCA on graph Laplacians we let $\boldsymbol{v}_k = \exp_\nu^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_k))$, where $\boldsymbol{\nu} = \mathrm{F}_\alpha(\hat{\eta})$ for $\hat{\eta}$ defined in (2.3.2) using either the Euclidean or Procrustes power metric. Then we define $\boldsymbol{S} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{v}_k \boldsymbol{v}_k^T$, which is an estimated covariance matrix. Suppose $\boldsymbol{S}$ is of rank r, with non-zero eigenvalues, $\lambda_1, \ldots, \lambda_r$, then the corresponding eigenvectors $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_r$, are the PCs in the tangent space, and the PC scores are

$$s_{kj} = \boldsymbol{\gamma}_j^T \boldsymbol{v}_k, \quad \text{for } k = 1, \ldots, n, \quad j = 1, \ldots, r. \tag{2.5.1}$$

The path of the jth PC in $\mathcal{L}_m$ is

$$\boldsymbol{L}(c) = \mathrm{P}_{\mathcal{L}}(\mathrm{F}_\alpha^{-1}(\exp_\nu(c\lambda_j^{\frac{1}{2}}\boldsymbol{\gamma}_j))), \quad c \in \mathbb{R}. \tag{2.5.2}$$

When the Euclidean power metric is used and $\alpha = 1$ is chosen, the importance of the ith node in the principal component $\boldsymbol{\gamma}$ is the proportion of the sum of the absolute diagonals each node has in a principal components when it is projected back into the embedding space, given as

$$\frac{\exp_\nu(\boldsymbol{\gamma})_{ii}}{(\sum_{j=1}^{m} \exp_\nu(\boldsymbol{\gamma})_{jj})}, \text{ for } 1 \leq i \leq m. \tag{2.5.3}$$

These importances can be negative. A large negative importance indicates the node is indicative a negative coordinate for the principal component. This method for finding node importances does not hold for any metric other than d_1. This is as the principal components in the embedding space for the other metrics do not have an interpretable connection to each node in a network as the inverse embedding is not just the identity projection as it is when d_1 is used. When using a metric other than d_1 importances of nodes can be found by extrapolating along the PC path of networks.

After the PC space has been found for the collection of graph Laplacians $\boldsymbol{L}_k$ for $1 \leq k \leq n$ it is useful to project other graph Laplacians into this space, for example in Section 5.1.2. A graph Laplacian $\boldsymbol{L}_{new}$, that was not used to find the PC space, can be projected into it and will have the jth PC score as

$$s_j = \gamma_j^T \boldsymbol{v}_{new},$$

where $\boldsymbol{v}_{new} = \exp_\nu^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_{new}))$.

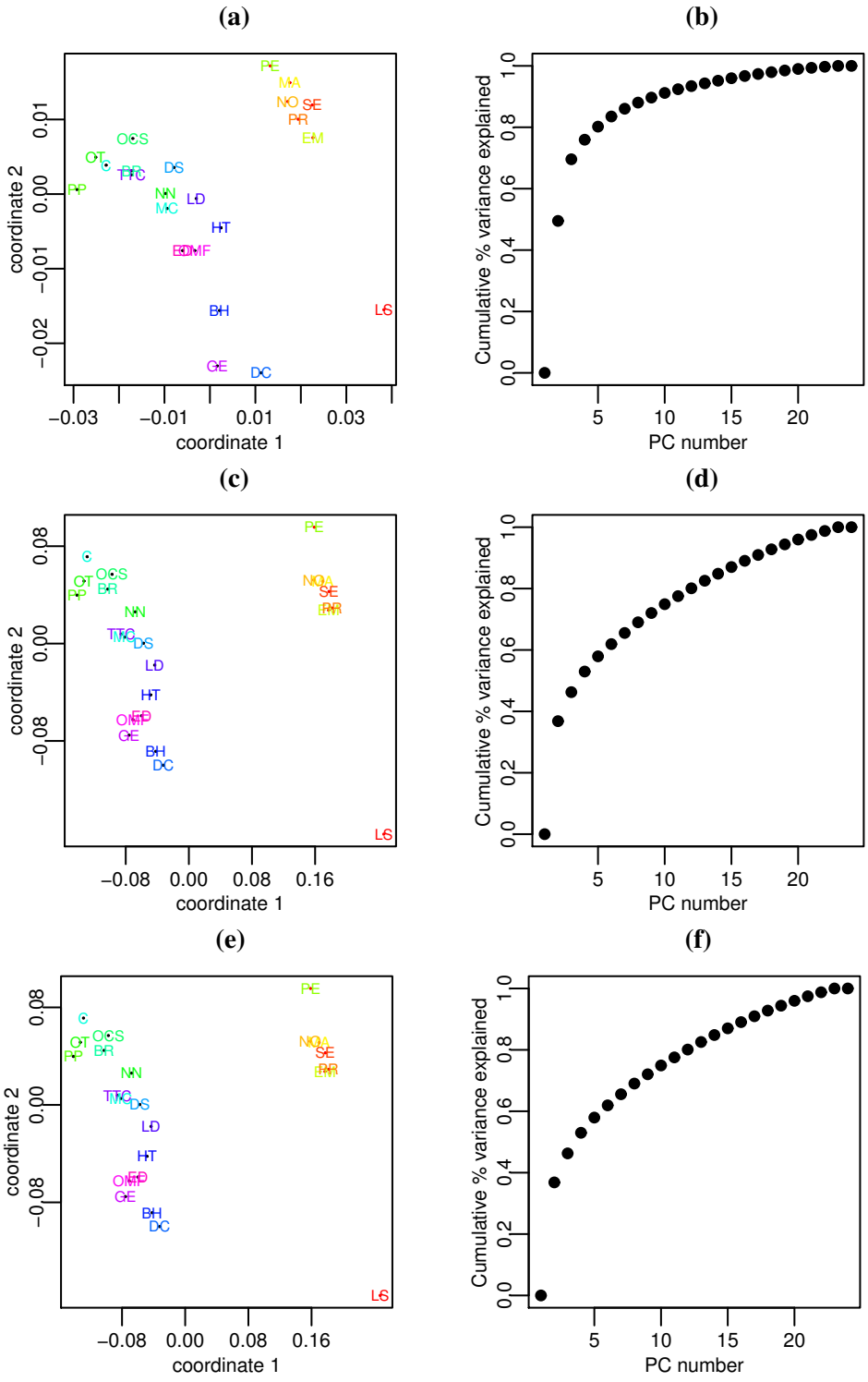


Figure 2.8: *Plot of PC 1 and PC 2 scores (left) for the Austen and Dickens novels, coloured in time order (red to green for Austen novels and green to violet for Dickens novels) and plot of the cumulative variance explained by each PC (right), using the (top to bottom) Euclidean, square root Euclidean and Procrustes size-and-shape metric. The abbreviations for novels are found in Table 1.2.*

Example 2.5.1: PCA applied to the Austen and Dickens novel data

We now apply the methods of PCA to the Austen and Dickens text data, for $m = 1000$. The first and second PC scores are plotted in Figure 2.8 for the Euclidean, square root Euclidean and Procrustes size-and-shape metric. The plots look very similar for all metrics, in fact they appear visibly identical between the square root and Procrustes metrics. The cumulative variance explained by each PC for each metric is also in Figure 2.8. The variance explained by PC1 and PC 1 and 2 together is 49% and 70%, 37% and 46% and 37% and 46% for the Euclidean, square root Euclidean and Procrustes size-and-shape respectively. Clearly the Euclidean metric is minimising the variance best when using 2 coordinates. A benefit of the square root Euclidean and Procrustes size-and-shape metric is clear here as they separate the Austen and Dickens novels with a large gap on PC1 where as *David Copperfield* (DC) and *Persuasion* (PE) are very close in PC1 for the Euclidean. For all metrics we can see Lady Susan looks like an anomaly for Jane Austen's writing as it very far from the cluster of Austen's other works. We now analyse the Euclidean PCs in more detail, to interpret what the principal components are actually measuring.

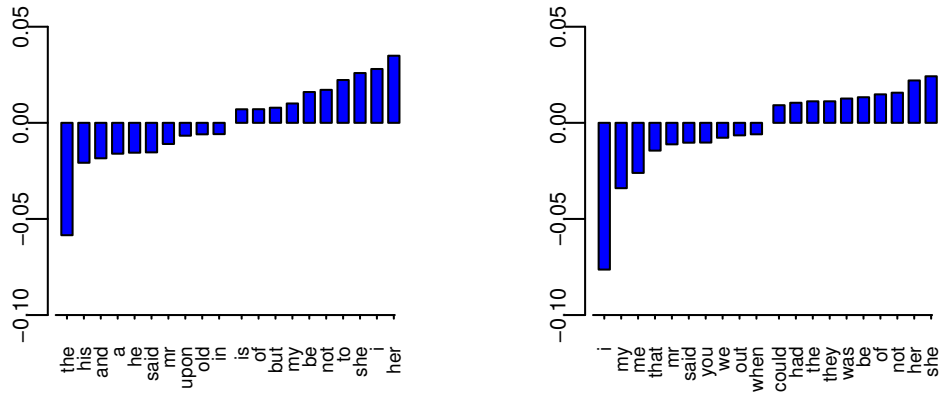


Figure 2.9: *The importance of each word given by (2.5.3) in (left) PC 1 and (right) PC 2.*

Figure 2.9 contains plots representing the importance and sign of each word in the first and second Euclidean PC. From Figure 2.8 a more positive PC 1 score is indicative of an Austen novel whilst a more negative one a Dickens novel. For a positive PC1 score the nodes '*her*' and '*she*' have importance whilst for a negative score words such as '*his*', and '*he*' have more importance, which is expected as Austen writes with more

female characters. The second PC actually is similar to a fitted regression line which we describe in Chapter 3, but even without this we can see from the colouring of the novels that Austen novels over time have the second PC increasing, as *Lady Susan* (LS) and *Persuasion* (PE) are her earliest and latest novels respectively. This is the opposite to Dickens where PC2 decreases with time. *Pickwick papers* (PP) is Dickens earliest and *The Mystery of Edwin Drood* (ED) his latest. The second PC has feminine words like '*her*' and '*she*' as the most positive words, but more first and second person words, such as '*I*', '*my*' and '*you*' as negative words. This is consistent with Austen increasingly using a stylistic device called "free indirect speech" in her later novels (Shaw, 1990). "Free indirect speech" has the property the third person pronouns, such as '*she*' and '*her*' are used instead of first person pronouns, such as '*I*' and '*my*'.

Example 2.5.2: PCA applied to M-money transaction data
We also apply our PCA method to the M-money data networks. The plots for PC 1 and 2 scores and the cumulative variance explained by each PC are found in Figure 2.10, for the Euclidean, square root Euclidean and Procrustes size-and-shape metrics. The plots have the networks corresponding to Saturdays and Sundays colored red and blue respectively, as we hypothesis networks on these days will differ to weekdays, something we investigate further in Example 4.6.2. When using the Euclidean metric the networks for Sundays seem to cluster in the bottom left, but for the other two metrics the Sundays do not cluster as clearly. For Saturdays there appears to be no clustering regardless of the metric used. From the cumulative variance plots it is clear that the first two PCs are not explaining a large percentage of the variance for any metric, and so a lot of information is being lost in the 2D plots.

2.6 Summary

In this chapter we have proposed a novel framework for the statistical analysis of networks by representing them as graph Laplacians. This framework is very general and whilst we have only defined it for two types of metrics, the Euclidean power and Procrustes power metric it generalises to other metrics. We have shown how using this framework we can perform standard statistical methods on samples of graph Laplacians such as calculating the mean and PCA. We shall use this framework in the following chapters to consider regression, two-sample testing and classification of samples of networks.

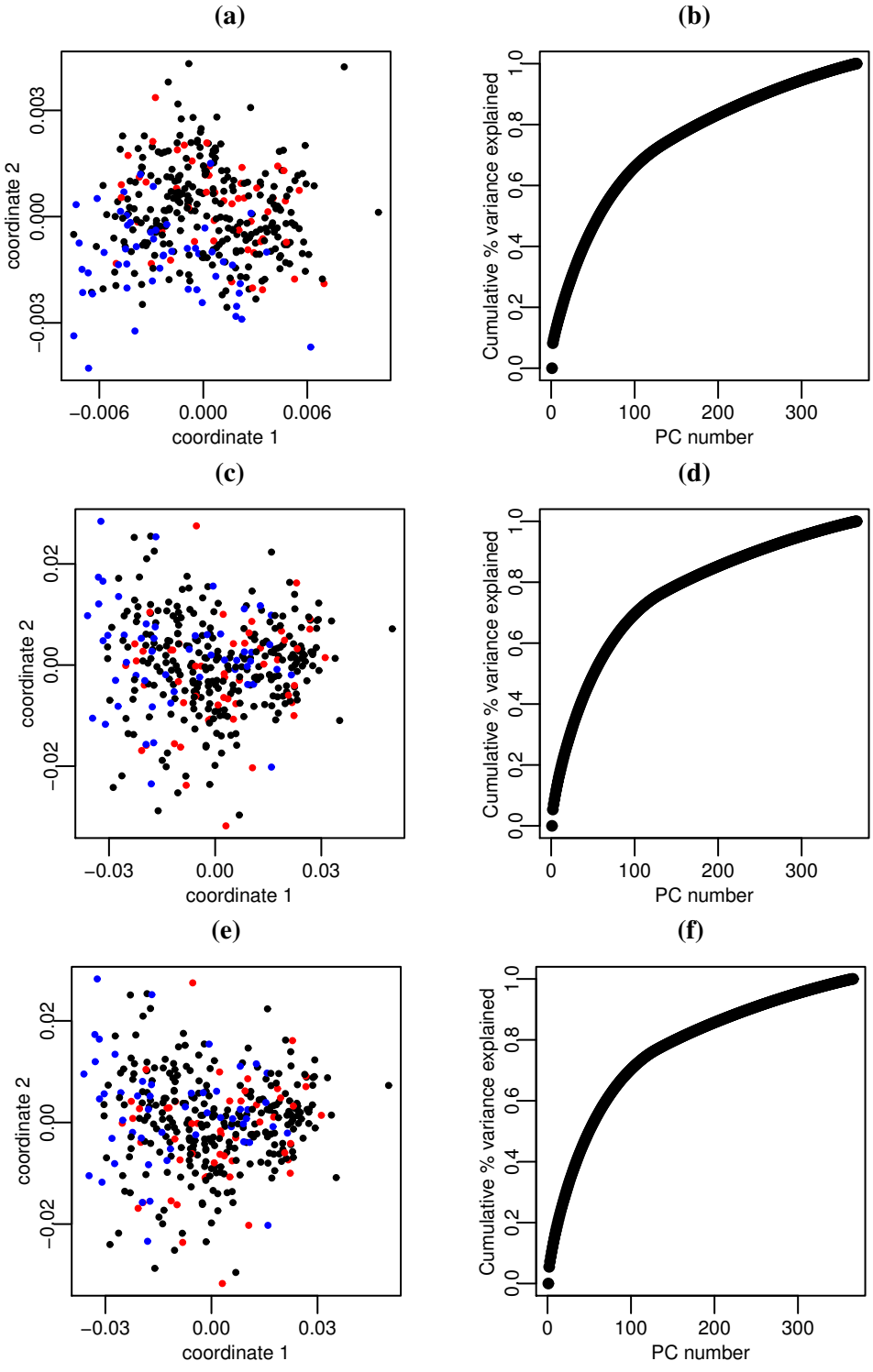


Figure 2.10: *(left) Plot of PC 1 and PC 2 scores for the M-money networks, weekdays are coloured black, Saturdays red and Sundays blue and (right) plot of the cumulative variance explained by each PC , using the (top to bottom) Euclidean, square root Euclidean and Procrustes size-and-shape metric.*

We have seen in the examples so far that the square root Euclidean and Procrustes size-and-shape metrics often give visibly similar if not identical results. This is suggesting that the rotation term allowed in the Procrustes size-and-shape metric is often similar to the identity matrix as very little rotation is occurring. In the following chapters if the square root Euclidean metric and Procrustes size-and-shape metric give visibly identical results we shall only include a plot for one and then we shall state the similarity between the two.

CHAPTER 3

Regression of graph Laplacians

In this chapter we will use the statistical framework we set up for graph Laplacians to study a broad range of regression problems. A main motivation of the regression methods we define is for regression over time for dynamic network datasets. The study of dynamic networks has recently increased as more data of this type is becoming available (Rastelli et al., 2018). An example of previous work on dynamic network data is Friel et al. (2016) which embedded nodes of bipartite dynamic networks in a latent space. Friel et al. (2016) used this embedding to study the interlocks in a bipartite network over time, motivated by networks representing the connection of leading Irish companies and board directors, where interlocking represents a director simultaneously sitting on multiple company boards. The dynamic networks motivating the regression models we shall now define, using our graph Laplacian framework, do not have bipartite constraints. The motivating datasets are the Austen and Dickens novels, described in Section 1.3.1, which we have the year each novel was first written for, and the Enron networks, described in Section 1.3.4, which each correspond to a specific month.

We have seen in Section 2.5 for networks with a time structure, PCA can be used to visualise this structure in a lower dimensional space. By performing PCA on the novel dataset, in Example 2.5.1, we hypothesise a linear regression model may be suitable for the novel networks. We also describe a Nadaraya-Watson non-parametric regression model for networks. For the Enron data we will show a limitation of using PCA in visualising the regression and suggest a solution to this. At the end of this chapter we will briefly look at regression for spatial networks.

Throughout this chapter we assume the data are the pairs $\{\boldsymbol{L}_k, \boldsymbol{t}_k\}$, for $1 \leq k \leq n$ in which the $\boldsymbol{L}_k \in \mathcal{L}_m$ are graph Laplacians to be regressed on covariate vectors $\boldsymbol{t}_k =$

$(t_k^1, \ldots, t_k^u)$. For the majority of this chapter we consider a one dimensional covariate ($u = 1$) that we often think of as time however the methods we define will generalise to any covariates.

3.1 Linear regression

Linear regression is a simple model for regression that supposes a linear relationship between the response and the covariates. We saw this may be a reasonable assumption for the novel graph Laplacians changing with time in Example 2.5.1. Using our framework for graph Laplacians we will fit a linear model extrinsically, meaning we fit a linear model in the tangent space (see Section 2.2.4). Using this model we can predict the graph Laplacians for specific covariates by obtaining a prediction in the tangent space and using our framework to transform it back into the graph Laplacian space, $\mathcal{L}_m$.

The linear model for graph Laplacians regression error model differs for the Euclidean power metric, defined in (2.2.2) and the Procrustes power metric, defined in (2.2.3). For the Euclidean power metric the regression error model is

$$\exp_\nu^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_k)) = \mathrm{vech}^*(\boldsymbol{D}_0 + \sum_{w=1}^{u} t_k^w \boldsymbol{D}_w) + \boldsymbol{\epsilon}, \tag{3.1.1}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}_{m(m-1)/2}(\mathbf{0}, \boldsymbol{\Omega}), \tag{3.1.2}$$

where vech^* is defined in (0.0.3). This regression model is in the tangent space and we take $\boldsymbol{\nu} = \mathbf{0}$. In general $\boldsymbol{\Omega}$ has a large number of elements, so in practice it is necessary to restrict $\boldsymbol{\Omega}$ to be diagonal or even isotropic, $\boldsymbol{\Omega} = \omega^2 \boldsymbol{I}_{m(m-1)/2}$. Recall that for the novels $m = 1000$. The estimated parameters $\{\hat{\boldsymbol{D}}_0, \ldots, \hat{\boldsymbol{D}}_u\}$ in (3.1.1) are the least squares solution to

$$(\hat{\boldsymbol{D}}_0, \ldots, \hat{\boldsymbol{D}}_u) = \arg\min_{\boldsymbol{D}_0, \ldots, \boldsymbol{D}_u} \sum_{k=1}^{n} \| \exp_\nu^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_k)) - \mathrm{vech}^*(\boldsymbol{D}_0 + \sum_{w=1}^{u} t_k^w \boldsymbol{D}_w)\|^2, \tag{3.1.3}$$

which are also the maximum likelihood estimates when $\boldsymbol{\Omega}$ is diagonal. The predicted

graph Laplacian for the covariate t_k is given by

$$\boldsymbol{f}(\boldsymbol{t}_k) = \hat{\boldsymbol{L}}_k = \mathrm{P}_{\mathcal{L}}\left(\mathrm{F}_{\alpha}^{-1}\left(\exp_{\nu}\left(\mathrm{vech}^{*}\left(\hat{\boldsymbol{D}}_0 + \sum_{w=1}^{u} t_k^w \hat{\boldsymbol{D}}_w\right)\right)\right)\right) \in \mathcal{L}_m, \quad (3.1.4)$$

and so $\hat{\boldsymbol{L}}_k$ is the fitted graph Laplacian for the covariate $\boldsymbol{t}_k$. The optimisation in (3.1.3) is a convex optimisation problem, defined in (1.2.7), and the parameters of the regression line are found using the standard least squares approach in the tangent space. As the tangent space has dimension $m(m-1)/2$ for the Euclidean power metrics then the optimisation reduces element-wise for $1 \leq i, j \leq m$, to $m(m-1)/2$ independent optimisations.

For the Procrustes power metric the regression error model is

$$\exp_{\nu}^{-1}(\mathrm{F}_{\alpha}(\boldsymbol{L}_k)) = \mathrm{vec}(\boldsymbol{D}_0 + \sum_{w=1}^{u} t_k^w \boldsymbol{D}_w) + \boldsymbol{\epsilon}, \quad (3.1.5)$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}_{(m-1)^2}(\boldsymbol{0}, \boldsymbol{\Omega}), \quad (3.1.6)$$

where vec is defined in (0.0.1). The only difference using the Procrustes power metric has with the Euclidean power metric in the model is the vech* is changed to vec, and we take $\boldsymbol{\nu} = \mathrm{F}_{\alpha}(\hat{\eta})$. These changes are due to the difference in the definition of the tangent space, in (2.2.5), when using the Procrustes power metric. Just like for the Euclidean power metric in (3.1.3) the parameters $\{\hat{\boldsymbol{D}}_0, \ldots, \hat{\boldsymbol{D}}_u\}$ for the Procrustes power metric are found by minimising the least squares error, which is a convex optimisation.

Once a linear regression line is fitted, of interest is to test if there is significant evidence of linear regression with a covariate, meaning the corresponding $\boldsymbol{D}$ value is not a matrix of 0s. To test for the significance of covariate t^w the hypotheses are $H_0 : \boldsymbol{D}_w = \boldsymbol{0}$ and $H_1 : \boldsymbol{D}_w \neq \boldsymbol{0}$. By Wilks' Theorem (Wilks, 1962), if H_0 is true then the likelihood ratio test statistic is

$$T^{\ell} = -2\log\Delta = -2\left(\sup_{\mathcal{D}, \boldsymbol{D}_w = \boldsymbol{0}} \ell(\mathcal{D}) - \sup_{\mathcal{D}, \boldsymbol{D}_w \neq \boldsymbol{0}} \ell(\mathcal{D})\right) \sim \chi^2_{\frac{m(m-1)}{2}}, \quad (3.1.7)$$

approximately when n is large, where $\mathcal{D} = \{\boldsymbol{D}_0, \ldots, \boldsymbol{D}_u, \boldsymbol{\Omega}\}$ and ℓ is the log-likelihood function of $\phi(\exp_{\nu}^{-1}(\mathrm{F}_{\alpha}(\boldsymbol{L}_k))$ under the distribution from (3.1.1), which is a multivariate normal distribution. Using (3.1.7) H_0 is rejected in favour of H_1 at the $100\alpha\%$ significance level if T^{ℓ} is greater than the $(1-\alpha)$ quantile of $\chi^2_{\frac{m(m-1)}{2}}$, in which case

there is evidence for linear regression.

Example 3.1.1: Linear regression applied to the Austen and Dickens novel data For the Austen and Dickens data each novel, represented by a graph Laplacian $\boldsymbol{L}_k$, is paired with the year, t_k, the novel was written. We regress the $\{\boldsymbol{L}_k\}$ on the $\{t_k\}$ using the method in Section 3.1 for each author's novels, using the Euclidean and square root Euclidean metrics with $u = 1$. To visualise the regression lines in Figure 3.1 we find $\hat{\boldsymbol{L}}(t_k)$ for t_k at year intervals for the specific metrics and project these to the PC1 and PC2 space. For each metric the regression lines seem to fit Austen's data well, and could be used to see how her writing style has changed over time. As we noted in Example 2.5.1, for the PCA on the novels, Austen uses the stylistic device "free indirect speech", which corresponded to PC 2, more in later novels which the regression line also reflects. The regression lines fit the Dickens' novels less well, for example the novel *A Tale of Two Cities* is appearing closer to earlier novels on the regression line when it was in fact one of Dickens later novels. Unlike other Dickens novels, *A Tale of Two Cities* is a historical novel and so it may be expected that it does not fit in the temporal sequence of graph Laplacians as the other novels by Dickens.

To test for regression we estimated $\boldsymbol{\Omega}$ by assuming it was diagonal and then performed our test for regression, defined in (3.1.7) on the novels. The p-values were extremely small ($< 10^{-16}$) for both the Austen and Dickens regression lines, for both the Euclidean and square root Euclidean metrics. Hence there is very strong evidence to believe that the writing style of both authors changes with time, regardless of which metric we choose.

3.2 Nadaraya-Watson regression of graph Laplacians versus Euclidean covariate

The linear regression model we have just defined for graph Laplacians is an example of a parametric regression model. Parametric models may not always be appropriate especially if the underlying model for the data is unknown and cannot be sensibly approximated. In these cases non-parametric regression models are preferable. A popular choice is Nadaraya Watson regression described in Section 1.2.4. The Nadaraya Watson regression model can be adapted to work for graph Laplacians and the metrics we have defined for them.

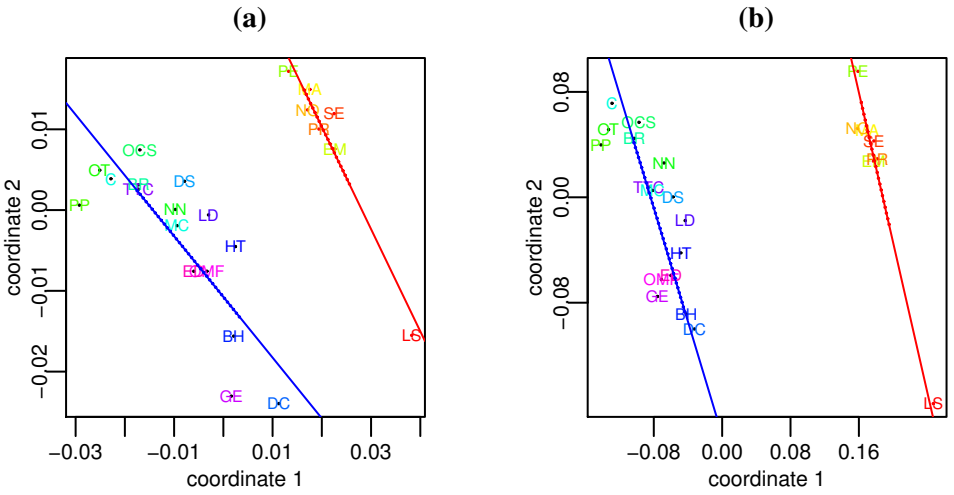


Figure 3.1: *Plot of PC 1 and PC 2 scores for the Austen and Dickens novels, coloured in time order (red to green for Austen novels and green to violet for Dickens novels) with extrinsic regression lines for Dickens novels (blue) and Austen novels (red) using the a) Euclidean and b) square root Euclidean metric. The abbreviations for novels are found in Table 1.2.*

The standard Nadaraya-Watson estimate defined in (1.2.8) for predicting graph Laplacians from given Euclidean covariates, $\boldsymbol{t}$, is given by

$$\hat{\boldsymbol{L}}(\boldsymbol{t}) = \frac{\sum_{i=1}^{n} K_h(\boldsymbol{t} - \boldsymbol{t}_i)\boldsymbol{L}_i}{\sum_{i=1}^{n} K_h(\boldsymbol{t} - \boldsymbol{t}_i)}, \tag{3.2.1}$$

where K_h is a kernel function with bandwidth $h > 0$. A common choice of kernel function is the Gaussian kernel given as

$$K_h(\boldsymbol{u}) = \frac{1}{h\sqrt{2\pi}} \exp(-\frac{\|\boldsymbol{u}\|^2}{2h^2}). \tag{3.2.2}$$

Any kernel function by definition is guaranteed to be non negative and therefore the standard Nadaraya-Watson estimate of a graph Laplacian is always the sum of positively weighted graph Laplacians. From Results 2.1.2 and 2.1.3 we know the space $\mathcal{L}_m$ is a convex cone meaning the sum of positively weighted graph Laplacians, and hence the estimate in (3.2.1) will be a graph Laplacian i.e. $\hat{\boldsymbol{L}}(\boldsymbol{t}) \in \mathcal{L}_m$.

The estimate in (3.2.1) is just the graph Laplacian which minimises the sum of the Euclidean distance, d_1, between the graph Laplacians weighted by K_h, given by

$$\hat{\boldsymbol{L}}(\boldsymbol{t}) = \arg\min_{\boldsymbol{L} \in \mathcal{L}_m} \sum_{i=1}^{n} K_h(\boldsymbol{t} - \boldsymbol{t}_i) d_1(\boldsymbol{L}_i, \boldsymbol{L})^2. \tag{3.2.3}$$

We can generalise this to give a more general Nadaraya-Watson estimate suitable for minimising any distance between graph Laplacians (Davis et al., 2010). This general Nadaraya-Watson estimate is the projected matrix that minimises the given distance, d, between weighted graph Laplacians, given as,

$$\hat{\boldsymbol{L}}(\boldsymbol{t}) = \mathrm{P}_{\mathcal{L}}(\arg \min_{\boldsymbol{L} \in \mathcal{PSD}_m} \sum_{i=1}^{n} K_h(\boldsymbol{t} - \boldsymbol{t}_i) d(\boldsymbol{L}_i, \boldsymbol{L})^2). \tag{3.2.4}$$

This is an extrinsic method hence the projection is needed and the constraint $\boldsymbol{L} \in \mathcal{PSD}_m$ is needed for the distance to be defined. For the Euclidean power metric this becomes

$$\hat{\boldsymbol{L}}(\boldsymbol{t}) = \mathrm{P}_{\mathcal{L}} \left(\mathrm{F}_{\alpha}^{-1} \left(\frac{\sum_{i=1}^{n} \mathrm{F}_{\alpha}(K_h(\boldsymbol{t} - \boldsymbol{t}_i)\boldsymbol{L}_i)}{\sum_{i=1}^{n} K_h(\boldsymbol{t} - \boldsymbol{t}_i)} \right) \right), \tag{3.2.5}$$

note when $\alpha = 1$ this estimate simplifies to that in (3.2.1).

To solve (3.2.4) for the Procrustes metric, the algorithm for weighted generalised Procrustes mean given in Dryden and Mardia (2016, Chapter 7) would be implemented.

Example 3.2.1: Nadaraya-Watson regression of the Austen and Dickens novel data with time

We apply the Nadaraya-Watson model to the Charles Dickens and Jane Austen novels separately to predict their writing styles at different times. We compared using the metrics d_1 and $d_{\frac{1}{2}}$. For each author a Nadaraya-Watson estimate was produced for each year within the period the author was writing. We compared different bandwidths, h, in the Gaussian Kernel. The results are shown in Figure 3.2 plotted on the first and second principal component space for all the novels. Using the bandwidth $h = 2$ seems preferred for both metrics as when $h = 1$ the regression lines are not at all smooth and when $h = 5$ both regression lines are not fitting to the curve of the data at all.

For both metrics with $h = 2$ the regression lines for Dickens appears to show a turning point around the years 1850 and 1851. In the year 1851 Dickens had a tragic year including his wife having a nervous breakdown, his father dying and his youngest child dying. It is possible that the turning point is corresponding to these significant events (Charles Dickens Info, 2018). As there are far fewer novels written by Austen it is far less obvious if there is any turning point in her writing, however it is still clear that Lady Susan still appears to be an anomaly, not fitting with the regression line consistent with Austen's other works.

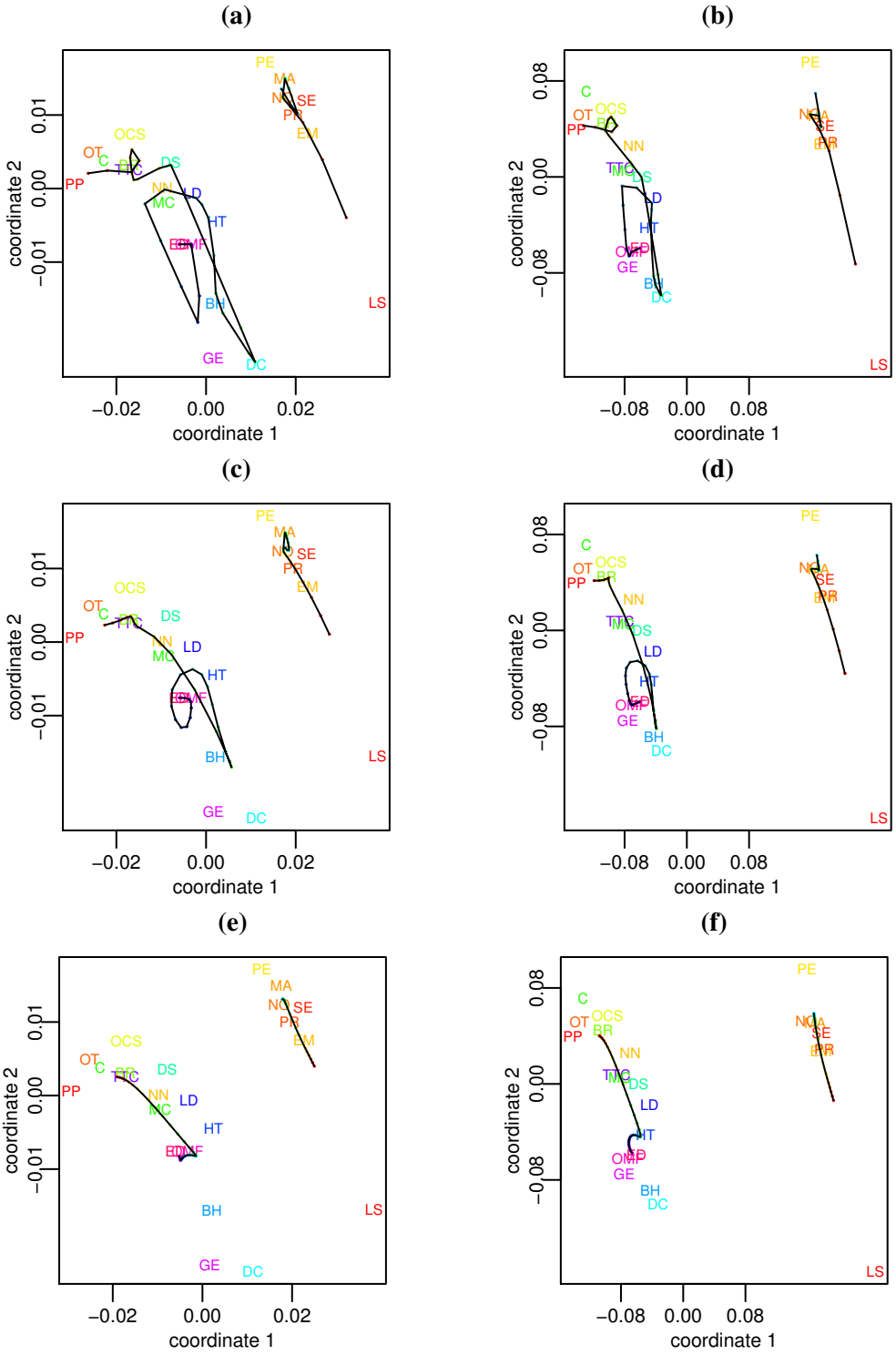


Figure 3.2: *Regression paths for the Dickens novels, coloured in time order green to violet between the years 1836 to 1870, and Austen novels ,coloured in time order red to green between the years 1794 to 1815, using (left to right) $d = d_1$ and $d = d_{\frac{1}{2}}$, with bandwidth (top to bottom) $h = 1, 2, 5$. The abbreviations for novels are found in Table 1.2.*

3.3 Nadaraya-Watson regression of Euclidean response versus graph Laplacian covariate

The Nadaraya Watson can also be applied in a reverse setting where some variable $\boldsymbol{t}_i$ is dependent on the graph Laplacian $\boldsymbol{L}_i$, this can be written as $\boldsymbol{t}_i = \boldsymbol{t}(\boldsymbol{L}_i)$. This could be used if, for example, one had the times networks were produced and then wanted to predict the time a new network was produced. In this case the Nadaraya-Watson estimator is a linear combination of known $\boldsymbol{t}$ values, weighted by the graph Laplacian distances, given by

$$\hat{\boldsymbol{t}}(\boldsymbol{L}) = \frac{\sum_{i=1}^{n} K_h(d(\boldsymbol{L}, \boldsymbol{L}_i))\boldsymbol{t}_i}{\sum_{i=1}^{n} K_h(d(\boldsymbol{L}, \boldsymbol{L}_i))}, \tag{3.3.1}$$

where d can be any metric between two graph Laplacians. Just as before a common kernel to use, and the one we shall choose, is the Gaussian kernel defined in (3.2.2).

Example 3.3.1: Nadaraya-Watson regression of times on the Austen and Dickens novel data

We apply this method to predict the year a novel was written given its graph Laplacian. As there are only 7 Austen novels we only applied this method to the Dickens novels as we did not feel there was sufficient data to get sensible results for the Austen novels. For a specified metric the Nadaraya-Watson estimate for time for a novel was found using all novels except the one of interest, and this was repeated for all 16 of Dickens' novels. We used the Gaussian kernel and for each metric the Nadaraya-Watson method was run repeatedly for bandwidths with intervals of 0.0001 between 0 and 0.1 and then the bandwidth that gave the smallest overall error of the predictions, measured by the root mean square deviation, was chosen. This gave the bandwidths 0.0048 for the Euclidean metric, 0.0524 for the square root Euclidean and 0.0523 for the Procrustes metric. These bandwidths correspond to the root mean square deviations 8.151, 7.262 and 7.260 years for the Euclidean, square root Euclidean and Procrustes size-and-shape metric respectively.

The predicted time for every Dickens novel for both the Euclidean and the square root Euclidean metrics are found in Figure 3.3. The plot for the predicted time when using the Procrustes size-and-shape metric is not included as this is visibly identical to the plot produced for the square root Euclidean metric. The linear regression line between the predicted and true times is included in the plots in Figure 3.3, for an optimal pre-

diction the line would be $y = x$ i.e. have gradient 1 and pass through the origin, as the prediction would equal the true year. For all metrics the rough ordering of the novels is maintained and the linear regression lines seem close to $y = x$.

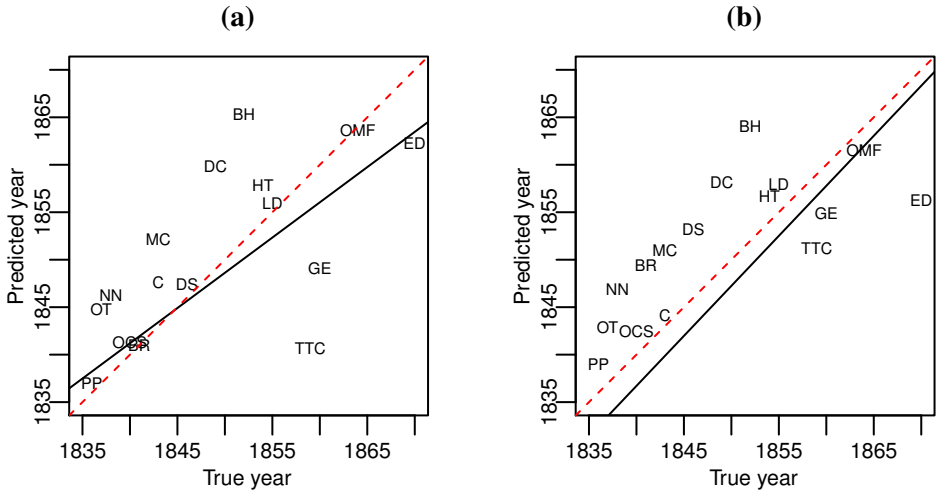


Figure 3.3: *The true and predicted times corresponding when each Dickens novel was written when using the Nadaraya-Watson model, using a) the Euclidean metric and b) the square root Euclidean metric. The linear regression line between the predicted and true times is plotted in black and the line $y = x$ is plotted in red. The abbreviations for novels are found in Table 1.2.*

3.4 Horseshoe effect

We defined in Section 2.5 how principal component analysis can be applied for graph Laplacians to reduce dimensions and produce 2D plots. We will see now how this method may produce a common but unwanted and often ignored phenomenon called the horseshoe effect when the data has a time structure. To investigate this effect we consider the the Enron networks and plot in Figure 3.4 the PC plots using the Euclidean and square root Euclidean metric for this data. The plot for the Procrustes size-and-shape metric is not included as it is visibly identical to the plot for the square root Euclidean metric. For all plots a horseshoe or arc shape can be seen suggesting there is a change point in the Enron data at the 'tip' of the arc, this is in fact an example of the horseshoe effect, and to conclude there is a change point in the data may be misleading (Kendall, 1970).

The horseshoe effect is present in many datasets including in political roll call votes (Diaconis et al., 2008), archaeology seriation data (Kendall, 1971) and microbiome data

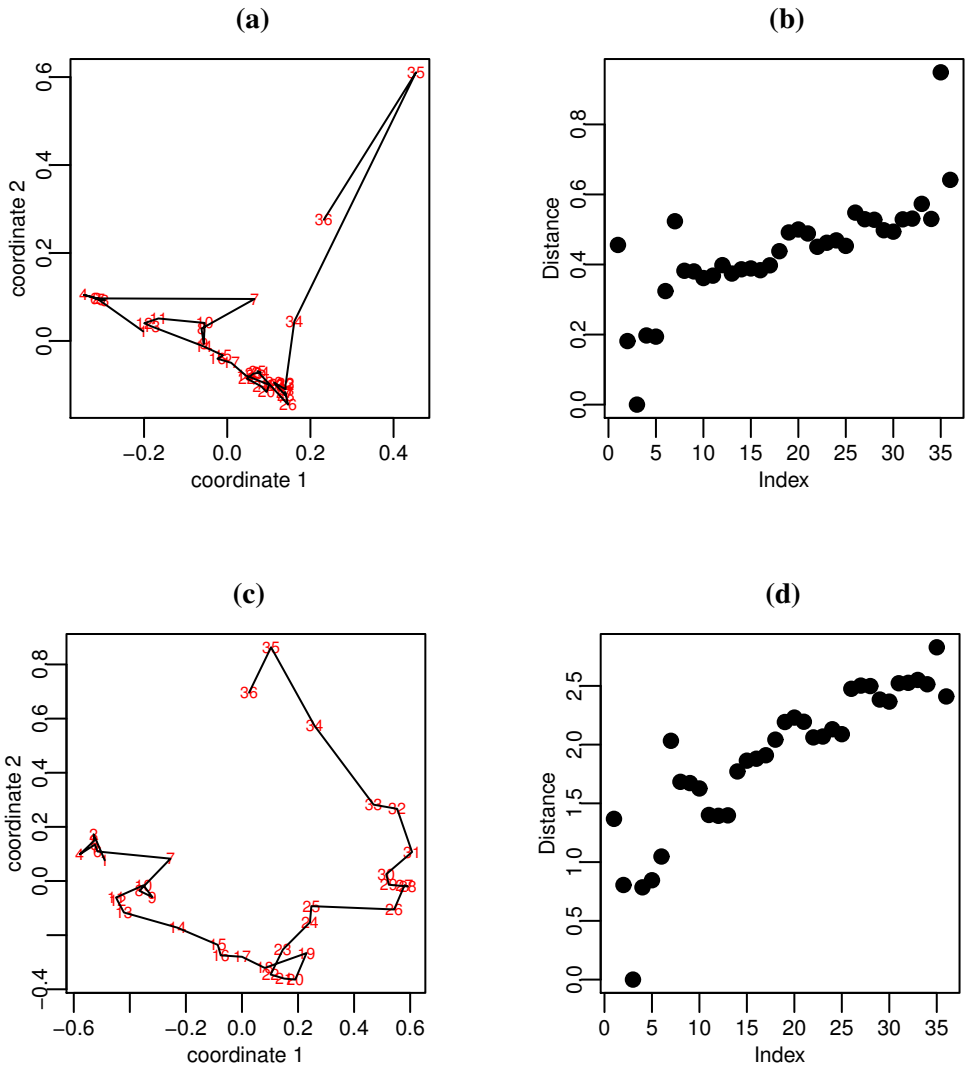


Figure 3.4: *PC plot for Enron network using (a) Euclidean metric and (c) Square root Euclidean metric. The red digits indicate the month of the data. Plots for distance of the 3rd network with each other network for the (b) Euclidean and (d) square root Euclidean metric.*

(Morton et al., 2017). Explained in Mardia et al. (1979, page 412) the horseshoe effect occurs when the distances which are 'large', between data points, appear the same as those that are 'moderate'. To investigate this effect we can contrive example datasets, that we can think of as in the graph Laplacian tangent space, that illustrate the horseshoe effect.

Based on the example in Morton et al. (2017), an example showing the horseshoe effect clearly, is the dataset $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$, where $\boldsymbol{v}_k = \exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_k))$, with $n = 100$ and $m = n + 2 = 102$ where

$$(\boldsymbol{v}_k)_j = \begin{cases} 1 \text{ if } j = k, k+1, k+2, \\ 0 \text{ otherwise.} \end{cases} \tag{3.4.1}$$

There clearly is a time structure to this data which can be visualised in (3.4.2). For each increment in time the data is shifted by 1 row (Morton et al., 2017). The PCA plot for this example is (a) in Figure 3.5 which clearly shows the horseshoe effect.

$$\begin{aligned} \boldsymbol{v}_1^T &= (1, 1, 1, 0, 0, 0, \dots) \\ \boldsymbol{v}_2^T &= (0, 1, 1, 1, 0, 0, \dots) \\ \boldsymbol{v}_3^T &= (0, 0, 1, 1, 1, 0, \dots) \end{aligned} \tag{3.4.2}$$

Another example is for an autoregressive model, where $\boldsymbol{v}_1 = \mathbf{0}$ and for $2 \leq k \leq n$

$$\begin{aligned} \boldsymbol{v}_k &= c + \rho \boldsymbol{v}_{k-1} + \epsilon_k \\ \text{where } \epsilon_k &\sim \mathcal{N}_m\left(\mathbf{0}, \sigma^2 \boldsymbol{I}\right), \end{aligned} \tag{3.4.3}$$

where $\boldsymbol{v}_k = \exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_k))$. Again there is clearly a strong time structure to this data. The PCA plot for this is show in (b) of Figure 3.5 for $c = 0$, $\rho = 0.99$ and $\sigma = 1$ and again this clearly is showing the horseshoe effect.

For both the examples we have presented, the PCA plots show horseshoe shapes, even though the data has no change point in. The reason for this horseshoe can be explained most clearly by considering the distance between $\boldsymbol{v}_1$ and $\boldsymbol{v}_k$ for both examples shown in (c) and (d) of Figure 3.5. For (c) as k increases the distance increases rapidly until it stabilise and therefore the distance between $\boldsymbol{v}_1$ and $\boldsymbol{v}_k$ become almost identical for all $k > K$. This same effect is less obvious in (d) but we can still see the gradient in (d) is negative showing the rate the distance is increasing is decreasing. We see the same is

true for the Enron data in Figure 3.4 as the distances rapidly increase until they begin to stabilise for both metrics. For the Enron data we look at the distance with the third network as this still shows the same effect but happens to be less noisy then using the first network. The horseshoe effect is explained by these distance plots as the distance metric between say time 1 and a 'large' time is around the same as between time 1 and a 'medium' time. Morton et al. (2017) described this as a "saturation property" of the metric, and so on the PCA plot the point corresponding to a 'large' time is pulled in closer to time 1 than we intuitively would expect.

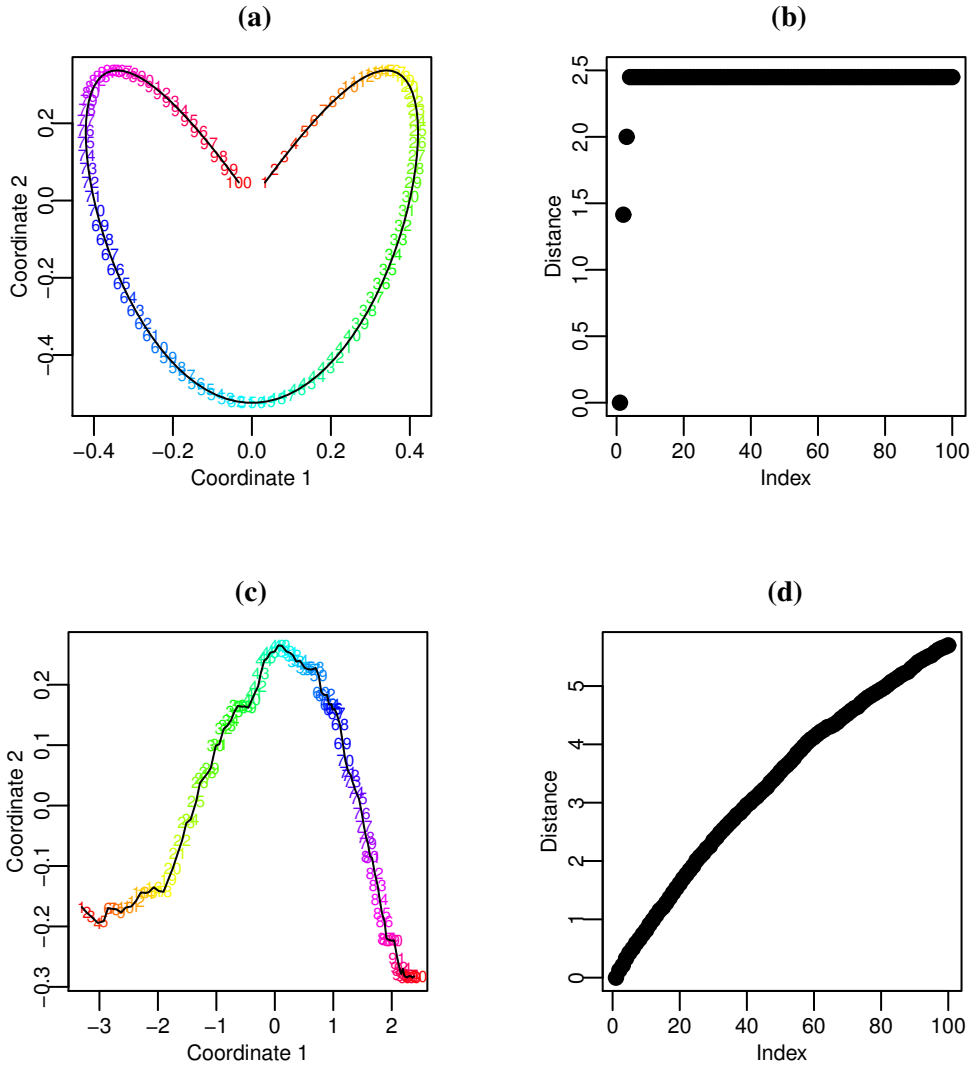


Figure 3.5: *Euclidean PCA plots and plots for the Euclidean distance between $\boldsymbol{v}_1$ and $\boldsymbol{v}_k$, for $2 \leq k \leq n$, for ((a) & (b)) Model 3.4.1 and ((c) & (d)) Model 3.4.3.*

When considering data such as the Enron data the horseshoe effect is potentially misleading as it can lead to the conclusion there is a change point in the data when in fact

there is not one. This motivates wanting to create a method for visualising the data in a low dimensional space in such a way that avoids producing the misleading horseshoe effect. When using the Euclidean metric PC and MDS plots are identical (Williams, 2002, Section 2.2), and hence if a Euclidean PC plot exhibits the horseshoe effect so will the Euclidean MDS plot. For the other metrics PCA and MDS do not give identical plots, although generally they do give plots that look similar. Because for the Euclidean metric PC and MDS plots are identical instead of altering our PCA method we shall use MDS to remove the horseshoe effect by defining a new distance metric between graph Laplacians. This metric is chosen to be unsaturating by using prior knowledge of the ordering of the data. The metric we choose is an adaptation of the Mahalanobis metric in the embedding space (Mahalanobis, 1936). The Mahalanobis distance only provides an adaptation for the Euclidean power metrics, defined in (2.2.2), and does not provide an adaptation to the Procrustes power metric, defined in (2.2.3). The adapted Mahalanobis distance between two graph Laplacians $\boldsymbol{L}_k$ and $\boldsymbol{L}_l$, at times k and l respectively, is

$$\sqrt{(\exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_k)) - \exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_l)) - \boldsymbol{\mu})^T \boldsymbol{\Sigma}_{kl}^{-1} (\exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_k)) - \exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_l)) - \boldsymbol{\mu})},$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_{kl}$ are the mean and covariance matrix of $\exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_k)) - \exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_l))$ respectively. These values are not feasible to estimate from the data and therefore we must assume a model for the $\boldsymbol{L}_k$s. For the Enron data, and many time structure datasets, a sensible model is an autoregressive model, given by

$$\begin{aligned} \exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_k)) &= \boldsymbol{c} + \rho \exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_{k-1})) + \boldsymbol{\epsilon}_k, \\ &\text{where } \boldsymbol{\epsilon}_k \sim \mathcal{N}_m\left(\mathbf{0}, \sigma^2 \boldsymbol{I}_{\frac{m(m-1)}{2}}\right). \end{aligned} \tag{3.4.4}$$

This model has weak stationarity and hence the means must satisfy

$$\mathrm{E}[\exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_k))] = \mathrm{E}[\exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_{k+1}))].$$

Using (3.4.4) we get

$$\mathrm{E}[\exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_k))] = \frac{\boldsymbol{c}}{1-\rho}.$$

To simplify this we assume $\boldsymbol{c} = \mathbf{0}$ hence $\mathrm{E}[\exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_k))] = \mathbf{0}$ and $\boldsymbol{\mu} = \mathrm{E}[\exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_k))] - \mathrm{E}[\exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_{k-1}))] = \mathbf{0}$. The autocovariance for this au-

toregressive model is known to be

$$\boldsymbol{\Sigma}_{kl} = \frac{\sigma^2 \rho^{|k-l|}}{1-\rho} \boldsymbol{I}_{\frac{m(m-1)}{2}},$$

this is diagonal matrix where the diagonal elements are the variance of elements and we have assumed a 0 covariance between any other elements.

To estimate the value of ρ we firstly rewrite (3.4.4) by

$$\boldsymbol{y}_k = \rho \boldsymbol{v}_k + \boldsymbol{\epsilon}_k,$$

where $\boldsymbol{y}_k = \exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_k))$ and $\boldsymbol{v}_k = \exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_{k-1}))$. We can then estimate ρ by

$$\rho = \arg_{\rho^*} \min \sum_{k=2}^{n} (\boldsymbol{y}_k - \rho^* \boldsymbol{v}_k)^T (\boldsymbol{y}_k - \rho^* \boldsymbol{v}_k) \tag{3.4.5}$$

$$= \arg_{\rho^*} \min f(\rho^*). \tag{3.4.6}$$

To solve this equation we differentiate with respect to ρ^* giving

$$\frac{df}{d\rho^*} = \sum_{k=2}^{n} (-2\boldsymbol{y}_k^T \boldsymbol{v}_k + 2\rho \boldsymbol{v}_i^T \boldsymbol{v}_k),$$

this equation is set to 0 to find the value of ρ^* which minimises (3.4.5)

$$\rho = \frac{\sum_{k=2}^{n} (\boldsymbol{y}_k^T \boldsymbol{v}_k)}{\sum_{k=2}^{n} (\boldsymbol{v}_k^T \boldsymbol{v}_k)}.$$

The Mahalanobis metric between graph Laplacians, $\boldsymbol{L}_k$ and $\boldsymbol{L}_l$, can now be written as

$$\begin{aligned}
&= \sqrt{\frac{1-\rho}{\sigma^2 \rho^{|k-l|}} (\exp_0^{-1}((\mathrm{F}_\alpha(\boldsymbol{L}_k)) - \exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_l)))^T (\exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_k)) - \exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_l)))} \\
&= \frac{1}{\sigma} \sqrt{\frac{1-\rho}{\rho^{|k-l|}}} \| \exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_k)) - \exp_0^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_l)) \|. \\
&= \frac{1}{\sigma} \sqrt{\frac{1-\rho}{\rho^{|k-l|}}} \| \mathrm{F}_\alpha(\boldsymbol{L}_k) - \mathrm{F}_\alpha(\boldsymbol{L}_l) \|,
\end{aligned}$$

using Result 2.2.4. As $\frac{1}{\sigma}$ is just a positive constant this is just consistently scaling the distance, and has no effect except scale in the MDS plots, we remove it to prevent us from having to estimate σ values.

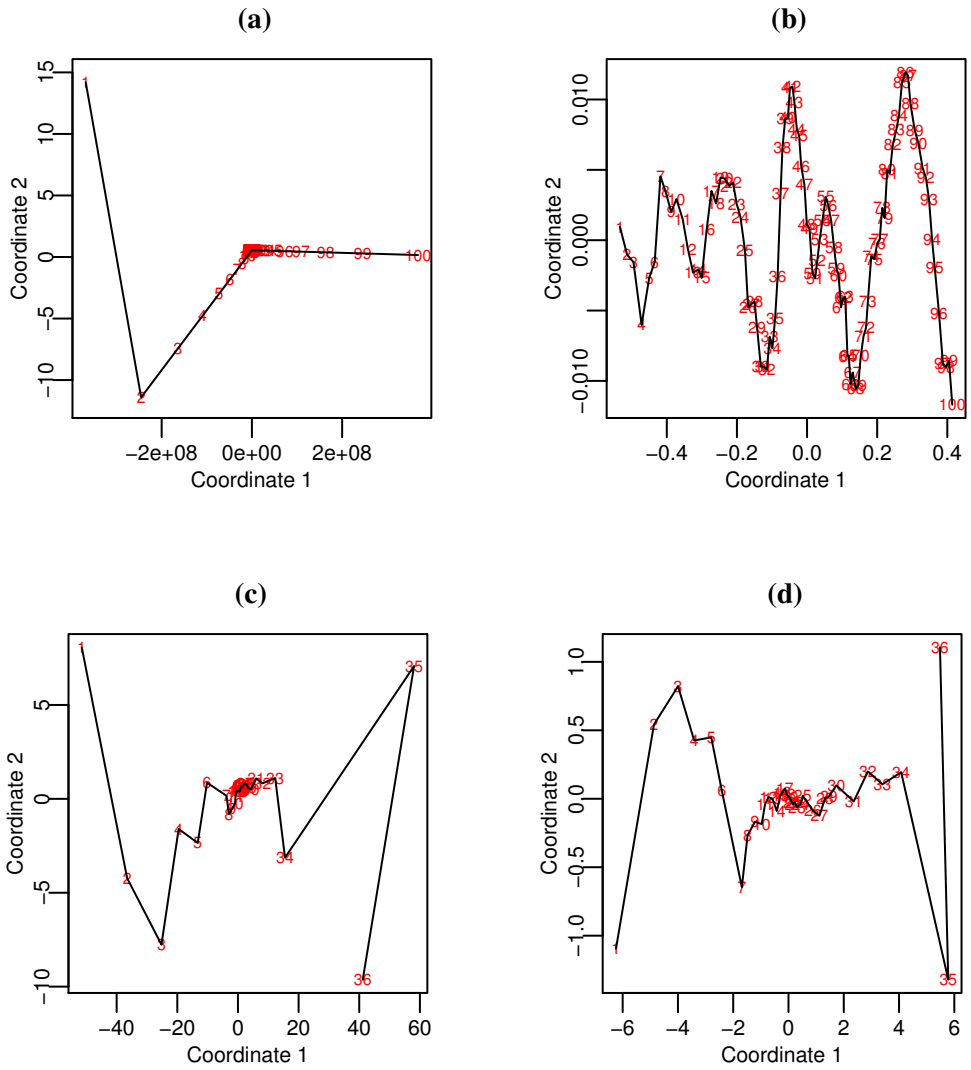


Figure 3.6: *MDS plots using the Mahalanobis metric for (a) Model 3.4.1 with* $\alpha = 1$*, (b) Model 3.4.3 with* $\alpha = 1$*, (c) the Enron data with* $\alpha = 1$ *and (d) the Enron data with* $\alpha = \frac{1}{2}$

Figure 3.6 shows the MDS plots when using the Mahalanobis distance for the simulated data and the Enron data. For (b) where the data is simulated from an autoregressive model and so we know our assumptions are valid the MDS plot looks sensible and has removed the horseshoe effect. In all the other cases where the assumption of an autoregressive model may be violated we see the middle values are clumping together and these plots are not sensible. For (a) the first coordinates are extremely large and this is as the ρ estimate is around 0.66 so the distances between data points further away becomes extremely large. The second coordinates are much smaller and the shape of the points is unexpected and suggests the the second coordinate is not acting sensibly. For the Enron data we look at the adaptation of the Euclidean and square root Euclidean metrics by using the Mahalanobis distance with $\alpha = 1$ and $\frac{1}{2}$. When using the Mahalanobis distance for the Enron data, found in plots (c) and (d) of Figure 3.6, the clumping of the middle data points is very extreme.

It is possible that when the autoregressive model assumptions are violated the estimates of ρ are not sensible. We therefore look at the effect of choosing a ρ value that maximises the variance explained by the first coordinate for each example, shown in Figure 3.7. The plot for the autoregressive model in 3.4.3 has remained the same due to the fact assumptions were not violated for it. For the simulated data of model 3.6 the MDS plot still contains unexplained turning points, this is most likely due to the fact the model has completely violated the assumptions required to use the Mahalanobis distance. For the Enron data the plots with ρ maximising variance seem more sensible and the 3rd, 7th, 34th, 35th and 36th month seem to stand out as true change points. Of these the 7th and 35th also stand out in the original plots in Figure 3.4. The 7th month corresponds to December 1999, this is picked out to be an anomaly in Wang et al. (2014), believed to coincide with Enron's tentative sham energy deal with Merrill Lynch created to meet profit expectations and boost the stock price. Month 34 and 35 correspond to March and April 2002 these correspond to the former Enron auditor, Arthur Andersen, being indicted for obstruction of justice (The Guardian, 2006). For the Enron data the data points have been coloured by two clusters found by hierachial clustering of the MDS coordinates. The change in clusters is between July and September 2000 for both metrics, and this seems to correspond to Enron shares hitting an all-time high, so could be sensible that a change would take place after this (The Guardian, 2006).

To see if the results when using the Mahalanobis distance for the Enron data are sensible we plot the consecutive distance between monthly networks of the Enron data from

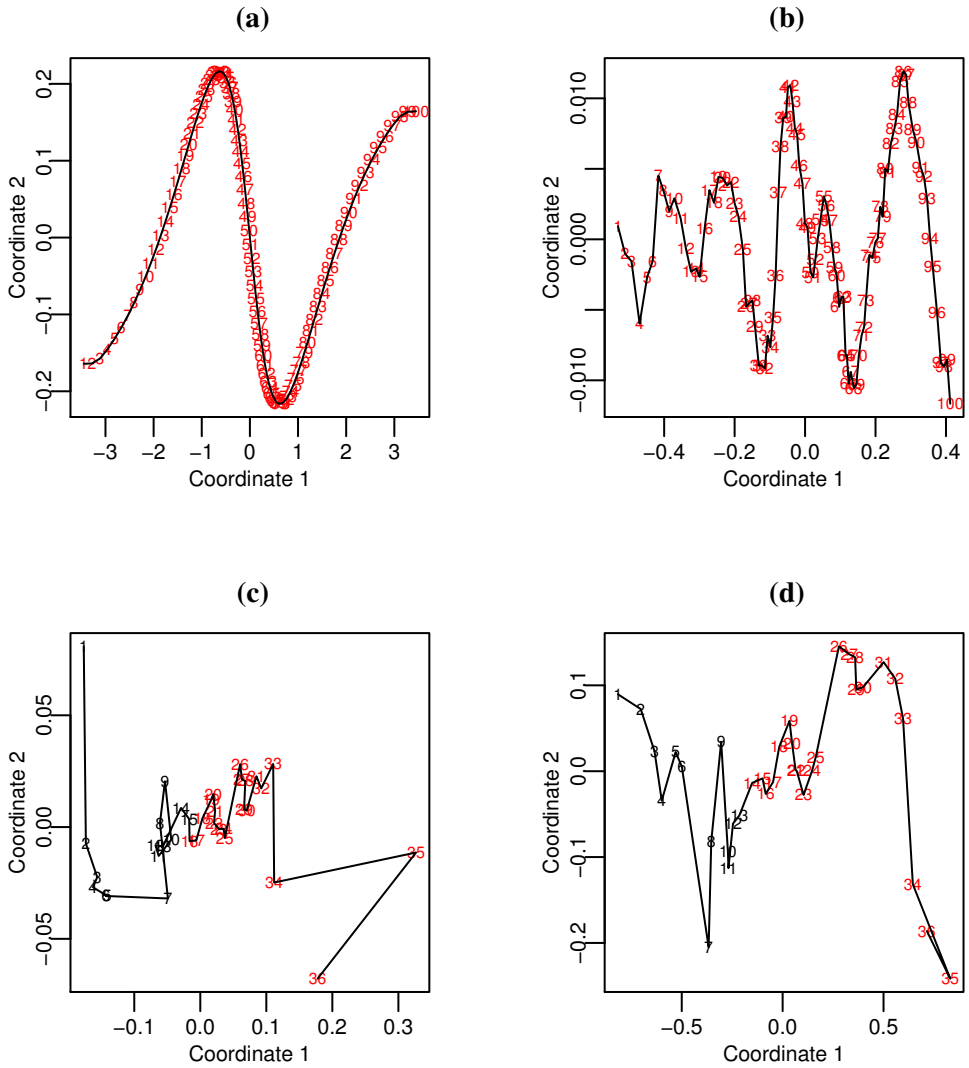


Figure 3.7: *MDS plots using the Mahalanobis metric with ρ chosen to maximise the variance explained by PC 1, for (a) Model 3.4.1 with $\alpha = 1$, (b) Model 3.4.3 with $\alpha = 1$, (c) the Enron data with $\alpha = 1$ and (d) the Enron data with $\alpha = \frac{1}{2}$.*

the months June 1999 to May 2002 for the Euclidean and square root Euclidean metric in Figure 3.8. The consecutive distances give us an idea of which months may be anomalies or turning points as these are likely to be ones with a large distance from the month previous. This method of detecting anomalies and turning points in networks has been used in Koutra et al. (2013), although with a different metric between networks. From our plots it looks like month 7, 34 and 35 may be anomalies. These all correspond to anomalies we picked out before, when using the Mahalanobis distance, suggesting using the Mahalanobis distance is sensible. Hence we have provided a sensible method for producing low-dimensional visualisations of data with a time structure that avoids the horseshoe effect. We have however had to impose quite strong modelling assumptions and so this method may not always be as suitable to use as it is for the Enron data.

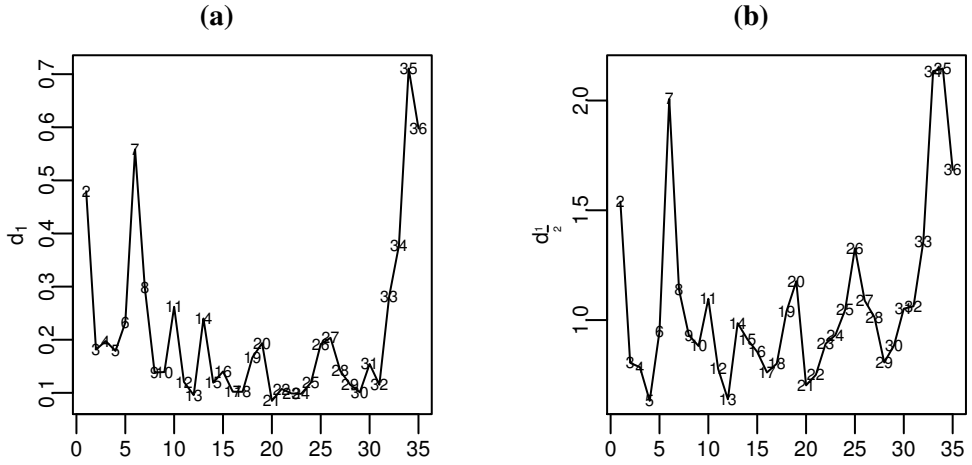


Figure 3.8: *Consective distances between the Enron networks for each month for the a) Euclidean metric and b) square root Euclidean metric.*

3.5 Kriging

To conclude this chapter we finally consider the case where graph Laplacians are dependent on spatial coordinates. We can denote this as $\boldsymbol{L}_i = \boldsymbol{L}(\boldsymbol{x}_i)$, where $\boldsymbol{x}_i \in \mathbb{R}^K$ are coordinates. We will adapt the commonly used spatial method, Kriging, described in Section 1.2.4, so we can estimate graph Laplacians for known coordinates. Whilst we could use our method in Section 3.2 to tackle this problem we believe it is better to use Kriging which makes modelling assumptions based on the spatial structure. We will apply Kriging on our tangent space, and so we denote the tangent space vector

$\boldsymbol{v}_i = \exp_\nu^{-1}(\mathrm{F}_\alpha(\boldsymbol{L}_i))$ as $\boldsymbol{v}_i = \boldsymbol{v}(\boldsymbol{x}_i)$. Applying Kriging in a tangent space to a manifold is seen in Pigoli et al. (2016), this focused on the manifold of positive definite symmetric matrices whereas we shall focus on the manifold of graph Laplacians.

As our graph Laplacians have a spatial structure we will assume the graph Laplacians in the tangent space are from a stationary random field. Data on a stationary random field have a constant mean over the field and the covariance between data points is only dependant on the distance between data points. These seem like sensible assumptions for our tangent space. Formally these assumptions are

$$\begin{aligned} &\mathrm{E}(\boldsymbol{v}(\boldsymbol{x})) = \mathrm{E}(\boldsymbol{v}(\boldsymbol{y})) = \boldsymbol{\mu}, && \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{C} \\ &\mathrm{Cov}_{v_j}(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{Cov}_{v_j}(|\boldsymbol{x} - \boldsymbol{y}|, 0), && \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{C} \text{ and } 1 \leq j \leq \frac{m(m-1)}{2} \end{aligned}$$

where $\mathrm{Cov}_{v_j}(\boldsymbol{x}, \boldsymbol{y})$ represents the covariance between $(\boldsymbol{v}(\boldsymbol{x}))_j$ and $(\boldsymbol{v}(\boldsymbol{y}))_j$. For simplicity we will also assume $\mathrm{Cov}_{\boldsymbol{v}_j}(\boldsymbol{x}, \boldsymbol{y})$ are not dependent on j and that each element in $\boldsymbol{v}(\boldsymbol{x})$ is independent. Again these assumptions seem sensible and can be formally written as

$$\begin{aligned} \mathrm{Cov}_{v_j}(\boldsymbol{x}, \boldsymbol{y}) &= \mathrm{Cov}_{v}(\boldsymbol{x}, \boldsymbol{y}) \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^K \\ \mathrm{Cov}(\boldsymbol{v}_j(\boldsymbol{x}), \boldsymbol{v}_w(\boldsymbol{x})) &= 0 \quad \forall \boldsymbol{x} \in \mathbb{R}^K \text{ when } j \neq w. \end{aligned}$$

We also assume $\mathrm{Cov}_{\boldsymbol{v}_j}(\boldsymbol{x}, \boldsymbol{y})$ is known, in practice this would not generally be true and would need to be estimated. In general the expectation of the graph Laplacians in the tangent space, $\boldsymbol{\mu}$, will not be known and therefore we implement ordinary Kriging, the method designed for when $\boldsymbol{\mu}$ is unknown.

From (1.2.9) the estimate for $\boldsymbol{L}_0 = \boldsymbol{L}(\boldsymbol{x}_0)$ when using Kriging is of the form $\hat{\boldsymbol{L}}_0 = \mathrm{P}_{\mathcal{L}}(\mathrm{F}_\alpha^{-1}(\exp_\nu(\boldsymbol{v}_0)))$ with $\boldsymbol{v}_0 = \sum_{i=1}^n W_i \boldsymbol{v}_i$, for a sample of graph Laplacians with known coordinates $\{\boldsymbol{L}_1, ..., \boldsymbol{L}_n\}$ and tangent vector $\{\boldsymbol{v}_1, ..., \boldsymbol{v}_n\}$. We choose the estimator to be unbiased and due to the stationarity of the field under the model we have

$$\begin{aligned} \boldsymbol{\mu} = \mathrm{E}(\hat{\boldsymbol{v}}_0) = \mathrm{E}(\sum_{i=1}^n W_i \boldsymbol{v}_i) = \sum_{i=1}^n W_i \boldsymbol{\mu} \\ \sum_{i=1}^n W_i = 1. \end{aligned}$$

The estimator is also chosen to have minimum variance, so to find the W_i values we

minimise $\text{Var}(\sum_{i=1}^{n} W_i \boldsymbol{v}_i - \hat{\boldsymbol{v}}_0)$. This gives the estimator as

$$\hat{\boldsymbol{v}}_0 = \arg\min_{\boldsymbol{v}_0} \text{Var}(\sum_{i=1}^{n} W_i \boldsymbol{v}_i - \boldsymbol{v}_0)$$
$$\text{subject to} \sum_{i=1}^{n} W_i = 1.$$

Ordinary Kriging uses Lagrange multipliers to solve the minimisation, hence the solution can be written

$$\begin{bmatrix} w_1 \\ \vdots \\ w_n \\ \lambda \end{bmatrix} = \begin{bmatrix} \text{Cov}_v(\boldsymbol{x}_1, \boldsymbol{x}_1) & \dots & \text{Cov}_v(\boldsymbol{x}_1, \boldsymbol{x}_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \text{Cov}_v(\boldsymbol{x}_n, \boldsymbol{x}_1) & \dots & \text{Cov}_v(\boldsymbol{x}_n, \boldsymbol{x}_n) & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} \text{Cov}_v(\boldsymbol{x}_1, \boldsymbol{x}_0) \\ \vdots \\ \text{Cov}_v(\boldsymbol{x}_n, \boldsymbol{x}_0) \\ 1 \end{bmatrix},$$

where λ is a Lagrange multiplier. The weights are now known so the estimate is $\hat{\boldsymbol{L}}_0 = \text{P}_{\mathcal{L}}(\text{F}_{\alpha}^{-1}(\exp_{\nu}(\boldsymbol{v}_0)))$ where $\hat{\boldsymbol{v}}_0 = \sum_{i=1}^{n} W_i \boldsymbol{v}_i$.

Example 3.5.1: Kriging applied to simulated graph Laplacian data

We demonstrate the use of Kriging on graph Laplacians by a simulation study. We consider graph Laplacians dependent on 2D coordinates this could be representative of networks specific to places with latitude and longitude coordinates. We only consider $\alpha = 1$ for this example and we chose graph Laplacians with dimension $m = 5$. Each element in the tangent vector is modelled by a Gaussian process with $\text{E}(\boldsymbol{v}) = \boldsymbol{\mu} = \mathbf{1}$ for $1 \leq j < \frac{m(m-1)}{2}$ and $\text{Cov}_v(\boldsymbol{x}, \boldsymbol{y}) = s - s(1 - \exp(\frac{-|\boldsymbol{x}-\boldsymbol{y}|)}{r})$, with sill $s = 0.025$ and range $r = 50$. Graph Laplacians were generated for each grid point in a 50 by 50 grid, leading to 2500 graph Laplacians. Figure 3.9 provides an illustration of the networks in a small section of the field, we can see how along the field the networks seem to vary smoothly.

To test our method for Kriging of networks we split the set of graph Laplacians randomly into a training set of n =1875 graph Laplacians and coordinates and a test set of 625 graph Laplacians and coordinates, as explained in Section 1.2.4. For each coordinate in the test set Kriging was performed to give a predicted graph Laplacian. This was run 10 times and the mean squared prediction error was 0.003, which is very small, showing that when the assumptions are met Kriging provides a good predictor for the response. The Kriging method in future should be tested on real data when the assump-

tions may be violated to see how the method performs then.

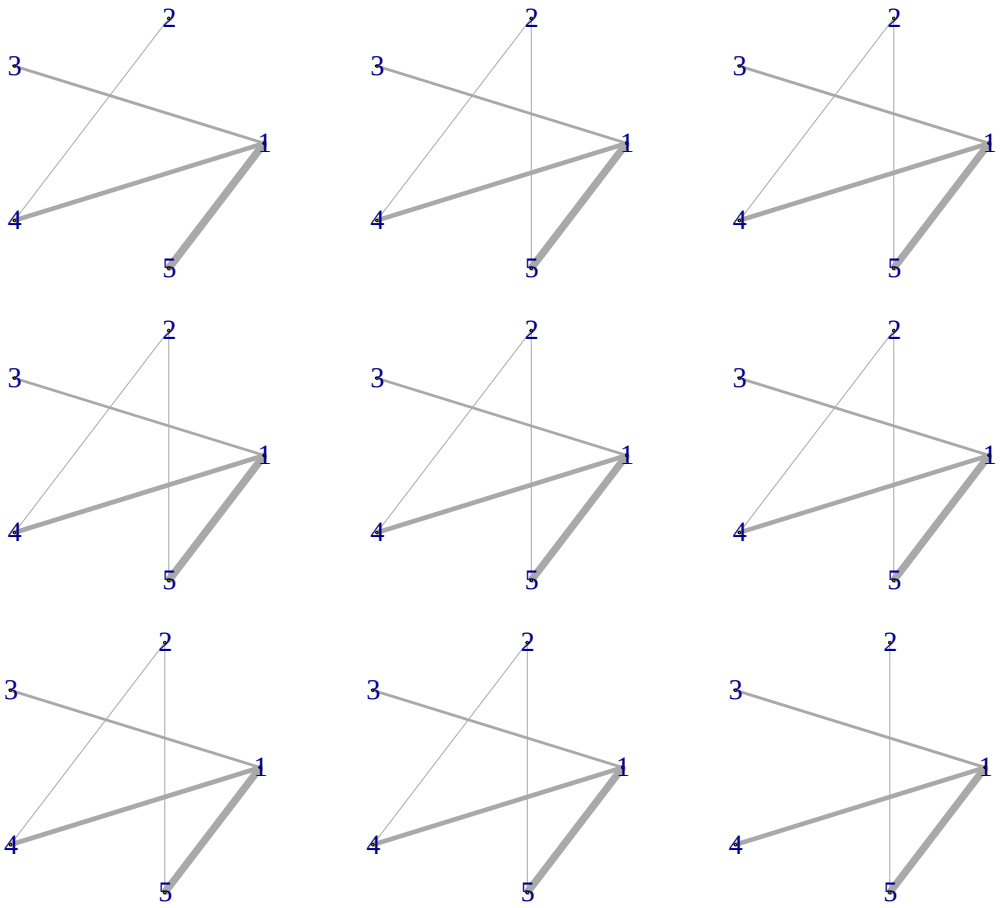


Figure 3.9: *Example networks along Kriging field in Example 3.5.1, corresponding to the field's x coordinates (left to right) 1, 3, 5 and y coordinates (top to bottom) 1, 3 and 5.*

3.6 Summary

In this chapter we have proposed several regression models for networks using the general framework defined in Chapter 2. We have used both parametric and non-parametric models to predict the graph Laplacians from covariates, namely time and spatial coordinates. The non-parametric model was the Nadaraya-Watson regression model whilst the parametric model was the linear regression model, more complex parametric models such as quadratic regression could be easily adapted from our framework too.

Not only did we use Nadaraya-Watson regression to estimate graph Laplacians from known Euclidean covariates, we used it also to estimate Euclidean responses from

known graph Laplacians. A very similar method to this can be used to estimate the probability a known graph Laplacian belongs to a certain class which we shall use in Chapter 5. Chapter 5 looks at the further regression problem of classification, where the aim is to predict a discrete outcome, not a continuous one as we have done in this chapter.

We studied the horseshoe effect on graph Laplacians that occurs on PC plots when the data has a time structure by considering the Enron dataset. We proposed a method to remove the horseshoe effect, however this method required quite strong assumptions for the data. Whilst this method gave promising results for the Enron data, as the method was motivated for Enron data the method could be too specific to this data and so it would be of interest to see if this method seems appropriate to other datasets and if the assumptions seem valid for more datasets.

For graph Laplacians with a spatial structure we adapted the classical method of Kriging to predict graph Laplacians for known spatial coordinates. We currently have only demonstrated this on a simulation study where all assumptions are met and so our estimator did well. It is of interest for future work to use this model on real data with a spatial structure to see if the assumptions we have made hold for real data and how the method performs in more contexts.

CHAPTER 4

Two-sample hypothesis tests for graph Laplacian data

In this chapter we define a formal two-sample test to test for a difference in population mean network given two samples of networks. This topic is seen already to be an interesting challenge in Tang et al. (2017), which proposes a hypothesis test for a certain model of network, named random dot product networks, using adjacency matrices. We use our graph Laplacian framework to define our two-sample test. Ginestet et al. (2017) also uses graph Laplacians with a central limit theorem to develop a hypothesis test for networks, which we shall compare with our two-sample test.

To define our two-sample test we consider two populations $\mathcal{A}$ and $\mathcal{B}$ of $m \times m$ graph Laplacians with corresponding population means $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$ and unprojected means $\boldsymbol{\eta}_A$ and $\boldsymbol{\eta}_B$ defined in (2.3.1). Given two random samples $\{\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_{n_A}\}$ and $\{\boldsymbol{B}_1, \boldsymbol{B}_2, \ldots, \boldsymbol{B}_{n_B}\}$ respectively from $\mathcal{A}$ and $\mathcal{B}$, the goal is to test the hypotheses

$$H_0 : \boldsymbol{\mu}_A = \boldsymbol{\mu}_B \text{ and } H_1 : \boldsymbol{\mu}_A \neq \boldsymbol{\mu}_B. \tag{4.0.1}$$

A suitable test statistic for this test is $T = d(\mathrm{P}_{\mathcal{L}}(\hat{\boldsymbol{A}}), \mathrm{P}_{\mathcal{L}}(\hat{\boldsymbol{B}}))^2$ for a suitable metric d, where $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ are defined as $\hat{\boldsymbol{\eta}}$ in (2.3.2), hence $\mathrm{P}_{\mathcal{L}}(\hat{\boldsymbol{A}})$ and $\mathrm{P}_{\mathcal{L}}(\hat{\boldsymbol{B}})$ are the sample population means for populations $\mathcal{A}$ and $\mathcal{B}$, respectively. However for this test statistic, as the projection is included, no central limit theorem is immediately available for the test statistics and so the test must be developed non-parametrically, for example using a permutation test. A permutation test requiring the projected means is computationally not appealing as the projection would need to be performed many times and so the test would become quite slow for larger graph Laplacians. Although this test is not entirely

prohibitive, especially for smaller dimensional graph Laplacians, we instead choose to test

$$H_0 : \boldsymbol{\eta}_A = \boldsymbol{\eta}_B \text{ and } H_1 : \boldsymbol{\eta}_A \neq \boldsymbol{\eta}_B, \tag{4.0.2}$$

with a far less computationally intensive test statistic:

$$T = d(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})^2, \tag{4.0.3}$$

which is, in general, a squared extrinsic distance. Note for all these cases $\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}} \in \mathcal{PSD}_m$ so theses distances are well-defined, due to the fact stated in Section 2.2.2 that our distances in $\mathcal{L}_m$ hold more generally for $\mathcal{PSD}_m$. Whilst testing the equality of the $\boldsymbol{\eta}$ values in (4.0.2) is not equivalent to testing equality of the $\boldsymbol{\mu}$s in (4.0.1) we shall show it is a sensible approximation of this test. The tests are not equivalent as $\boldsymbol{\eta}_A \neq \boldsymbol{\eta}_B$ does not imply $\boldsymbol{\mu}_A \neq \boldsymbol{\mu}_B$ due to the projection being many to one. It is possible for $\boldsymbol{\mu}_A = \mathrm{P}_{\mathcal{L}}(\boldsymbol{\eta}_A) = \mathrm{P}_{\mathcal{L}}(\boldsymbol{\eta}_B) = \boldsymbol{\mu}_B$ but $\boldsymbol{\eta}_A \neq \boldsymbol{\eta}_B$.

To confirm our test in (4.0.2), using the test statistic in (4.0.3), is a suitable approximation of the test in (4.0.1) we compare $d^2_\alpha(\boldsymbol{\mu}_A, \boldsymbol{\mu}_B)$ with $d^2_\alpha(\boldsymbol{\eta}_A, \boldsymbol{\eta}_B)$ for some simple examples. In Figure 4.1 two random samples, $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ with $k = 1, \ldots, 100$, were generated from Erdös-Renyi networks described in Section 1.2.5 and $d^2_\alpha(\hat{\boldsymbol{\mu}}_A, \hat{\boldsymbol{\mu}}_B)$ and $d^2_\alpha(\hat{\boldsymbol{\eta}}_A, \hat{\boldsymbol{\eta}}_B)$ were found. This was repeated 1000 times for 16 values of $p_B \in [0.025, 0.2]$, the probability of an edge in the second sample. The probability of an edge in the first sample was set as $p_A = 0.1$. Figure 4.1 shows how $d^2_{\frac{1}{2}}(\hat{\boldsymbol{\mu}}_A, \hat{\boldsymbol{\mu}}_B)$ and $d^2_{\frac{1}{2}}(\hat{\boldsymbol{\eta}}_A, \hat{\boldsymbol{\eta}}_B)$ are always very similar for all the simulations. Importantly when $d^2_{\frac{1}{2}}(\hat{\boldsymbol{\eta}}_A, \hat{\boldsymbol{\eta}}_B) \neq 0$, so $\hat{\boldsymbol{\eta}}_A \neq \hat{\boldsymbol{\eta}}_B$, it is clear $d^2_{\frac{1}{2}}(\hat{\boldsymbol{\mu}}_A, \hat{\boldsymbol{\mu}}_B) \neq 0$ meaning $\hat{\boldsymbol{\mu}}_A \neq \hat{\boldsymbol{\mu}}_B$. An equivalent result is seen when using $d^2_{\frac{1}{2},S}$. Therefore it is sensible to assume in practice if $\boldsymbol{\eta}_A \neq \boldsymbol{\eta}_B$ implies $\boldsymbol{\mu}_A \neq \boldsymbol{\mu}_B$ and so test we shall use in (4.0.2) is usually equivalent to the desired test in (4.0.1). We do not compare the distances for d_1, as when using d_1 the projection is not required as the $\boldsymbol{\eta}$ values are intrinsic means so are already guaranteed to belong to $\mathcal{L}_m$. In this case the test we use in (4.0.2) is identical to the desired test in (4.0.1) and $d_1(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})^2 = d_1(\mathrm{P}_{\mathcal{L}}(\hat{\boldsymbol{A}}), \mathrm{P}_{\mathcal{L}}(\hat{\boldsymbol{B}}))^2$.

Any Euclidean or Procrustes power metric is suitable to use in the test statistic in (4.0.3). However we will just consider the Euclidean,

$$T_E = d_1(\hat{\boldsymbol{A}}_E, \hat{\boldsymbol{B}}_E)^2, \tag{4.0.4}$$

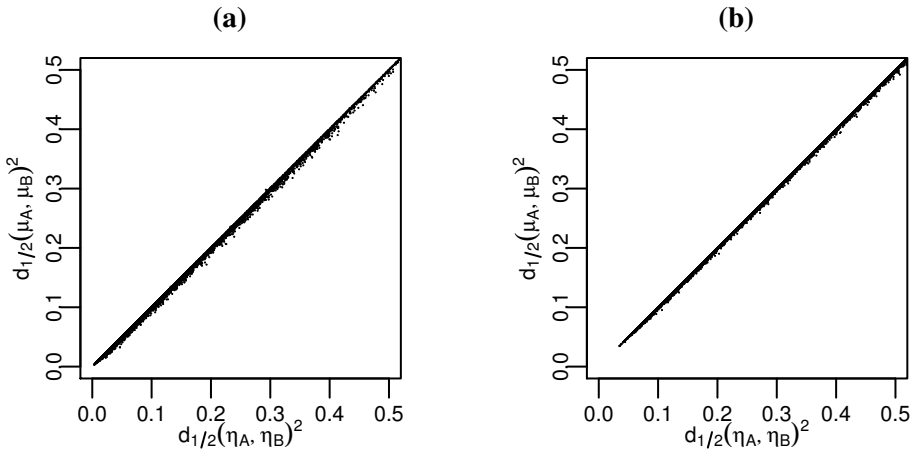


Figure 4.1: *Comaprison of $d^2_{\frac{1}{2}}(\boldsymbol{\mu}_A, \boldsymbol{\mu}_B)$ with $d^2_{\frac{1}{2}}(\boldsymbol{\eta}_A, \boldsymbol{\eta}_B)$ for graph Laplacians generated from Erdös-Renyi networks with a) m=5 and b) m=10.*

the square root Euclidean,

$$T_H = d_{\frac{1}{2}}(\hat{\boldsymbol{A}}_H, \hat{\boldsymbol{B}}_H)^2, \tag{4.0.5}$$

and the Procrustes size-and-shape,

$$T_S = d_{\frac{1}{2},S}(\hat{\boldsymbol{A}}_S, \hat{\boldsymbol{B}}_S)^2, \tag{4.0.6}$$

where the subscripts $\{E, H, S\}$ refer to whether the Euclidean, square root or Procrustes size-and-shape means have been used, respectively. For these test statistics we will derive a general central limit theorem that will lead to an asymptotic distribution for the test statistic. We will also provide a method for a non-parametric test when assumptions about the data's distribution cannot be made and so the distribution of the test statistic is unknown.

The likelihood ratio test for regression with test statistic $-2\log\Delta$ in Section 3.1 gives an alternative test for equality of means when the covariates are group labels. However if we were to use this test from Section 3.1 an additional assumption of normality for the observations needs to be made, and therefore the two-sample test we have defined in the current chapter is preferred.

An alternative two-sample test is that proposed in Ginestet et al. (2017) which we shall define and compare with the test statistics we have defined, on a variety of different datasets.

4.1 Ginestet two-sample test

The two-sample test statistic from Ginestet et al. (2017) is

$$T_G = \frac{n_A n_B}{n_A + n_B}(\phi(\hat{\boldsymbol{A}}_E) - \phi(\hat{\boldsymbol{B}}_E))^T \hat{\boldsymbol{\Sigma}}'^{-1}(\phi(\hat{\boldsymbol{A}}_E) - \phi(\hat{\boldsymbol{B}}_E)) \xrightarrow{D} \chi^2_{\binom{m}{2}},$$
$$\text{with } \hat{\boldsymbol{\Sigma}}' = \frac{n_A \hat{\boldsymbol{\Sigma}}'_A + n_B \hat{\boldsymbol{\Sigma}}'_B}{n_A + n_B - 2},$$

where $\hat{\boldsymbol{\Sigma}}'_A$ and $\hat{\boldsymbol{\Sigma}}'_B$ are the estimated covariance matrices for $\phi(\boldsymbol{A})$ and $\phi(\boldsymbol{B})$ respectively using a shrinkage estimator from Schäfer and Strimmer (2005) and ϕ is defined in (0.0.4). This distribution holds under the assumption $\text{E}(\boldsymbol{\mu}^E_{Aij}) \neq 0$ and $\text{E}(\boldsymbol{\mu}^E_{Bij}) \neq 0$ for $i \neq j$ and also assumes $\boldsymbol{\Sigma}'_A = \boldsymbol{\Sigma}'_B = \boldsymbol{\Sigma}'$, which may not always be true.

The test in Ginestet et al. (2017) is defined by ignoring the diagonal element of the graph Laplacian and so it equates to just using the adjacency matrix, defined in (1.2.2). The test statistic when including the diagonal is

$$T'_G = \frac{n_A n_B}{n_A + n_B}(\text{vech}(\hat{\boldsymbol{A}}_E) - \text{vech}(\hat{\boldsymbol{B}}_E))^T \hat{\boldsymbol{\Sigma}}''^{-}(\text{vech}(\hat{\boldsymbol{A}}_E) - \text{vech}(\hat{\boldsymbol{B}}_E)),$$

where $\boldsymbol{\Sigma}''_A = \text{Cov}(\text{vech}(\boldsymbol{A}))$, $\boldsymbol{\Sigma}''_B = \text{Cov}(\text{vech}(\boldsymbol{B}))$, $\hat{\boldsymbol{\Sigma}}'' = \frac{n_A \hat{\boldsymbol{\Sigma}}''_A + n_B \hat{\boldsymbol{\Sigma}}''_B}{n_A + n_B - 2}$ and vech is defined in (0.0.2). As $\hat{\boldsymbol{\Sigma}}''$ will not in general be full rank, $\hat{\boldsymbol{\Sigma}}''^{-}$ represents the Moore-Penrose inverse (Penrose, 1955). Two interesting results hold for T_G and T'_G.

Result 4.1.1. *T'_G has an identical asymptotic distribution to T_G.*

Result 4.1.2. *$T_G = T'_G$ when $n_A - 1, n_B - 1 \geq \frac{m(m-1)}{2}$ and the standard unbiased estimator for covariance matrices is used instead of the shrinkage estimator from Schäfer and Strimmer (2005).*

These results show that for simplicity it is fine to ignore the diagonal as done in T_G. The proofs of these results are found in Section 4.8.1.

4.2 A central limit theorem

Similarly to Ginestet et al. (2017) a central limit theorem will be used to find the asymptotic distribution of the test statistic of our hypothesis test when using the Euclidean power metric.

Result 4.2.1. *Consider independent identically distributed random observations $\boldsymbol{A}_k$*

where $F_\alpha(\boldsymbol{A}_k)$, $k = 1, \ldots, n$, has a distribution with mean $\mathbb{E}[F_\alpha(\boldsymbol{A})]$, where F_α is defined in (2.2.1). Then for any Euclidean power metric

$$\sqrt{n}\left(\phi\left(F_\alpha(\hat{\boldsymbol{\eta}})\right) - \phi\left(F_\alpha(\boldsymbol{\eta})\right)\right) \xrightarrow{D} \mathcal{N}_{\frac{m(m-1)}{2}}\left(\mathbf{0}, \boldsymbol{\Sigma}\right),$$

as $n \to \infty$, where $\phi(\boldsymbol{A})$ is defined in (0.0.4) and $\boldsymbol{\Sigma}$ is a finite variance matrix.

This central limit theorem holds, under the condition $\text{var}(F_\alpha(\boldsymbol{A}))_{ij})$ is finite, for all $1 \leq i, j \leq m$, as the embedded extrinsic mean when using the power Euclidean metric, $\text{F}_\alpha(\hat{\boldsymbol{\eta}})$, is just the arithmetic mean in the embedding space.

When $\alpha = 1$ this result is similar to that in Ginestet et al. (2017) although they work directly in $\mathcal{L}_m$ whereas we work in the embedding space. When considering the Procrustes power metric similar central limit theorem results follow involving a rotation term, providing the more stringent conditions of Bhattacharya and Patrangenaru (2005, Section 3) hold.

Consider two independent random samples $\boldsymbol{A}_k, k = 1, \ldots, n_A$, and $\boldsymbol{B}_k, k = 1, \ldots, n_B$, where $F_\alpha(\boldsymbol{A}_k)$ and $F_\alpha(\boldsymbol{B}_k)$ have distributions with mean $\mathbb{E}[F_\alpha(\boldsymbol{A})]$ and $\mathbb{E}[F_\alpha(\boldsymbol{B})]$ respectively. When finding the distribution for the test statistic of our hypothesis test for these samples we rely on the central limit theorem in Result 4.2.1 that is only valid for the Euclidean power metric. Hence we only consider the distribution for the test statistic in (4.0.3) when using the Euclidean power metric, $T = d_\alpha(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})$. Using our central limit theorem in Result 4.2.1 we have when $n_A, n_B \longrightarrow \infty$,

$$\begin{aligned} n_A^{\frac{1}{2}}(\phi(\text{F}_\alpha(\hat{\boldsymbol{A}})) - \phi(\text{F}_\alpha(\boldsymbol{\eta}_A)))\dot{\sim}\mathcal{N}_{\frac{m(m-1)}{2}}(\mathbf{0}, \boldsymbol{\Sigma}_A), \\ n_B^{\frac{1}{2}}(\phi(\text{F}_\alpha(\hat{\boldsymbol{B}})) - \phi(\text{F}_\alpha(\boldsymbol{\eta}_B)))\dot{\sim}\mathcal{N}_{\frac{m(m-1)}{2}}(\mathbf{0}, \boldsymbol{\Sigma}_B), \end{aligned}$$

where $\boldsymbol{\Sigma}_A = \text{Cov}(\phi(\text{F}_\alpha(\boldsymbol{A})))$ and $\boldsymbol{\Sigma}_B = \text{Cov}(\phi(\text{F}_\alpha(\boldsymbol{B})))$. These distributions are independent of one another. We can then write

$$\begin{aligned} \boldsymbol{x} = \left(\frac{n_A n_B}{n_A + n_B}\right)^{\frac{1}{2}} & (\phi(\text{F}_\alpha(\hat{\boldsymbol{A}})) - \phi(\text{F}_\alpha(\hat{\boldsymbol{B}}))) \\ & \dot{\sim}\mathcal{N}_{\frac{m(m-1)}{2}}((\frac{n_A n_B}{n_A + n_B})^{\frac{1}{2}}(\phi(\text{F}_\alpha(\boldsymbol{\eta}_A)) - \phi(\text{F}_\alpha(\boldsymbol{\eta}_B)), \boldsymbol{\Sigma}), \end{aligned} \tag{4.2.1}$$

where $\boldsymbol{\Sigma} = \frac{n_B\boldsymbol{\Sigma}_A + n_A\boldsymbol{\Sigma}_B}{n_A + n_B}$ as $n_A, n_B \longrightarrow \infty$ and $\frac{n_A}{n_B} \to r \in (0, \infty)$. To find the distribution of T under the null hypothesis we first note the Euclidean power distance squared can be expressed as the quadratic form of normal random variables,

Lemma 4.2.1. *When $n_A, n_B \longrightarrow \infty$ and $\frac{n_A}{n_B} \to r \in (0, \infty)$,*

$$T = d_\alpha(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})^2 = \frac{n_A + n_B}{n_A n_B} \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x}. \tag{4.2.2}$$

where $\boldsymbol{x}$, defined in (4.2.1), is normally distributed.

Proof. If we write $\mathrm{F}_\alpha(\hat{\boldsymbol{A}}) = (\hat{a}_{ij})$ and $\mathrm{F}_\alpha(\hat{\boldsymbol{B}}) = (\hat{b}_{ij})$, then the Euclidean power distance squared between the sample means $\mathrm{F}_\alpha(\hat{\boldsymbol{A}})$ and $\mathrm{F}_\alpha(\hat{\boldsymbol{B}})$ can be written in terms of these elements as,

$$\begin{aligned} d_\alpha(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})^2 &= \|(\mathrm{F}_\alpha(\hat{\boldsymbol{A}}) - \mathrm{F}_\alpha(\hat{\boldsymbol{B}})\|^2 \\ &= \sum\sum\nolimits_{i \neq j} (\hat{a}_{ij} - \hat{b}_{ij})^2 + \sum_{i=1}^{m} (\hat{a}_{ii} - \hat{b}_{ii})^2. \end{aligned} \tag{4.2.3}$$

The summands satisfy $\mathrm{F}_\alpha(\hat{\boldsymbol{A}})\mathbf{1}_m = \mathbf{0}_m$ and $\mathrm{F}_\alpha(\hat{\boldsymbol{A}}) = \mathrm{F}_\alpha(\hat{\boldsymbol{A}})^T$, and similarly for $\hat{\boldsymbol{B}}$ too, hence we can write $\hat{a}_{ii} = -\sum_{i \neq j} \hat{a}_{ij}$ and $\hat{b}_{ii} = -\sum_{i \neq j} \hat{b}_{ij}$, and $\hat{a}_{ij} = \hat{a}_{ji}$ and $\hat{b}_{ij} = \hat{b}_{ji}$. We substitute these in to remove the $\hat{a}_{ij}$ and $\hat{b}_{ij}$ with $i \geq j$ and rewrite Equation (4.2.3) as

$$\begin{aligned} d_1(\hat{\boldsymbol{A}}_E, \hat{\boldsymbol{B}}_E)^2 &= \sum\sum\nolimits_{i \neq j} (\hat{a}_{ij} - \hat{b}_{ij})^2 + \sum_{i=1}^{m} (\sum_{j \neq i} (-\hat{a}_{ij} + \hat{b}_{ij}))^2, \\ &= 2\sum\sum\nolimits_{i \neq j} (\hat{a}_{ij} - \hat{b}_{ij})^2 + \sum_{i=1}^{m} \sum_{p \neq i, p \neq j, j \neq i} (\hat{b}_{ij} - \hat{a}_{ij})(\hat{b}_{ip} - \hat{a}_{ip}), \\ &= 4\sum\sum\nolimits_{i < j} (\hat{a}_{ij} - \hat{b}_{ij})^2 \\ &+ \sum_{i<j, k<p, (i,j) \neq (k,p)} \sum_{i=k, i=p, j=k \text{ or } j=p} (\hat{b}_{ij} - \hat{a}_{ij})(\hat{b}_{kp} - \hat{a}_{kp}), \end{aligned} \tag{4.2.4}$$

for which all the a and b terms are independent.

For simplification of notation from now on we work with the difference matrix, $\boldsymbol{D}$, of $\mathrm{F}_\alpha(\hat{\boldsymbol{A}})$ and $\mathrm{F}_\alpha(\hat{\boldsymbol{B}})$,

$$\begin{aligned} \boldsymbol{D} &= \mathrm{F}_\alpha(\hat{\boldsymbol{A}}) - \mathrm{F}_\alpha(\hat{\boldsymbol{B}}) = (\delta_{ij}) \\ \delta_{ij} &= \hat{a}_{ij} - \hat{b}_{ij}. \end{aligned}$$

Hence the distance in (4.2.4) can be written in terms of the elements from this difference

matrix as,

$$d_\alpha(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})^2 = 4\sum\sum_{i<j}(\delta_{ij})^2 + \sum_{i<j,k<p,(i,j)\neq(k,p)} \sum_{i=k,i=p,j=k \text{ or } j=p} (\delta_{ij})(\delta_{kp}). \quad (4.2.5)$$

The $\frac{m(m-1)}{2}$ column vector $\boldsymbol{x}$ from (4.2.1) can be written as,

$$\boldsymbol{x} = \left(\frac{n_A n_B}{n_A + n_B}\right)^{\frac{1}{2}} \phi(\boldsymbol{D}),$$

in terms of which the squared distance can be expressed as the quadratic form,

$$d_\alpha(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})^2 = \frac{n_A + n_B}{n_A n_B} \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x},$$

where the $\boldsymbol{Q}$ matrix, illustrated below, has each row and column corresponding to a δ_{ij} value. The value of q_{rs} corresponding to row δ_{ij} and column δ_{kp} is the coefficient of $\delta_{ij}\delta_{kp}$ in (4.2.5). Written out in full,

$$\boldsymbol{Q} = \quad (4.2.6)$$

	δ_{12}	δ_{13}	$\dots$	δ_{1m}	δ_{23}	$\dots$	δ_{2m}	$\dots$	$\delta_{(m-2)m}$	$\delta_{(m-1)m}$
δ_{12}	4	1	1...1	1	1	1...1	1	0...0	0	0
δ_{13}	1	4	1...1	1	1	0...0	0	$\dots$	0	0
$\dots$	$\begin{matrix}1\\\vdots\\1\end{matrix}$	$\begin{matrix}1\\\vdots\\1\end{matrix}$	$4 \ddots 4$	$\begin{matrix}1\\\vdots\\1\end{matrix}$	$\begin{matrix}0\\\vdots\\0\end{matrix}$	$\ddots$	$\begin{matrix}0\\\vdots\\0\end{matrix}$	$\ddots$	$\vdots$	$\vdots$
δ_{1m}	1	1	1...1	4	0	$\dots$	1	$\dots$	1	1
δ_{23}	1	1	0...0	0	4	1...1	1	$\dots$	0	0
$\dots$	$\begin{matrix}1\\\vdots\\1\end{matrix}$	$\begin{matrix}0\\\vdots\\0\end{matrix}$	$\ddots$	$\vdots$	$\begin{matrix}1\\\vdots\\1\end{matrix}$	$4 \ddots 4$	$\begin{matrix}1\\\vdots\\1\end{matrix}$	$\ddots$	$\vdots$	$\vdots$
δ_{2m}	1	0	0...0	1	1	1...1	4	$\dots$	1	1
$\dots$	$\begin{matrix}0\\\vdots\\0\end{matrix}$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\ddots$	$\vdots$	$4 \ddots 4$	$\vdots$	$\vdots$
$\delta_{(m-2)m}$	0	0	$\dots$	1	0	$\dots$	1	$\dots$	4	1
$\delta_{(m-1)m}$	0	0	$\dots$	1	0	$\dots$	1	$\dots$	1	4

The values of q_{rs} can be determined element-wise by first finding the i, j, k, p values

where the q_{rs} value equals the coefficient of $\delta_{ij}\delta_{kp}$; these are,

$$\begin{aligned} i &= f(r,m), \\ j &= g(r,i,m), \\ k &= f(s,m), \\ p &= g(s,k,m), \end{aligned} \tag{4.2.7}$$

$$\text{where } f(r,m) = m - 1 - \text{floor}\left(\frac{\sqrt{(-8(r-1)+4m(m-1)-7}}{2} - \frac{1}{2}\right),$$

$$g(r,i,m) = r + i - \frac{m(m-1)}{2} + \frac{(m-i+1)(m-i)}{2}.$$

The function f and g map an index of a vector to indices of the upper diagonal of a matrix, with dimension $m \times m$, running through row by row. The function f gives the row index and the function g gives the column index. Using Equation (4.2.5) the matrix in (4.2.8) can hence be summarised by ,

$$q_{rs} = \begin{cases} 4, & \text{if } r = s \\ 1, & \text{if } i = k \text{ or } i = p \text{ or } j = k \text{ or } j = p \\ 0, & \text{otherwise.} \end{cases} \tag{4.2.8}$$

□

As we have now shown that the test statistic, when using the Euclidean power metric, can be written as the quadratic form of normal random variables we can now find the distribution of this test statistic under the null hypothesis. When H_0 is true $\phi(\mathrm{F}_\alpha(\boldsymbol{\eta}_A)) = \phi(\mathrm{F}_\alpha(\boldsymbol{\eta}_B))$ and so from Equation (4.2.1) we can see $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Hence the distribution of the test statistic is just the distribution of a quadratic form of normal random variables with mean $\mathbf{0}$, and so the distribution of the tests statistic is as follows:

Result 4.2.2. *Consider independent random samples of networks of size n_A and n_B. For the power Euclidean metrics under the null hypothesis, H_0: $\boldsymbol{\eta}_A = \boldsymbol{\eta}_B$, as $n_A, n_B \longrightarrow \infty$, such that $n_A/n_B \to r \in (0, \infty)$:*

$$\frac{n_A n_B}{n_A + n_B} T = \frac{n_A n_B}{n_A + n_B} d_\alpha(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})^2 \xrightarrow{D} \sum_{i=1}^{m(m-1)/2} \lambda_i \chi_1^2, \tag{4.2.9}$$

in which each χ_1^2 *is independent and* λ_i *are the* $m(m-1)/2$ *non-zero eigenvalues of* $\boldsymbol{\Sigma Q}$.

Proof. The result follows directly from Box (1954) which provides the distribution of a quadratic form of normal random variables with mean $\mathbf{0}$. □

When the value of $\boldsymbol{\Sigma}$ is known, or can be sensibly approximated, this distribution can be used to find the critical value, $T_{100a\%}$, such that the null hypothesis is rejected when $T > T_{100a\%}$, where $T_{100a\%}$ is chosen to give a significance level of $100a\%$.

The quantiles of the distribution in Result 4.2.2 and hence the critical value can be found easily through large simulations. To find quantiles without simulation the distribution can be approximated by a singular chi squared distribution or a Gaussian distribution (Box, 1954). The distribution in Result 4.2.2 can be approximated, using results from Box (1954), by

$$\begin{aligned} T &\dot{\sim} g\chi_h^2 \\ \text{where } g = \frac{n_A + n_B}{n_A n_B} \frac{\sum_{i=1}^{l} \lambda_i^2}{\sum_{i=1}^{l} \lambda_i} &\text{ and } h = \frac{(\sum_{i=1}^{l} \lambda_i)^2}{\sum_{i=1}^{l} \lambda_i^2}. \end{aligned} \tag{4.2.10}$$

As this chi-squared distribution is the sum of independent random variables with finite mean and variance then by the central limit theorem this distribution can then be approximated by

$$T \dot{\sim} \frac{n_A + n_B}{n_A n_B} \mathcal{N}\left(\sum_{i=1}^{l} \lambda_i, 2 \sum_{i=1}^{l} \lambda_i^2 \right). \tag{4.2.11}$$

In practice $\boldsymbol{\Sigma}$ will generally not be known and so needs to be estimated. In our application using the novel dataset with $m = 1000$, $\boldsymbol{\Sigma}$ is a symmetric matrix with $M(M+1)/2$ parameters where $M = m(m-1)/2 = 499500$. The $\boldsymbol{\Sigma}$ matrix hence is often very highly dimensional, which can lead to issues estimating it, especially from relatively small samples. It is seen in Preston and Wood (2011) that for the smaller samples we will deal with, using regularised versions of tests statistics will often perform better. One approach is to use the shrinkage estimator from Schäfer and Strimmer (2005) to estimate $\boldsymbol{\Sigma}$, as employed by Ginestet et al. (2017), but this is still impractical for our application with $m = 1000$. If we assume a diagonal matrix $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^*$ then the λ_i correspond to the variances of individual components of the difference in means, and

these can be estimated consistently from method of moments estimators. A further very simple model that enables us to write the distribution more explicitly for $\alpha = 1$ is an isotropic covariance matrix with covariance matrix $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_{m(m-1)/2}$, which only requires estimation of a single variance parameter σ^2, we consider this model in Section 4.2.1. We shall also consider, in Section 4.2.2, how we can calculate $\boldsymbol{\Sigma}$ when using the Euclidean power metric with $\alpha = 1$ for two specific models, the stochastic block model and Erdös-Renyi model, and how this enables us to write the distribution of the test statistic more precisely. Alternatively we will also consider non-parametric methods that do not require large covariance matrices to be estimated or models for the data to be prescribed.

4.2.1 A parametric test assuming isotropic covariance matrix

If we assume isotropic covariance matrices for both sets $\mathcal{A}$ and $\mathcal{B}$ we can write the distribution of the test statistic when the Euclidean power metric is used. In this case for $\boldsymbol{x}$ as defined in (4.2.1) we have $\boldsymbol{\Sigma}_A = \sigma_A^2 \boldsymbol{I}_{\frac{m(m-1)}{2}}$ and $\boldsymbol{\Sigma}_B = \sigma_B^2 \boldsymbol{I}_{\frac{m(m-1)}{2}}$ and we also must assume that under H_0 $\sigma_A^2 = \sigma_B^2 = \sigma_0^2$, meaning under the null hypothesis the populations have equal covariance matrices. Then under H_0

$$\boldsymbol{x} \dot{\sim} \mathcal{N}\left(\boldsymbol{0}, \sigma_0^2 \boldsymbol{I}_{\frac{m(m-1)}{2}}\right). \tag{4.2.12}$$

As the covariance matrix is a scaled identity matrix, the eigenvalues of $\boldsymbol{\Sigma Q}$ in Result 4.2.2 are the eigenvalues of $\sigma_0^2 \boldsymbol{Q}$ which are known if the eigenvalues of $\boldsymbol{Q}$ are known.

Result 4.2.3. *The eigenvalues of $\boldsymbol{Q}$ are $\{2m, \underbrace{m, \ldots, m}_{m-1 \text{ times}}, \underbrace{2, \ldots, 2}_{\frac{m(m-3)}{2} \text{ times}}\}$, for $m \geq 3$.*

This is proved in Section 4.8.2. Now the distribution of T can be written in a closed form solution under H_0.

Result 4.2.4. *Under the null hypothesis,*

$$T = d(\hat{\boldsymbol{A}}_E, \hat{\boldsymbol{B}}_E)^2 \sim 2m\tau_0^2 \chi_1^2 + m\tau_0^2 \chi_{m-1}^2 + 2\tau_0^2 \chi^2_{\frac{m(m-3)}{2}}$$

approximately for large n_A and n_B where $n_A/n_B \to r \in (0, \infty)$, $\tau_0^2 = \frac{n_A + n_B}{n_A n_B}\sigma_0^2$ and the χ^2 terms are independent.

The proof is given in Section 4.8.3.

Of course the value of σ_0 would not be known in practice and hence would still need to

be estimated. Under H_0 we estimate σ_0^2 as the variance of the off-diagonal elements of the graph Laplacians $\{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_{n_A}, \boldsymbol{B}_1, \ldots \boldsymbol{B}_{n_B}\}$. A motivation for when the isotropic covariance matrix assumptions are approximately valid is if both samples belong to sets of Erdös-Renyi random network models defined in Section 1.2.5. For this case the isotropic covariance is derived in (4.2.15). The Erdös-Renyi random network model is a special case of a stochastic block model network that we consider next.

4.2.2 A parametric test assuming stochastic block model

For a stochastic block model defined in Section 1.2.5 the value of $\boldsymbol{\Sigma}$ is known when $\alpha = 1$, which we can use to write a distribution for our test statistic. We will firstly consider the most general case, in which every node is in its own block, hence $k = m$ and we can therefore write the probability of an edge between nodes i and j as p_{ij}. We define $P = (p_{ij})$ and set all $p_{ii} = 0$ to prevent loops.

We consider now the two sets $\mathcal{A}$ and $\mathcal{B}$ being modelled by stochastic block models with probability matrices $P_A = (p_{ij}^A)$ and $P_B = (p_{ij}^B)$ respectively. In this case the population Euclidean graph Laplacian mean for each set is entirely dependent on its probability matrix, hence the hypotheses to test become H_0: $P_A = P_B = P = (p_{ij})$ and H_1: $P_A \neq P_B$. The distribution of $\boldsymbol{x}$ under the stochastic block model is

$$\begin{aligned}
\boldsymbol{x} &\dot{\sim} \mathcal{N}\left(\left(\frac{n_A n_B}{n_A + n_B}\right)^{\frac{1}{2}} \boldsymbol{\mu}', \boldsymbol{\Sigma}\right) \\
\text{where } \boldsymbol{\mu}' &= (p_{12}^B - p_{12}^A, \ldots, p_{m-1\,m}^B - p_{m-1\,m}^A)^T, \\
\text{and } \boldsymbol{\Sigma} &= \text{diag}\left(\frac{n_B\, p_{12}^A\,(1 - p_{12}^A) + n_A\, p_{12}^B\,(1 - p_{12}^B)}{n_A + n_B}, \ldots, \right. \\
&\qquad \left. \frac{n_B\, p_{m-1\,m}^A\,(1 - p_{m-1\,m}^A) + n_A\, p_{m-1\,m}^B\,(1 - p_{m-1\,m}^B)}{n_A + n_B}\right),
\end{aligned} \tag{4.2.13}$$

the working for which is found in Section 4.8.4. Under the null hypothesis we see

$$\boldsymbol{\mu}' = (0, \ldots, 0)^T \text{ and } \boldsymbol{\Sigma} = \text{diag}\left(p_{12}(1 - p_{12}), \ldots \ , p_{m-1\,m}(1 - p_{m-1\,m})\right),$$

these can be substituted into Result 4.2.2 enabling us to find the approximate distribution of T_E for a specific P.

For the special case of a stochastic block model, where there is only 1 block the sets $\mathcal{A}$ and $\mathcal{B}$ are made of graph Laplacians representing Erdös-Renyi random networks, defined in Section 1.2.5, with probabilities p_A and p_B of any edge occurring respectively. In this case the $\boldsymbol{\mu}'$ and $\boldsymbol{\Sigma}$ in (4.2.13) become

$$\boldsymbol{\mu}' = (p_B - p_A, \ldots, p_B - p_A)^T, \tag{4.2.14}$$

$$\boldsymbol{\Sigma} = \frac{n_B\, p_A\,(1-p_A) + n_A\, p_B\,(1-p_B)}{n_A + n_B} \boldsymbol{I}_{\frac{m(m-1)}{2}}. \tag{4.2.15}$$

It is clear the population Euclidean graph Laplacian mean is entirely dependent on the probability of an edge occurring for an Erdös-Renyi random network, hence for the two-sample test the hypotheses simplify to H_0: $p_A = p_B = p$ and H_1: $p_A \neq p_B$. Under H_0 $\boldsymbol{x}$ has an isotropic covariance matrix just as in (4.2.12) and so the distribution of T_E can be written under H_0 using Result 4.2.4,

Result 4.2.5. *Under the null hypothesis for samples of Erdös-Renyi random networks,*

$$T_E = d_1(\hat{\boldsymbol{A}}_E, \hat{\boldsymbol{B}}_E)^2 \sim 2m\tau_0^2\chi_1^2 + m\tau_0^2\chi_{m-1}^2 + 2\tau_0^2\chi^2_{\frac{m(m-3)}{2}}$$

approximately for large n_A and n_B where $n_A/n_B \to r \in (0,\infty)$. Where $\tau_0^2 = \frac{n_A+n_B}{n_A n_B}p(1-p)$ and the χ^2 terms are all independent.

For the Erdös-Renyi random networks we in fact can also compute explicitly the distribution for H_1.

Result 4.2.6. *Under the alternative hypothesis for samples of Erdös-Renyi random networks,*

$$\begin{aligned} T_E = d_1(\hat{\boldsymbol{A}}_E, \hat{\boldsymbol{B}}_E)^2 \sim & 2m\tau_1^2\chi_1^2\left(\frac{(p_B-p_A)^2 m(m-1)}{2\tau_1^2}\right) \\ & + m\tau_1^2\chi_{m-1}^2 + 2\tau_1^2\chi^2_{\frac{m(m-3)}{2}}, \end{aligned}$$

approximately for large n_A and n_B where $n_A/n_B \to r \in (0,\infty)$. Where $\tau_1^2 = \left(\frac{n_A n_B}{n_A+n_B}\right)^{-1}\frac{n_B\, p_A\,(1-p_A)+n_A\, p_B\,(1-p_B)}{n_A+n_B}$ and $\chi^2(a)$ is the non-central χ^2 distribution with non-centrality parameter a, these χ^2 terms are all independent.

The proof is in Section 4.8.5.

In practice we would not know the values of p_A, p_B and p and hence we need to estimate these. Under H_0 we estimate p as the arithmetic mean of the off-diagonal elements of the graph Laplacians $\{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_{n_A}, \boldsymbol{B}_1, \ldots \boldsymbol{B}_{n_B}\}$, which is a consistent es-

timator under H_0 and the Erdös-Renyi model. Similarly under H_1 we can estimate p_A and p_B as the arithmetic mean of the off-diagonal elements of the graph Laplacians $\{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_{n_A}\}$ and $\{\boldsymbol{B}_1, \ldots \boldsymbol{B}_{n_B}\}$ respectively which are again consistent estimators under H_1 and the Erdös-Renyi model.

4.3 Non-parametric tests

The approach in the preceding sections relied on either estimating $\boldsymbol{\Sigma}$ directly or imposing strong parametric modelling assumptions, and it was also limited to the case with d being the Euclidean power metric, i.e. it was not appropriate when d is the Procrustes power metric. An alternative approach to avoid these limitation is to develop a non-parametric test. We use a random permutation test similar to that in Preston and Wood (2010), which we define in Algorithm 1 for r permutations.

Algorithm 1 *Random permutation test to test the equality of means for two samples of graph Laplacians, $\hat{\mathcal{A}} = \{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_{n_A}\}$ and $\hat{\mathcal{B}} = \{\boldsymbol{B}_1, \ldots \boldsymbol{B}_{n_B}\}$, using the test statistic T.*

1: Calculate the test statistics between $\hat{\mathcal{A}}$ and $\hat{\mathcal{B}}$, given by $T = T(\hat{\mathcal{A}}, \hat{\mathcal{B}})$.
2: Generate random sets $\hat{\mathcal{A}}^*$ and $\hat{\mathcal{B}}^*$ of size n_A and n_B respectively, by randomly sampling without replacement from $\hat{\mathcal{A}} \cup \hat{\mathcal{B}}$.
3: Compute the test statistic of sets $\hat{\mathcal{A}}^*$ and $\hat{\mathcal{B}}^*$, given by $T^* = T(\hat{\mathcal{A}}^*, \hat{\mathcal{B}}^*)$.
4: Repeat steps 2 and 3 r times, to give test statistics $T_1^*, T_2^*, \ldots, T_r^*$.
5: Order the test statistics $T^*(1) \leq T^*(2) \leq \cdots \leq T^*(r)$.
6: Calculate the p-value, which is $1 - \frac{j}{r}$ for the minimum $1 \leq j \leq r-1$ satisfying $T^*(j) < T \leq T^*(j+1)$, unless $T \leq T^*(1)$, in which case the p-value is 1 or if $T > T^*(r)$, in which case the p-value is 0.

The random permutation test approximates the distribution of T under H_0 by finding multiple T values for the permuted samples where H_0 holds. With this approximated distribution we can then find the significance of the test statistic for the true samples. A limitation of using the permutation test is it assumes exchangeability of the observations under the null hypothesis (Amaral et al., 2007). This means under the null hypothesis the populations of sets $\mathcal{A}$ and $\mathcal{B}$ are assumed identical which is not always the case. When we apply our methods to data we will consider an example of simulated data where null exchangeability does not hold, to examine the consequences.

4.4 Comparing the test statistics

We will now compare our three test statistics, T_E, T_H and T_S, defined in (4.0.4), (4.0.5) and (4.0.6), along with the Ginestet et al. (2017) test statistic, T_G. We investigate how good an approximation the asymptotic distributions are for T_E and T_G, found in Sections 4.2 and 4.1, for synthetic and real data. Problems with the convergence are likely when the covariance estimate is poor from lack of data, or when the off-diagonal elements are near 0 in expectation. We will describe a good indication of the adequacy of the approximation by checking the size of the test. We will see in Section 4.5 cases when the approximation is poor and in these it is better to perform this test using a random permutation test to simulate the distribution under H_0.

For an approximation of the distribution of a test statistic, T, under H_0 to be suitable then the *empirical size* of the test, P(reject $H_0|H_0$ true), should be close to the *nominal size* of the test, $100a\%$. The empirical size of the test can be rewritten as P(p-value $<$ $a|H_0$ true). The nominal size is set before the test whilst the empirical size is calculated after a test and it is calculated differently for simulated and real data.

For simulated data Monte Carlo simulations are used to simulate the datasets repeatedly, for the v^{th} Monte Carlo realisation p_v is the p-value computed from Algorithm 1 or an asymptotic distribution, the empirical size is then given as

$$\frac{1}{M}\sum_{v=1}^{M}\mathbb{1}_{\{\mathrm{p}_v\leq a\}}, \tag{4.4.1}$$

where $\mathbb{1}$ is the indicator function.

For real data, when Monte Carlo simulations cannot be used, under tests with asymptotic distributions where the cut-off value, $T_{100a\%}$ is known, the empirical size is found as

$$\frac{1}{r}\sum_{v=1}^{r}\mathbb{1}_{\{T_v^*>T_{100a\%}\}}, \tag{4.4.2}$$

from running Algorithm 1 once. Essentially this calculates the size by comparing the asymptotic distribution with a distribution calculated from permutations.

A test is considered successful if it produces a high power provided the size is correct, where power is P(reject $H_0|H_1$ true) . The power of a test can only be calculated when

the distribution of the test statistic, $T \in \{T_E, T_H, T_S, T_G\}$, under H_1 can be found. We can only calculate the distribution under H_1 for our synthetic data as we can simulate the distribution under different alternate simple hypotheses. For the synthetic data the distribution under H_1 can be approximated by running step 1 of Algorithm 1 M times for sets under H_1, to create the collection of test statistics $\{T^1, \ldots, T^M\}$. The distribution of the collection $\{T^1, \ldots, T^M\}$ will then approximate the distribution of a T under H_1 and the power can be estimated by finding the probability of rejecting H_0, i.e. $T > T_{100a\%}$, under the simulated alternative distribution.

We now compare the test statistics for different data. We will see and explain from the synthetic and neuroimaging data that our test statistic using a permutation test consistently out performs using the tests with asymptotic distributions and so for the later data in this chapter we shall only consider the permutation tests.

4.5 Simulation study

We now apply our two-sample test to synthetic data, with $n = n_A = n_B$. Two of the network models we use for the simulated data are the Erdös-Renyi random networks (E.R) and the Watts-Strogatz small-world model (W.S) described in Section 1.2.5. For the W.S model we fix $nei_A = nei_B = nei$ as the neighbourhood sizes and p_A and p_B are the respective rewiring probability. We also use the normal model (N), defined in Section 1.2.5, which produces networks with weights $w_{ij} \sim \mathcal{N}(p_B, \sigma^2)$, for $1 \leq i, j \leq m$. For all models the hypotheses simplify to H_0: $p_A = p_B = p$ and H_1: $p_A \neq p_B$. The aim of the simulation study is to check the convergence for the cases we have an asymptotic distribution for the test statistic and to compare the powers for the different test statistics.

To check the convergence of T_E and T_G distributions under H_0 we look at the values of the empirical size, P(reject $H_0|H_0$ true), found in Table 4.1 for different synthetic data. We expect the empirical size to be $\approx 100a\%$ and we set $100a\% = 5\%$. We run $M = 1000$ Monte Carlo simulations to find the empirical size, given in (4.4.1), of the tests using T_E and T_G for both tests with asymptotic distributions and non-parametric tests. For the non-parametric test we use Algorithm 1 with $r = 100$ permutations. The distribution used in Asy (4.2.4) is that of Result 4.2.4 where the cut-off is found by simulating this distribution for 100000 values. The Asy (4.2.4) distribution requires the

estimation of a parameter, when the graph Laplacian represent networks which model E.R networks the p in Result 4.2.5 is estimated, whilst for all other cases the σ_0 value of Result 4.2.4 is estimated.

From Table 4.1 when T_G is used with its asymptotic distribution, we can see the empirical size rarely matches 5% for higher dimensional networks; this indicates the approximation of the distribution for T_G is poor for these examples, most likely through poor covariance matrices estimates. When using the asymptotic distribution for T_E from Result 4.2.4 the assumptions are met for the E.R and N models and in these cases the size is approximately 5%, therefore the approximation seems good. The assumptions for Result 4.2.4 are not met for the W.S model and in these cases the size is not around 5% showing the distribution is a poor fit. Even for the more general asymptotic distribution from Result 4.2.2 for T_E the size is not near 5% for the W.S model, but is close to 5% for the other models.

Model	**m**	**n**	**Variables**	**Empirical Size** ($\%$)					
				T_E			T_H	T_G	
				Asy (4.2.4)	Asy (4.2.2)	Perm	Perm	Asy	Perm
E.R	5	100	$p = 0.5$	**6.0**	**6.4**	**6.2**	**5.9**	**5.1**	6.6
E.R	5	100	$p = 0.1$	**4.4**	**4.9**	**5.3**	**4.8**	**4.9**	**5.7**
E.R	10	100	$p = 0.5$	**4.4**	**4.5**	**5.3**	**5.5**	**4.1**	**5.2**
E.R	10	100	$p = 0.1$	**3.9**	**4.8**	**4.8**	**5.3**	**4.8**	**5.5**
E.R	50	100	$p = 0.5$	**4.4**	**3.9**	**6.3**	**5.9**	1.5	**5.5**
E.R	50	100	$p = 0.1$	**5.1**	**3.8**	**5.8**	**5.6**	0.8	**4.9**
W.S	5	100	$p = 0.1, nei = 1$	0.0	2.2	**6.3**	6.4	**4.4**	**6.2**
W.S	5	100	$p = 0.5, nei = 1$	1.3	3.4	**5.2**	**5.1**	**3.9**	**5.0**
W.S	40	100	$p = 0.1, nei = 1$	0.0	0.0	**6.1**	**6.2**	0.0	**4.4**
W.S	40	100	$p = 0.5, nei = 1$	0.0	0.5	**6.0**	**6.3**	1.6	**6.0**
E.R&N	10	100	$p = 0.5, \sigma = 0.01$	**4.7**	**5.2**	**5.8**	87.6	**4.9**	6.6
E.R&N	10	100	$p = 0.5, \sigma = 0.1$	**4.8**	**5.2**	**5.9**	82.8	**4.0**	6.6
E.R&N	20	100	$p = 0.5, \sigma = 0.01$	**4.5**	**5.5**	**5.3**	77.8	2.5	7.7
E.R&N	20	100	$p = 0.5, \sigma = 0.1$	**5.9**	6.5	**5.7**	69.1	3.3	7.0

Table 4.1: *A table comparing the test statistics T_E, T_H and T_G. Asy indicates the size was found using the asymptotic distribution of H_0, Asy (4.2.4) and Asy(4.2.2) are when the asymptotic distribution used is from Result 4.2.4 and Result 4.2.2 respectively. Perm indicates a permutation test was used. The critical value was set to give a nominal size of 5%, bold values indicate the empirical size is within 1.96 standard errors of 5%, which is $5\pm1.351\%$.*

We were interested how the non-parametric tests performed when null exchangeability

does not hold. When the E.R or the W.S models were used in Table 4.1 null exchangibility holds and the empirical size under the random permutation tests, for T_E, T_H and T_G, match up well to 5%. Null exchangibility does not hold for the last 4 rows in Table 4.1. For these examples the mean of both sets $\mathcal{A}$ and $\mathcal{B}$ are equal but the variance differs. We expected this may cause problems when using the random permutation test. When using T_G and T_H the empirical size does not match 5% well, however when T_E is used the empirical size is still close to 5%.

Figure 4.2 contains power plots for synthetic data to compare our three test statistics along with T_G. The power is calculated as described in Section 4.4. Random permutation tests are used to perform the test when using T_H and T_S. We include powers using the asymptotic distribution of T_E and T_G as well as using a random permutation tests for both, to allow for poor convergence. For the test statistic T_E we calculate it by the asymptotic distribution in Result 4.2.2. We set $r = 100$ for the random permutation tests and use $M = 1000$ to estimate the power. For Figure 4.2a and 4.2b the sets $\mathcal{A}$ and $\mathcal{B}$ contained graph Laplacians representing E.R networks with $p_A = p = 0.1$, $m = 5$ and 10 respectively. For these examples all the test statistics perform well. However T_E, using it both asymptotically and by permutation, and T_H perform best giving larger powers when $p_B \neq p_A$. For Figure 4.2c the graph Laplacians represent W.S models with $nei = 1$, $p_A = p = 0.1$ and $m = 5$. Here all of the tests perform very similarly, with T_G both used asymptotically and by permutation and T_E by permutation performing the best, giving higher power for $p_B \neq p_A$. We have already seen in Table 4.1 that the asymptotic distribution for T_G fits this model with $m = 5$ well, but the asymptotic distribution for T_E does not fit this model well, also when using T_E asymptotically we get the lowest powers.

When the mean degrees between the sets are different, found on the graph Laplacian diagonal, these will contribute a lot to the test statistic, however the test can also differentiate sets with similar or even identical mean degrees. For example, when $m = 10$, let set $\mathcal{A}$ be the set of E.R networks with $p_A = \frac{4}{9}$ and $\mathcal{B}$ be the set of W.S networks with $p = 0.1$ and $nei = 2$. The mean degree for all nodes is 4, but applying our two-sample test, with any $T \in \{T_E, T_H, T_S, T_G\}$, the power of the test is 1, meaning the null is always rejected. This shows an advantage of using the graph Laplacian matrix over the degree matrix, defined in (1.2.3).

From the simulation study we have seen for large dimensions the asymptotic distributions for T_E and T_G are poor. When the data fits having an isotropic covariance matrix

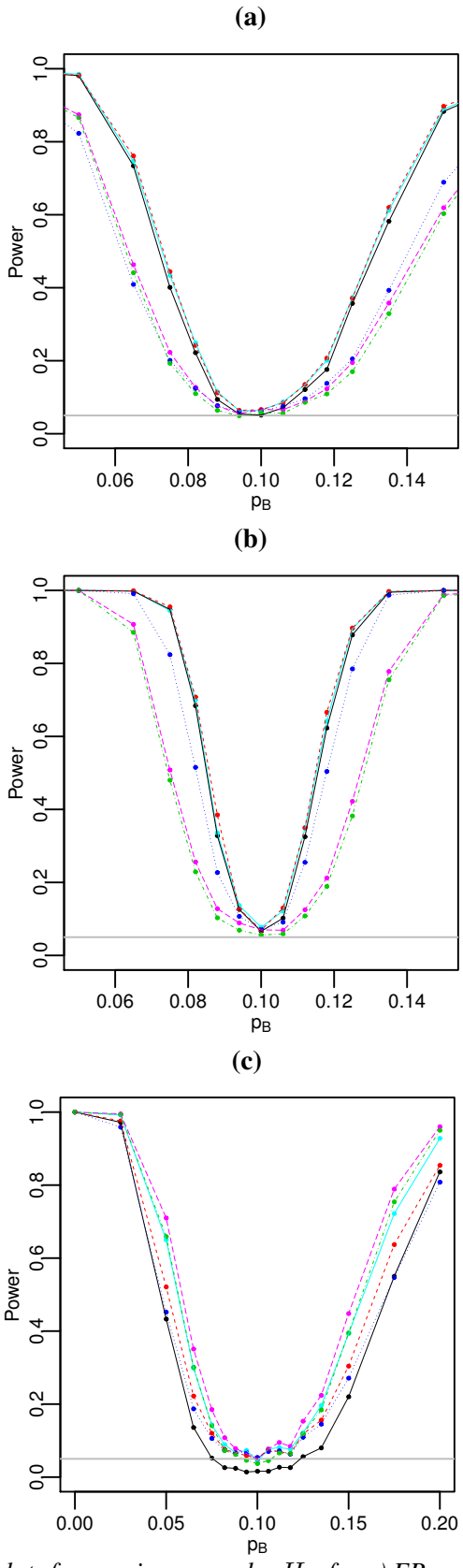


Figure 4.2: *Power plots for varying p_B under H_1, for a) ER model $p = 0.1$ and $m = 5$, b) ER model $p = 0.1$ and $m = 10$ and c) WS model $nei = 1$, $p = 0.1$ and $m = 5$. Black-T_E asymptotic, light blue-T_E permutation, red-T_H, dark blue-T_S, green-T_G asymptotic and pink-T_G permutation.*

the distribution for T_E fits well and gives powerful results, but this is too restrictive as few real datasets will fit this model. So for different datasets neither T_E or T_G used asymptotically will perform consistently well. Therefore it is more appropriate to use a test statistic with a permutation test, of these T_E consistently performs well, as well as being the most computationally appealing. We now consider real data, for which we will discuss how the conclusions from the simulation study similarly hold for this data too.

4.6 Application of the two-sample test to network data

We now apply our test to real network data. In Example 4.6.1 we perform our test on the FCP neuroimaging data described in Section 1.3.3. Example 4.6.2 uses the NLAB data described in Section 1.3.2. In Example 4.6.3 we perform the test on the novel data and explore in more detail the difference in mean between Austen and Dickens.

Example 4.6.1: Two-sample test applied to the neuroimaging data
For the FCP neuroimaging dataset, introduced in Section 1.3.3, it is of interest to test if there is a significant difference between functional connectivity for gender. We perform two-sample tests on gender using T_E, T_H, T_S and T_G. As stated in Section 1.3.3 the neuroimaging covariance matrices have a quantile thresholding value, c, to convert them to networks. We run our tests for different c values to ensure this value is not having a large effect on the performance of each test statistic.

For the two-sample tests we set the significance value at the 5% level, and therefore the empirical size is expected to be 5% for all test statistics. The random permutation tests used are those defined in Algorithm 1, using $r = 1000$. We run the test not only on the full sample of 462 males and 555 females, but also a sub-sample of 50 networks for each gender, as this is more representative of typical neuroimaging sample sizes. We find the empirical size of the test when using T_E and T_G asymptotically, by comparing the asymptotic distribution with the distribution from permutations, given in (4.4.1).

Table 4.2 provides the results for the different thresholding values. The size for T_G used asymptotically is never near 5%, this indicates the asymptotic distribution of T_G is a very poor fit for this dataset, this is almost certainly due to the small sample size estimating a large covariance matrix. When using T_E with its asymptotic distribution the size is very far from 5% for the large sample but surprisingly close for the smaller

sample. Because of the problems with the approximation of distribution for both T_G and T_E it is more appropriate to use these by permutation test. Using T_E, T_H, T_S or T_G's random permutation test give significant results at the 5% level. If we are to assume that there is a difference in brain activity for gender which seems to be the conclusion, then T_E, T_H, T_S and T_G's random permutation test all perform well. However due to the need to estimate and invert a large covariance matrix many times in T_G's random permutation test it is very computationally intensive, and so our test statistics seem favourable for larger dimensional graph Laplacians, particularly T_E which is the least computationally intensive.

Sample size	**Threshold**	**Size**		**P-value**					
		T_G Asy	T_E Asy	T_E Asy	T_E Perm	T_H Perm	T_S Perm	T_G Asy	T_G Perm
Full	0.2 Q	20.8	96.8	**0.000**	**0.002**	**0.006**	**0.006**	**0.000**	**0.000**
Full	0.4 Q	89.5	93.4	**0.000**	**0.002**	**0.011**	**0.011**	**0.000**	**0.000**
Full	0.6 Q	91.5	92.4	**0.000**	**0.001**	**0.006**	**0.006**	**0.000**	**0.000**
Full	0.8 Q	9.3	92.7	**0.000**	**0.000**	**0.001**	**0.000**	**0.000**	**0.000**
50	0.2 Q	0	6.6	0.920	**0.000**	**0.000**	**0.000**	0.995	**0.000**
50	0.4 Q	0	3.3	0.996	**0.000**	**0.000**	**0.000**	1.000	**0.000**
50	0.6 Q	0	3.0	0.999	**0.000**	**0.000**	**0.000**	0.406	**0.000**
50	0.8 Q	0	1.0	0.988	**0.000**	**0.000**	**0.000**	0.929	**0.000**

Table 4.2: *A table with the p-value of the two-sample test when using T_G, T_E and T_H for the FCP dataset. The empirical size is included when T_G and T_E are used asymptotically. Bold when the p-value is under 0.05.*

Example 4.6.2: Two-sample test applied to the M-money transaction data

For the M-money transaction networks, described in Section 1.3.2, we hypothesise that the networks for a weekday will be different to the networks on a weekend, as a weekday may correspond to more business transactions than the weekend. However it is not clear from the PCA plots in Example 2.5.2 if this hypothesis is true and so we use our two-sample test to test this hypothesis by testing if the mean network for a weekday is different to the mean network for a weekend day.We perform out test using the permutation test from Algorithm 1 with $r = 1000$, as we have seen that for large networks, like these, that the asymptotic tests are not appropriate. For each metric all permuted values were less than the observed test statistic leading to p-values of 0, meaning there is strong evidence for a difference in mean networks, for M-money transactions in Tanzania, on weekdays and weekends.

Example 4.6.3: Two-sample test applied to the novel data

It is interesting to test if different authors have significantly different writing styles and by representing text as graph Laplacians we can apply our two-sample test to address this question. We run our two-sample test for the sets of networks representing Austen and Dickens novels which gives the test statistics $T_E = 0.0011$, $T_H = 0.0690$, $T_S = 0.0689$. We compute the p-value from the permutation test from Algorithm 1 with $r = 1000$ permutations for each of T_E, T_H, T_S and in each case all permuted values were less than the observed statistics for the data. Hence, in each case the estimated p-value is zero, indicating very strong evidence for a difference in mean graph Laplacian for the authors.

In Example 2.5.1 we saw from the PC plots Dickens and Austen works were very well separated and so it is not surprising they have significantly different means. A question with a less obvious answer is does Dickens' work significantly differ to the whole collection of 19th century authors' work, described in Section 1.3.1, and we again use our two-sample test to address this question. The p-value is calculated from the permutation test with $r = 1000$, giving p-values 0.003, 0 and 0 for the Euclidean, square root Euclidean and Procrustes shape-and-size respectively, therefore in each case there is significant evidence that the means of Dickens novels is different to the other 19th century novels. We will look in more detail at the difference between Dickens and all the other 19th century authors again in Example 5.1.1.

4.6.1 Exploring difference between Austen and Dickens

Given that the Austen and Dickens novels are significantly different in mean we would like to explore how they differ. We provide a method of doing so in Severn et al. (2019). In particular we examine the off-diagonal elements of $P_{\mathcal{L}}(\hat{\boldsymbol{\eta}_{Dickens}}) - P_{\mathcal{L}}(\hat{\boldsymbol{\eta}_{Austen}})$, i.e. the differences in the mean weighted adjacency matrix, and compare them to appropriate measures of standard error of the differences using a z-statistic. The method of comparison we provide is multiple univariate tests, similar ideas have been used before for network analysis, for example in Ginestet et al. (2014).

The histograms of the off-diagonal individual graph Laplacians are heavy tailed, and a plot of sample standard deviations versus sample means, found in Figure 4.3, show an overall average linear increase with approximate slope $\beta = 0.2$, but with a large spread. We shall use this relationship in a regularised estimate of our choice of standard error. For a particular co-occurrence pair of words we have weighted adjacency

values $x_i, i = 1, \ldots, n_1$ and $y_j, j = 1, \ldots, n_2$ with sample means $\bar{x}$ and $\bar{y}$, and sample standard deviations s_x and s_y. For our analysis here we use the Euclidean mean graph Laplacians. We estimate the variance in our sample with a weighted average of the sample variance and an estimate based on the linear relationship between the mean and standard deviation, and in particular the population pooled variance is estimated by

$$s_p^2 = \frac{n_1(w_1 s_x^2 + (1-w_1)\beta^2\bar{x}^2) + n_2(w_2 s_y^2 + (1-w_2)\beta^2\bar{y}^2)}{(n_1 + n_2 - 2)},$$

where the weights are taken as $w_i = n_i/N, i = 1, 2$, where we take $N = 200$. If all values in one of the samples are 0 (due to no word co-occurrence pairings in any of that author's books) then we drop that word pairing from further analysis, as we are only interested in the relative usage of the word occurrences that are actually used by both authors. A univariate z-statistic for comparing adjacencies is then

$$z = \frac{\bar{x} - \bar{y}}{(\xi + s_p)\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{4.6.1}$$

where we include the regularizing offset $\xi > 0$ to avoid highlighting very small differences in mean adjacency with very small standard errors. The value for ξ is chosen as the median of all s_p values under consideration.

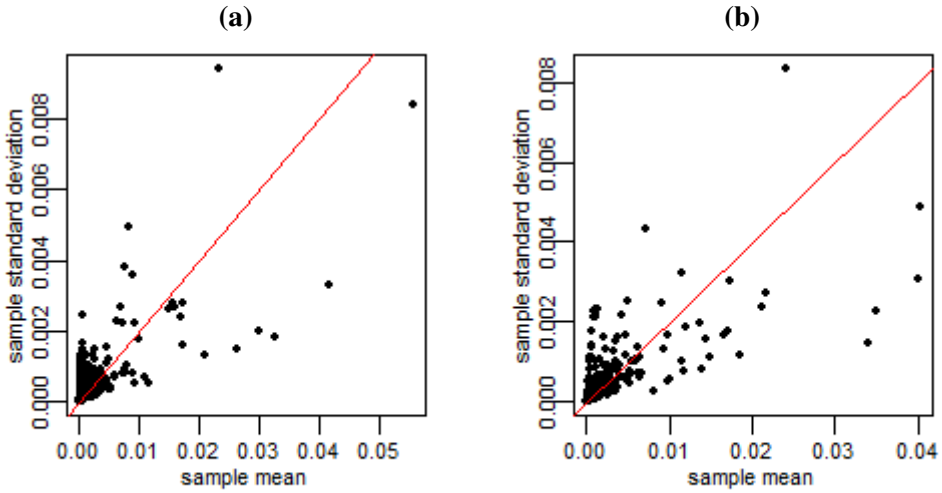
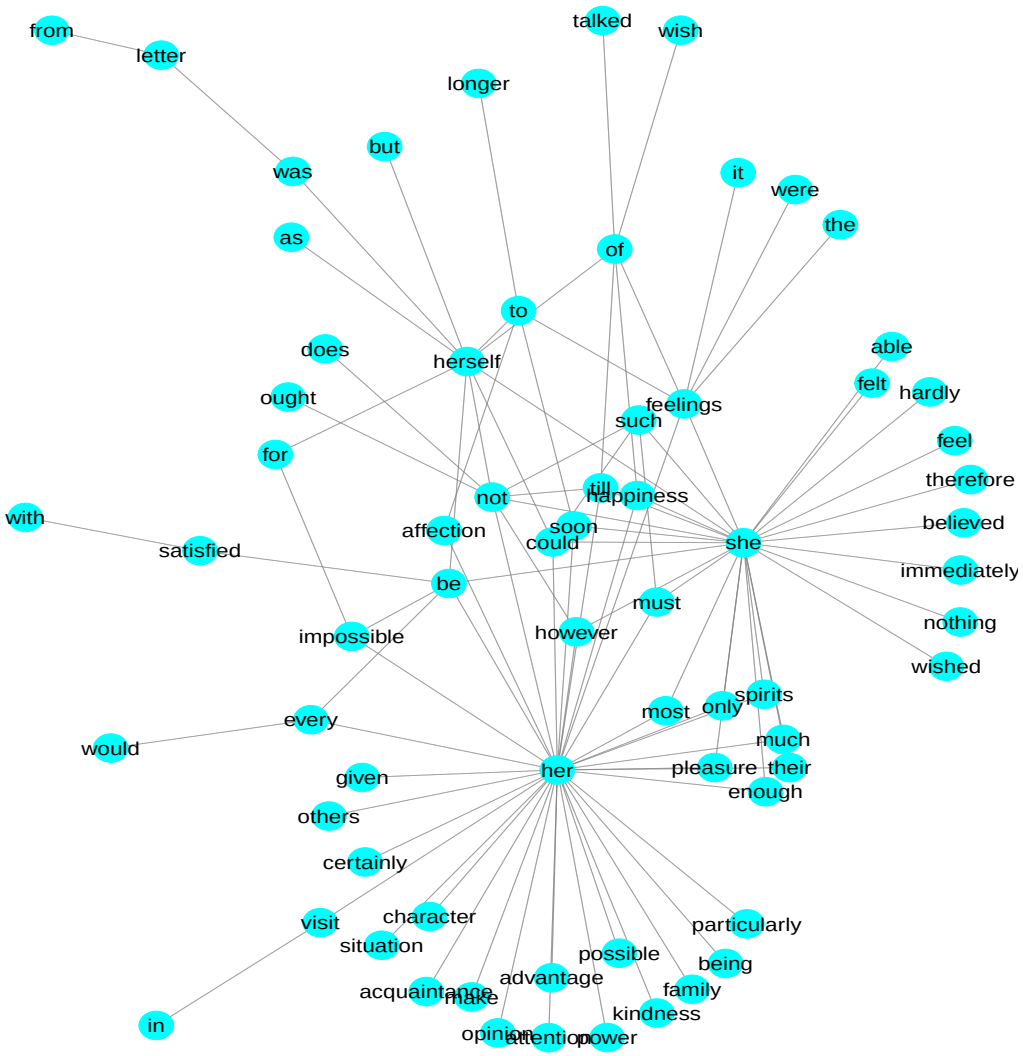


Figure 4.3: *The sample mean vs standard deviation for each off-diagonal element for the (a) Dickens novels and (b) Austen novels. The red line has intercept* 0 *and gradient* 0.2.

The exploratory graphical displays in Figure 4.3 illuminate striking differences between the novelists. For Austen there are very common pairings of words with '*her*', '*she*',


from
letter
talked
wish
longer
but
was
it
were
the
as
of
to
does
herself
able
felt
hardly
ought
feelings
such
feel
for
therefore
not
till
happiness
with
affection
soon
could
she
believed
satisfied
be
immediately
must
nothing
however
impossible
wished
spirits
every
most
only
much
would
her
pleasure
their
given
enough
others
certainly
visit
character
particularly
situation
possible
being
advantage
acquaintance
make
family
kindness
in
opinion
attention
power

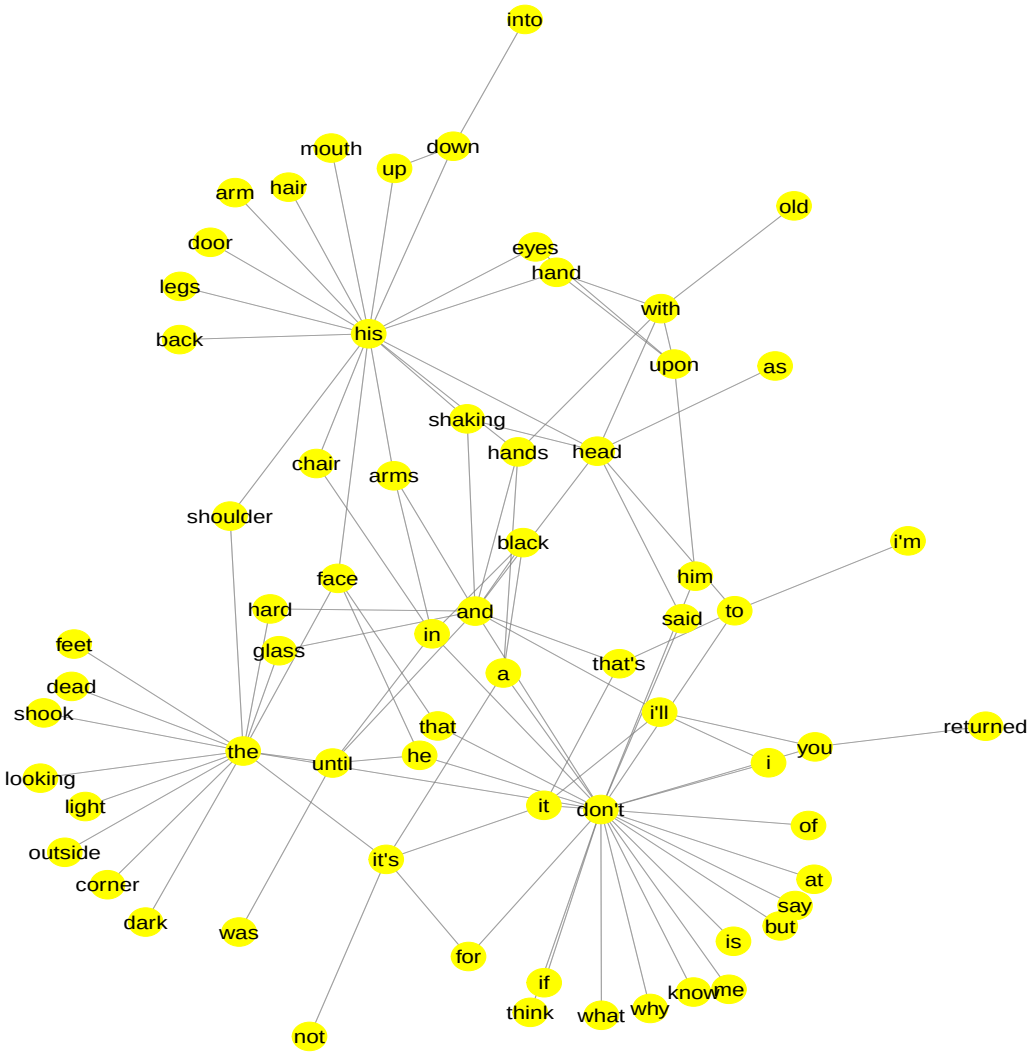


Figure 4.3: *Networks displaying the top 100 pairs of words ranked according to the z-statistic in (4.6.1), with more prominent co-occurrences used by Austen (top, in blue) and the more prominent co-occurrences used by Dickens (bottom, in yellow).*

'*herself*', which form important hubs in this network. Austen also pairs these hubs with more emotional words '*feelings*', '*felt*', '*feel*', '*kindness*', '*happiness*, '*affection*', '*pleasure* and stronger words '*power*', '*attention*', '*must*', '*certainly*', '*advantage*' and '*opinion*'. Also we see more use of '*letter*' in Austen, which is a literary device often used by the author. For Dickens there are more common uses of abbreviations, especially '*don't*' which is an important hub, and also '*it's*', '*i'll*' and '*that's*'. In contrast the Austen network highlights '*not*'. Dickens also more prominently pairs body parts *'arm', 'arms', 'eyes', 'feet', 'hair', 'hand', 'hands', 'head', 'mouth', 'face', 'shoulder', 'legs'* in combination with the strong hubs '*his*' and '*the*'. Dickens use of body parts is an interesting finding that has been noted and studied before in Mahlberg (2013). The hubs '*his*' and '*the*' are also paired with other objects, such as *'door', 'chair', 'glass'*. Finally, Dickens has the more prominent use of pairs with a sombre word, such as '*dark*', '*black*' and '*dead*', which might have been expected.

4.7 Summary

In this Chapter we have defined a two-sample test to test the equality of means of samples of graph Laplacians using our graph Laplacian framework. The two-sample test has a test statistic which is the distance squared between the sample's unprojected means. The two-sample test is general and could be easily adapted to be used with many different distance metrics between graph Laplacians. We specifically looked at the test when using the Euclidean, square root Euclidean and Procrustes size-and-shape metrics. For the Euclidean metric the distribution of the test statistic could be found asymptotically however unless simple models hold for the data then this asymptotic distribution requires the estimation of a large covariance matrix.

We compare our three tests with a similar test proposed in Ginestet et al. (2017) on simulated data and the neuroimaging data. Ginestet's test also requires the estimation of a large covariance matrix and for this reason we see that for graph Laplacians with large dimensions Ginestet's and the Euclidean tests with asymptotic distributions are not suitable to use, giving incorrect empirical sizes. Instead any of our tests used using a permutation test are more suitable. For our two-sample tests, using the Euclidean metric is computationally more appealing, as no square rooting or Procrustes analysis is needed, and as using the Euclidean metric consistently gives suitable results it seems that this test is favourable.

We apply the two-sample test also to the M-money transaction data where we can see that money transfer networks for weekends are significantly different than weekdays. We also apply the test to the 19th century author datasets, where the test shows significance difference between means for the different authors. To study what are the main differences in the means for the 19th century novels dataset we propose a method to determine the co-occurrences which differ the most significantly between authors. This method gave insightful results that agree with previous findings for the authors.

4.8 Calculations for Chapter 4

4.8.1 Alternative to T_G using the diagonal

Proof of equivalent distributions

Result 4.1.1. *T'_G has an identical asymptotic distribution to T_G.*

Proof. We know from Ginestet et al. (2017),

$$n_A^{\frac{1}{2}}(\text{vech}(\hat{\boldsymbol{A}}_E) - \text{vech}(\boldsymbol{\mu}_A^E))^T \dot{\sim} \mathcal{N}(0, \boldsymbol{\Sigma}''_A),$$
$$n_B^{\frac{1}{2}}(\text{vech}(\hat{\boldsymbol{B}}_E) - \text{vech}(\boldsymbol{\mu}_B^E))^T \dot{\sim} \mathcal{N}(0, \boldsymbol{\Sigma}''_B),$$

where vech is defined in (0.0.2), and $\boldsymbol{\Sigma}''_A = \text{Cov}(\text{vech}(\boldsymbol{A}))$ and $\boldsymbol{\Sigma}''_B = \text{Cov}(\text{vech}(\boldsymbol{B}))$, for $\boldsymbol{A} \in \mathcal{A}$ and $\boldsymbol{B} \in \mathcal{B}$. Note now $\boldsymbol{\Sigma}''_A$ and $\boldsymbol{\Sigma}''_B$ are $\frac{m(m+1)}{2} \times \frac{m(m+1)}{2}$ matrices. With simple manipulation we see,

$$\left(\frac{n_A n_B}{n_A + n_B}\right)^{\frac{1}{2}} (\text{vech}(\hat{\boldsymbol{A}}_E) - \text{vech}(\hat{\boldsymbol{B}}_E))^T \dot{\sim} \mathcal{N}\left(\left(\frac{n_A n_B}{n_A + n_B}\right)^{\frac{1}{2}} (\text{vech}(\boldsymbol{\mu}_A^E) - \text{vech}(\boldsymbol{\mu}_B^E)), \boldsymbol{\Sigma}''\right).$$

Under H_0 we set $\text{vech}(\boldsymbol{\mu}_A^E) = \text{vech}(\boldsymbol{\mu}_B^E)$, and assume $\boldsymbol{\Sigma}''_A = \boldsymbol{\Sigma}''_B = \boldsymbol{\Sigma}''$ and $\hat{\boldsymbol{\Sigma}}'' \to \boldsymbol{\Sigma}''$, giving

$$\left(\frac{n_A n_B}{n_A + n_B}\right)^{\frac{1}{2}} (\text{vech}(\hat{\boldsymbol{A}}_E) - \text{vech}(\hat{\boldsymbol{B}}_E))^T \dot{\sim} \mathcal{N}\left(\mathbf{0}, \hat{\boldsymbol{\Sigma}}''\right).$$

As we have now included the diagonal of the graph Laplacians in the test statistic the covariance matrix $\hat{\boldsymbol{\Sigma}}''$ will not be full rank ($\frac{m(m+1)}{2}$), due to the fact that the diagonal is dependant on the off-diagonal of a graph Laplacian. The covariance matrix is at most rank $\frac{m(m-1)}{2}$ and therefore $\hat{\boldsymbol{\Sigma}}''$ is singular. We can use the fact $\hat{\boldsymbol{\Sigma}}''$ is a symmetric square matrix to write by spectral decomposition,

$$\hat{\boldsymbol{\Sigma}}'' = \boldsymbol{U}\Lambda\boldsymbol{U}^T$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_{\frac{m(m-1)}{2}}, \underbrace{0, \ldots, 0}_{m \text{ times}})$ and λ_i are the non zero eigenvalues of $\hat{\boldsymbol{\Sigma}}''$, $\boldsymbol{U}$'s columns are the eigenvectors of $\hat{\boldsymbol{\Sigma}}''$ and $\boldsymbol{U} = \boldsymbol{U}^T$. The Moore-Penrose inverse is

$$\hat{\boldsymbol{\Sigma}}''^{-} = \boldsymbol{U}\Lambda^{-}\boldsymbol{U}^T, \tag{4.8.1}$$

where $\Lambda^{-} = \text{diag}(\lambda_1^{-1}, \ldots, \lambda^{-1}_{\frac{m(m-1)}{2}}, \underbrace{0, \ldots, 0}_{m \text{ times}})$. It is clear then that,

$$\left(\frac{n_A n_B}{n_A + n_B}\right)^{\frac{1}{2}} (\text{vech}(\hat{\boldsymbol{A}}_E) - \text{vech}(\hat{\boldsymbol{B}}_E))^T \boldsymbol{U}(\Lambda^{-})^{\frac{1}{2}} \dot{\sim} \mathcal{N}\left(\mathbf{0}, \text{diag}(\underbrace{1, \ldots, 1}_{\frac{m(m-1)}{2} \text{ times}}, \underbrace{0, \ldots, 0}_{m \text{ times}})\right),$$

and hence

$$T_G{}' = \frac{n_A n_B}{n_A + n_B}(\text{vech}(\hat{\boldsymbol{A}}_E) - \text{vech}(\hat{\boldsymbol{B}}_E))^T \hat{\boldsymbol{\Sigma}}''^{-}(\text{vech}(\hat{\boldsymbol{A}}_E) - \text{vech}(\hat{\boldsymbol{B}}_E)) \xrightarrow{D} \chi^2_{\frac{m(m-1)}{2}}.$$

□

Proof of equality of test statistics

Result 4.1.2. *$T_G = T'_G$ when $n_A - 1, n_B - 1 \geq \frac{m(m-1)}{2}$ and the standard unbiased estimator for covariance matrices is used instead of the shrinkage estimator from Schäfer and Strimmer (2005).*

Proof. When $n_A - 1 \geq \frac{m(m-1)}{2} \leq n_B - 1$ the estimate, $\hat{\boldsymbol{\Sigma}}'$, from the test in Ginestet et al. (2017), and $\hat{\boldsymbol{\Sigma}}''$ can be estimated by the sample covariance matrices without the need for the shrinkage estimator, as n_A and n_B are large enough for the covariance matrices to reach their maximum rank.

We will prove

$$T_G = \frac{n_A n_B}{n_A + n_B} \boldsymbol{z}^T \hat{\boldsymbol{\Sigma}}'^{-1} \boldsymbol{z} = \frac{n_A n_B}{n_A + n_B} \boldsymbol{y} \hat{\boldsymbol{\Sigma}}''^{-} \boldsymbol{y} = {T_G}',$$
$$\text{where } \overset{\frac{m(m-1)}{2} \times 1}{\boldsymbol{z}} = \phi(\hat{\boldsymbol{A}}_E) - \phi(\hat{\boldsymbol{B}}_E)$$
$$\overset{\frac{m(m+1)}{2} \times 1}{\boldsymbol{y}} = \text{vech}(\hat{\boldsymbol{A}}_E) - \text{vech}(\hat{\boldsymbol{B}}_E).$$

We also define $\overset{m \times 1}{\boldsymbol{d}} = \text{diag}(\hat{\boldsymbol{A}}_E) - \text{diag}(\hat{\boldsymbol{B}}_E)$. As the diagonal elements of a graph Laplacians are a linear combination of the upper triangular elements we can write $\boldsymbol{d} = \boldsymbol{F}\boldsymbol{z}$. As permutations of elements in the half vectorisation are irrelevant we can denote

$$\boldsymbol{y} = (\boldsymbol{d}^T, \boldsymbol{z}^T)^T$$
$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{I}_{\frac{m(m-1)}{2}} \end{bmatrix} \boldsymbol{z}.$$

The covariance of $\boldsymbol{y}$ is given as

$$\begin{aligned} \text{Cov}(\boldsymbol{y}) = \hat{\boldsymbol{\Sigma}}'' &= \begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{I}_{\frac{m(m-1)}{2}} \end{bmatrix} \text{Cov}(\boldsymbol{z}) \begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{I}_{\frac{m(m-1)}{2}} \end{bmatrix}^T \\ &= \begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{I}_{\frac{m(m-1)}{2}} \end{bmatrix} \hat{\boldsymbol{\Sigma}}' \begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{I}_{\frac{m(m-1)}{2}} \end{bmatrix}^T, \end{aligned}$$

this is not full rank. We can write $\begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{I}_{\frac{m(m-1)}{2}} \end{bmatrix} = \boldsymbol{W}\boldsymbol{D}\boldsymbol{V}^T$, by singular decomposition, where $\boldsymbol{W}$ and $\boldsymbol{V}$ are orthogonal, then set $\boldsymbol{P} = \boldsymbol{W}\boldsymbol{D}^{-1}\boldsymbol{V}^T$, where we know $\boldsymbol{D}^{-1}$ will exist as $\begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{I}_{\frac{m(m-1)}{2}} \end{bmatrix}$ has full column rank . Hence

$$\boldsymbol{I}_{\frac{m(m-1)}{2}} = \boldsymbol{P}^T \begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{I}_{\frac{m(m-1)}{2}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{I}_{\frac{m(m-1)}{2}} \end{bmatrix}^T \boldsymbol{P}.$$

Therefore the Moore-Penrose inverse for $\hat{\boldsymbol{\Sigma}}''$ is $\hat{\boldsymbol{\Sigma}}''^{-} = \boldsymbol{P}\hat{\boldsymbol{\Sigma}}'^{-1}\boldsymbol{P}^T$ as this satisfies the four conditions required to be a Moore-Penrose inverse in Penrose (1955). For example

the first condition is satisfied

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}''\hat{\boldsymbol{\Sigma}}''^{-}\hat{\boldsymbol{\Sigma}}'' &= \begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{I}_{\frac{m(m-1)}{2}} \end{bmatrix} \hat{\boldsymbol{\Sigma}}' \begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{I}_{\frac{m(m-1)}{2}} \end{bmatrix}^T \boldsymbol{P}\hat{\boldsymbol{\Sigma}}'^{-1}\boldsymbol{P}^T \begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{I}_{\frac{m(m-1)}{2}} \end{bmatrix} \hat{\boldsymbol{\Sigma}}' \begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{I}_{\frac{m(m-1)}{2}} \end{bmatrix}^T \\ &= \begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{I}_{\frac{m(m-1)}{2}} \end{bmatrix} \hat{\boldsymbol{\Sigma}}'\hat{\boldsymbol{\Sigma}}'^{-1}\hat{\boldsymbol{\Sigma}}' \begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{I}_{\frac{m(m-1)}{2}} \end{bmatrix}^T \\ &= \hat{\boldsymbol{\Sigma}}'',\end{aligned}$$

and similarly the other three hold

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}''^{-}\hat{\boldsymbol{\Sigma}}''\hat{\boldsymbol{\Sigma}}''^{-} &= \hat{\boldsymbol{\Sigma}}''^{-}, \\ (\hat{\boldsymbol{\Sigma}}''^{-}\hat{\boldsymbol{\Sigma}}'')^T &= \hat{\boldsymbol{\Sigma}}''^{-}\hat{\boldsymbol{\Sigma}}'', \\ (\hat{\boldsymbol{\Sigma}}''\hat{\boldsymbol{\Sigma}}''^{-})^T &= \hat{\boldsymbol{\Sigma}}''\hat{\boldsymbol{\Sigma}}''^{-}.\end{aligned}$$

Hence we get the result

$$\begin{aligned}T_G{}' &= \frac{n_A n_B}{n_A + n_B}\boldsymbol{y}\hat{\boldsymbol{\Sigma}}''^{-}\boldsymbol{y} \\ &= \frac{n_A n_B}{n_A + n_B}\boldsymbol{z}^T \begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{I}_{\frac{m(m-1)}{2}} \end{bmatrix}^T \hat{\boldsymbol{\Sigma}}''^{-} \begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{I}_{\frac{m(m-1)}{2}} \end{bmatrix} \boldsymbol{z} \\ &= \frac{n_A n_B}{n_A + n_B}\boldsymbol{z}^T \begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{I}_{\frac{m(m-1)}{2}} \end{bmatrix}^T \boldsymbol{P}\hat{\boldsymbol{\Sigma}}'^{-1}\boldsymbol{P}^T \begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{I}_{\frac{m(m-1)}{2}} \end{bmatrix} \boldsymbol{z} \\ &= \frac{n_A n_B}{n_A + n_B}\boldsymbol{z}^T\hat{\boldsymbol{\Sigma}}'^{-1}\boldsymbol{z} \\ &= T_G.\end{aligned}$$

□

4.8.2 Proof of the test statistic's asymptotic distribution when the covariance is isotropic

Eigenvalues of Q

Result 4.2.3. *The eigenvalues of* $\boldsymbol{Q}$ *are* $\{2m, \underbrace{m, \ldots, m}_{m-1 \text{ times}}, \underbrace{2, \ldots, 2}_{\frac{m(m-3)}{2} \text{ times}}\}$, *for* $m \geq 3$.

To prove the expressions for the eigenvalues of $\boldsymbol{Q}$ we firstly examine the eigenvalues of $\boldsymbol{Q}$ for different ms to see if there is a pattern to spot. For $m = 3$ the eigenvalues are $\{6, 3, 3\}$, for $m = 4$ the eigenvalues are $\{8, 4, 4, 4, 2, 2\}$ and for $m = 5$ the eigenvalues are $\{10, 5, 5, 5, 5, 2, 2, 2, 2, 2\}$. We have seen numerically that Result 4.2.3 is true for $m = 3, 4, 5$ and so now we will prove it for $m > 5$, hence in the following proofs the cases $m = 3, 4,$ and 5 are ignored as they are already shown to be true. First we will prove the three eigenvalues, 2, m and $2m$, are in fact eigenvalues by giving a corresponding eigenvector for each and then we will prove that we have the correct multiplicity of each eigenvalue. Many of the results stated within these proofs rely on visual observations from the visualisation of $\boldsymbol{Q}$ in (4.2.6). Throughout we denote $\boldsymbol{q}_r$ as the rth row of $\boldsymbol{Q}$.

Lemma 4.8.1. *$2m$ is an eigenvalue of $\boldsymbol{Q}$, for $m \geq 3$.*

Proof. For the eigenvalue $2m$ a corresponding eigenvector is $\boldsymbol{u}_1 = (1, \ldots, 1)^T$. This can be seen easily as in every row there is a four and the number of ones is $(m-2) + (m-2) = 2m - 4$, so the row sum will be $2m$ for every row. □

To find the other two eigenvalues it is useful to note there are three cases for the rth row

- $r = r_1 = 1$ where $q_{r_1 1} = 4$, this is the first row and is always of the form $(4, \underbrace{1, \ldots, 1}_{2(m-2) times}, \underbrace{0, \ldots, 0}_{\frac{(m-3)(m-2)}{2} times})$,
- $r = r_2$ where $1 < r_2 \leq (m-1) + (m-2) = 2m - 3$. $q_{r_2 1} = 1$, in $q_{r_2\, 2:2m-3}$ there is a 4, $m-2$ lots of 1s and $m-3$ lots of 0s, and in $q_{r_2\, 2m-2:\frac{m(m-1)}{2}}$ there are $m-3$ lots of 1s and $\frac{(m-3)(m-4)}{2}$ lots of 0s.
- $r = r_3$ where $2m - 3 < r_3 \leq \frac{m(m-1)}{2}$. Similarly $q_{r_3 1} = 0$, in $q_{r_3\, 2:2m-3}$ there are 4 lots of 1s and $2m-8$ lots of 0s, and in $q_{r_3\, 2m-2:\frac{m(m-1)}{2}}$ there is a 4 and there are $2m-8$ lots of 1s and $\frac{(m-5)(m-4)}{2}$ lots of 0s.

We use these cases to find a corresponding eigenvector to the eigenvalues m and 2.

Lemma 4.8.2. *m is an eigenvalue of $\boldsymbol{Q}$, for $m \geq 3$.*

Proof. For the eigenvalue m a corresponding eigenvector is $\boldsymbol{f}_1 = (1, \underbrace{\frac{m-4}{2(m-2)}, \ldots, \frac{m-4}{2(m-2)}}_{2(m-2) times}, \underbrace{\frac{-2}{m-2}, \ldots, \frac{-2}{m-2}}_{\frac{(m-3)(m-2)}{2} times})^T$. We use the cases to show

this is an eigenvector. First call $\boldsymbol{Q}\boldsymbol{f}_1 = \{\boldsymbol{q}_1\boldsymbol{f}_1, \ldots, \boldsymbol{q}_{\frac{m(m-1)}{2}}\boldsymbol{f}_1\}^T$, then

$$\begin{aligned}
\boldsymbol{q}_{r_1}\boldsymbol{f}_1 &= 4 + \frac{2(m-2)(m-4)}{2(m-2)} + 0, \\
&= m, \\
\boldsymbol{q}_{r_2}\boldsymbol{f}_1 &= 1 + \frac{(4+m-2)(m-4)}{2(m-2)} - \frac{2(m-3)}{m-2}, \\
&= m\frac{m-4}{2(m-2)}, \\
\boldsymbol{q}_{r_3}\boldsymbol{f}_1 &= 0 + \frac{4(m-4)}{2(m-2)} - \frac{2(4+2m-8)}{m-2}, \\
&= m\frac{-2}{m-2},
\end{aligned}$$

for $r_1 = 1$, $1 < r_2 \leq 2m-3$ and $2m-3 < r_3 \leq \frac{m(m-1)}{2}$. Clearly $\boldsymbol{Q}\boldsymbol{f}_1 = m\boldsymbol{f}_1$, and so $\boldsymbol{f}_1$ is an eigenvector for the eigenvalue m. □

Lemma 4.8.3. *2 is an eigenvalue of $\boldsymbol{Q}$, for $m \geq 4$.*

Proof. A corresponding eigenvector to the eigenvalue 2 is $\boldsymbol{w}_1 = (1, \underbrace{\frac{-1}{m-2}, \ldots, \frac{-1}{m-2}}_{2(m-2)times}, \underbrace{\frac{2}{(m-2)(m-3)}, \ldots, \frac{2}{(m-2)(m-3)}}_{\frac{(m-3)(m-2)}{2}times})^T$. We use the cases to show this is an eigenvector, first call $\boldsymbol{Q}\boldsymbol{w}_1 = \{\boldsymbol{q}_1\boldsymbol{w}_1, \ldots, \boldsymbol{q}_{\frac{m(m-1)}{2}}\boldsymbol{w}_1\}^T$, then

$$\begin{aligned}
\boldsymbol{q}_{r_1}\boldsymbol{w}_1 &= 4 - \frac{2(m-2)}{m-2} + 0, \\
&= 2, \\
\boldsymbol{q}_{r_2}\boldsymbol{w}_1 &= 1 - \frac{4+m-2}{m-2} + \frac{2(m-3)}{(m-2)(m-3)}, \\
&= 2\frac{-1}{m-2}, \\
\boldsymbol{q}_{r_3}\boldsymbol{w}_1 &= 0 - \frac{4}{m-2} + \frac{(4+2m-8)2}{(m-2)(m-3)}, \\
&= 2\frac{2}{(m-2)(m-3)},
\end{aligned}$$

for $r_1 = 1$, $1 < r_2 \leq 2m-3$ and $2m-3 < r_3 \leq \frac{m(m-1)}{2}$. Clearly $\boldsymbol{Q}\boldsymbol{w}_1 = 2\boldsymbol{w}_1$ meaning $\boldsymbol{w}_1$ is an eigenvector for the eigenvalue 2. □

We now prove the multiplicity of each eigenvalue is as stated. We define the multiplicity of the eigenvalues $2m$, m and 2 as γ_{2m}, γ_m and γ_2 respectively.

First look at the eigenvector found for m, due to the invariance of permutation of rows and columns of $\boldsymbol{Q}$ the values in the eigenvector $\boldsymbol{f}_1$ can be rearranged in a specific way and still be an eigenvector. If we create a matrix $\boldsymbol{F}$ where $\boldsymbol{F}[\boldsymbol{Q} = 4] = 1$, $\boldsymbol{F}[\boldsymbol{Q} = 1] = \frac{m-4}{2(m-2)}$ and $\boldsymbol{F}[\boldsymbol{Q} = 0] = \frac{-2}{m-2}$, then the first row is $\boldsymbol{f}_1^T$, and every other row is also an eigenvector, by the same logic. We define the lth row of $\boldsymbol{F}$ as $\boldsymbol{f}_l^T = (f_{l1}, f_{l2}, \ldots, f_{l\frac{m(m-1)}{2}})$. We prove the following lemma for the multiplicity of the eigenvalue m,

Lemma 4.8.4. *The multiplicity $\gamma_m \geq m-1$, for $m \geq 3$.*

Proof. We suppose for contradiction that the first $m-1$ eigenvectors, hence rows of $\boldsymbol{F}$, are not linearly independent, and so there exists $\theta_1, \ldots, \theta_{m-2} \in \mathcal{R}$ such that

$$\theta_1 \boldsymbol{f}_1^T + \cdots + \theta_{m-2} \boldsymbol{f}_{m-2}^T = \boldsymbol{f}_{m-1}^T. \tag{4.8.2}$$

It can be seen easily that for q_{rs} with $1 \leq r, s \leq m-1$ that $i = k = 1$ from Equations (4.2.7), meaning

$$q_{rs} = \begin{cases} 4, & \text{if } r = s \\ 1, & \text{if } r \neq s \end{cases} \tag{4.8.3}$$

$$f_{rs} = \begin{cases} 1, & \text{if } r = s \\ \frac{m-4}{2(m-2)}, & \text{if } r \neq s. \end{cases} \tag{4.8.4}$$

Equation (4.8.2) can be split to look at it component-wise, for $1 \leq t \leq \frac{m(m-1)}{2}$,

$$\theta_1 f_{1t} + \cdots + \theta_{m-2} f_{m-2t} = f_{m-1t}, \tag{4.8.5}$$

$$\theta_t f_{tt} + \sum_{l \neq t} \theta_l f_{lt} = f_{m-1t}. \tag{4.8.6}$$

Substituting from Equation (4.8.4) into Equation (4.8.6) for $1 \leq t \leq m-2$ gives

$$\theta_t + \sum_{l \neq t} \theta_l \frac{m-4}{2(m-2)} = \frac{m-4}{2(m-2)}. \tag{4.8.7}$$

If we subtract Equation (4.8.7) with $t = 2$ from this with $t = 1$ we get,

$$(1 - \frac{m-4}{2(m-2)})(\theta_1 - \theta_2) = 0,$$
$$\theta_1 = \theta_2,$$

this can be repeated for all $1 \leq t \leq m-2$, giving $\theta_1 = \theta_2 = \cdots = \theta_{m-2}$. Looking at Equation (4.8.6) for $t = m-1$ and substituting the result for the θs we get,

$$\frac{\theta_1(m-2)(m-4)}{2(m-2)} = 1,$$
$$\theta_1 = \frac{2}{m-4},$$

this is then substituted into Equation (4.8.7),

$$\frac{2}{m-4} + \frac{2(m-3)(m-4}{2(m-2)(m-4)} = \frac{m-4}{2(m-2)},$$
$$\frac{2}{m-4}(\frac{2(m-2) + (m-4)(m-3)}{2(m-2} = \frac{m-4}{2(m-2)},$$
$$m(m-2) = 0,$$
$$m = 0 \text{ or } 2,$$

this is a contradiction as we have been looking at $m > 5$, so the first $m-1$ rows in $\boldsymbol{F}$ are linearly independent eigenvectors, and so the multiplicity of m is at least $m-1$. □

Now looking at the eigenvalue 2, we can rearrange the eigenvector $\boldsymbol{w}_1$ to give more eigenvectors. We again create a matrix $\boldsymbol{W}$ where $\boldsymbol{W}[\boldsymbol{Q} = 4] = 1$, $\boldsymbol{W}[\boldsymbol{Q} = 1] = \frac{-1}{m-2}$ and $\boldsymbol{W}[\boldsymbol{Q} = 0] = \frac{2}{(m-2)(m-3)}$, the first row is $\boldsymbol{w}_1^T$, and every other row is also an eigenvector, by the same logic. We define the lth row of $\boldsymbol{W}$ as $\boldsymbol{w}_l^T = (w_{l1}, w_{l2}, \ldots, w_{l\frac{m(m-1)}{2}})$. We prove the following lemma for the multiplicity of the eigenvalue 2,

Lemma 4.8.5. *The multiplicity* $\gamma_2 \geq \frac{m(m-3)}{2}$*, for* $m \geq 3$*.*

Proof. Suppose for contradiction that the last $\frac{m(m-3)}{2}$ eigenvectors, hence rows of $\boldsymbol{W}$, are not linearly independent, so there exists $\kappa_{m+2}, \ldots, \kappa_{\frac{m(m-1)}{2}} \in \mathcal{R}$ such that

$$\kappa_{m+2}\boldsymbol{w}_{m+2}^T + \cdots + \kappa_{\frac{m(m-1)}{2}}\boldsymbol{w}_{\frac{m(m-1)}{2}}^T = \boldsymbol{w}_{m+1}^T, \tag{4.8.8}$$

component-wise this gives, for $1 \leq l \leq \frac{m(m-1)}{2}$

$$\kappa_{m+2} w_{m+2l} + \cdots + \kappa_{\frac{m(m-1)}{2}} w_{\frac{m(m-1)}{2}l} = w_{m+1l}. \tag{4.8.9}$$

We look at the q_{rs} values as they directly correspond to w_{rs} values. For $r \geq m+2$ q_{rs} is the coefficient for $\delta_{ij}\delta_{kp}$ where $\delta_{ij} \in \{\delta_{25}, \ldots, \delta_{2m}, \delta_{34}, \ldots, \delta_{m-1m}\}$. Note

$$\begin{aligned} q_{r,2} &\text{ is the coefficient of } \delta_{ij}\delta_{13}, \\ q_{r,3} &\text{ is the coefficient of } \delta_{ij}\delta_{14}, \\ q_{r,m} &\text{ is the coefficient of } \delta_{ij}\delta_{23}, \\ q_{r,m+1} &\text{ is the coefficient of } \delta_{ij}\delta_{24}, \end{aligned}$$

from this we see, for $r \geq m+2$ q_{rs},

$$\begin{aligned} w_{r2} = \frac{-1}{m-2} &\text{ iff } q_{r2} = 1 \\ &\text{iff } \delta_{ij} \in \{\delta_{34}, \ldots, \delta_{3m}\}, \text{ this corresponds to} \\ &\quad r \in \mathcal{R}_1 = \{r_{34}, \ldots, r_{3m}\}. \\ w_{r3} = \frac{-1}{m-2} &\text{ iff } q_{r3} = 1 \\ &\text{iff } \delta_{ij} \in \{\delta_{34}, \delta_{45}, \ldots, \delta_{4m}\}, \text{ this corresponds to} \\ &\quad r \in \mathcal{R}_2 = \{r_{34}, r_{45}, \ldots, r_{4m}\}. \\ w_{rm} = \frac{-1}{m-2} &\text{ iff } q_{rm} = 1 \\ &\text{iff } \delta_{ij} \in \{\delta_{25}, \ldots, \delta_{2m}, \delta_{34}, \ldots, \delta_{3m}\}, \text{ this corresponds to} \\ &\quad r \in \mathcal{R}_3 = \{r_{25}, \ldots, r_{2m}, r_{34}, \ldots, r_{3m}\}. \\ w_{r,m+1} = \frac{-1}{m-2} &\text{ iff } q_{r,m+1} = 1 \\ &\text{iff } \delta_{ij} \in \{\delta_{25}, \ldots, \delta_{2m}, \delta_{34}, \delta_{45}, \ldots, \delta_{4m}\}, \text{ this corresponds to} \\ &\quad r \in \mathcal{R}_4 = \{r_{25}, \ldots, r_{2m}, r_{34}, r_{45} \ldots, r_{4m}\}. \end{aligned}$$

The set difference of $\mathcal{R}_1$ and $\mathcal{R}_2$ is equal to that for $\mathcal{R}_3$ and $\mathcal{R}_4$. Using this and the fact that

$w_{r2}, wr_{r3}, w_{rm}, w_{rm+1} = \frac{-1}{m-2}$ or $\frac{2}{(m-2)(m-3)}$, we see for $r \geq m+2$

$$w_{r2} \neq w_{r3} \text{ iff } w_{rm} \neq w_{rm+1}.$$

The specific equations from Equation (4.8.9) that we use are

$$\kappa_{m+2} w_{m+2\,2} + \cdots + \kappa_{\frac{m(m-1)}{2}} w_{\frac{m(m-1)}{2}2} = w_{m+1\,2}, \tag{4.8.10}$$

$$\kappa_{m+2} w_{m+2\,3} + \cdots + \kappa_{\frac{m(m-1)}{2}} w_{\frac{m(m-1)}{2}3} = w_{m+1\,3}, \tag{4.8.11}$$

$$\kappa_{m+2} w_{m+2\,m} + \cdots + \kappa_{\frac{m(m-1)}{2}} w_{\frac{m(m-1)}{2}m} = w_{m+1\,m}, \tag{4.8.12}$$

$$\kappa_{m+2} w_{m+2\,m+1} + \cdots + \kappa_{\frac{m(m-1)}{2}} w_{\frac{m(m-1)}{2}m+1} = w_{m+1\,m+1}. \tag{4.8.13}$$

Due to the equal set difference it is clear to see the left hand side of Equation (4.8.10) subtract Equation (4.8.11) equals the left hand side of Equation (4.8.12) subtract Equation (4.8.13), therefore their right hands must be equal giving

$$w_{m+1\,2} - w_{m+1\,3} = w_{m+1\,m} - w_{m+1\,m+1}. \tag{4.8.14}$$

We find the value of each term in this equation,

$$w_{m+1\,2} = \frac{2}{(m-2)(m-3)} \text{ as } q_{m+1\,2} \text{ is the coefficient of } \delta_{24}\delta_{13} \text{ which is } 0,$$
$$w_{m+1\,3} = \frac{-1}{m-2} \text{ as } q_{m+1\,3} \text{ is the coefficient of } \delta_{24}\delta_{14} \text{ which is } 1,$$
$$w_{m+1\,m} = \frac{-1}{m-2} \text{ as } q_{m+1\,m} \text{ is the coefficient of } \delta_{24}\delta_{23} \text{ which is } 1,$$
$$w_{m+1\,m+1} = 1 \text{ as } q_{m+1\,m+1} = 4,$$

substituting this into Equation (4.8.14) gives,

$$\begin{aligned} \frac{2}{(m-2)(m-3)} + \frac{1}{m-2} &= \frac{-1}{m-2} - 1 \\ \frac{m-1}{m-3} &= 1 - m \\ (m-2)(m-1) &= 0 \\ m &= 1 \text{ or } 2, \end{aligned}$$

this is a contradiction as we are looking at $m > 5$ so the last $\frac{m(m-3)}{2}$ rows of $\boldsymbol{W}$ are linearly independent eigenvectors, so the multiplicity of 2 is at least $\frac{m(m-3)}{2}$. □

We are now able to prove the multiplicity of each eigenvalue.

Lemma 4.8.6. *The multiplicities are given by* $\gamma_{2m} = 1$, $\gamma_m = m - 1$ *and* $\gamma_2 = \frac{m(m-3)}{2}$.

Proof. The sum of the multiplicity cannot exceed the size of the matrix $\boldsymbol{Q}$, so $\gamma_{2m} + \gamma_m + \gamma_2 \leq \frac{m(m-1)}{2}$. For contradiction suppose $\gamma_{2m} \neq 1$ or $\gamma_m \neq m-1$ or $\gamma_2 \neq \frac{m(m-3)}{2}$, using Lemma 4.8.5 and 4.8.4 this means $\gamma_{2m} > 1$ or $\gamma_m > m-1$ or $\gamma_2 > \frac{m(m-3)}{2}$, leading to

$$\begin{aligned}\gamma_{2m} + \gamma_m + \gamma_2 &> 1 + m - 1 + \frac{m(m-3)}{2} \\ &> \frac{m(m-1)}{2}.\end{aligned}$$

Equating our two results gives,

$$\frac{m(m-1)}{2} < \gamma_{2m} + \gamma_m + \gamma_2 \leq \frac{m(m-1)}{2}$$

this is clearly a contradiction and so $\gamma_{2m} = 1$, $\gamma_m = m-1$ and $\gamma_2 = \frac{m(m-3)}{2}$. □

We have now proved Result 4.2.3, by showing $2m, m$ and 2 are eigenvalues of $\boldsymbol{Q}$ with multiplicity $1, m-1$ and $\frac{m(m-3)}{2}$ respectively, and as their multiplicities sum to the dimension of $\boldsymbol{Q}$ no other eigenvalues exist.

4.8.3 Distribution of T_E under H_0 when the covariance is isotropic

Under the null hypothesis, $\boldsymbol{\Sigma} = \sigma_0^2 \boldsymbol{I}_{\frac{m(m-1)}{2}}$. We know the distribution of $d(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})^2$, from Result 4.2.2, is $\frac{n_A+n_B}{n_A n_B} \sum_{i=1}^{l} \lambda_i \chi_1^2$, where λ_i are the l non-zero eigenvalues of $\boldsymbol{\Sigma Q}$. We see $\boldsymbol{\Sigma Q} = \sigma_0^2 \boldsymbol{Q}$, and so the eigenvalues of this are $\lambda_i = \sigma_0^2 \lambda_i^q$, where the λ_i^q's are the eigenvalues of $\boldsymbol{Q}$. From Result 4.2.3 we know the eigenvalues of $\boldsymbol{Q}$ and so we can substitute in our results, leading to,

$$\begin{aligned}d_1(\hat{\boldsymbol{A}}_E, \hat{\boldsymbol{B}}_E)^2 \dot{\sim} & 2m \frac{n_A + n_B}{n_A n_B} \sigma_0^2 \chi_1^2 \\ & + \underbrace{m \frac{n_A + n_B}{n_A n_B} \sigma_0^2 \chi_1^2 + \cdots + m \frac{n_A + n_B}{n_A n_B} \sigma_0 \chi_1^2}_{m-1 \text{times}} \\ & + \underbrace{2 \frac{n_A + n_B}{n_A n_B} \sigma_0^2 \chi_1^2 + .. + 2 \frac{n_A + n_B}{n_A n_B} \sigma_0^2 \chi_1^2}_{\frac{m(m-3)}{2} \text{ times}},\end{aligned}$$

as the chi-squared terms are independent this simplifies to Result 4.2.4.

4.8.4 Proof of the distribution of graph Laplacians from a stochastic block model

Proof. To calculate the distribution of $\boldsymbol{x}$ under the stochastic block model we first calculate the distribution of the elements in the mean graph Laplacians, $\hat{\boldsymbol{A}}_E$ and $\hat{\boldsymbol{B}}_E$, by using the distribution of the elements in a graph Laplacian in $\mathcal{A}$ and $\mathcal{B}$. From the definition of a Stochastic block model network we know, for the graph Laplacian $\boldsymbol{A}_k = (a_{ijk})$,

$$a_{ijk} \sim -\mathcal{B}(1, p_{ij}^A) \text{ for } i \neq j,$$

where $\mathcal{B}$ denotes the binomial distribution.

The Euclidean mean of the graph Laplacians in set $\mathcal{A}$ is $\hat{\boldsymbol{A}}_E = \frac{1}{n_A}\sum_{k=1}^{n_A} \boldsymbol{A}_k = (\hat{a}_{ij})$ therefore

$$\hat{a}_{ij} = \frac{1}{n_A}\sum_{k=1}^{n_A} a_{ijk}.$$

Now we can see as as $n_A \to \infty$ for $i \neq j$

$$\begin{aligned} n_A\hat{a}_{ij} = \sum_{k=1}^{n_A} a_{ijk} &\sim -\mathcal{B}(n_A, p_{ij}^A), \\ \frac{\sum_{k=1}^{n_A} a_{ijk}}{\sqrt{n_A}} &\sim \mathcal{N}\Big(-\sqrt{n_A}p_{ij}^A, p_A(1-p_{ij}^A)\Big), \\ \text{so } \hat{a}_{ij} &\sim \mathcal{N}\Big(-p_{ij}^A, \frac{1}{n_A}p_{ij}^A(1-p_{ij}^A)\Big). \end{aligned}$$

Similarly as $n_B \to \infty$,

$$\hat{b}_{ij} \sim \mathcal{N}\Big(-p_{ij}^B, \frac{1}{n_B}p_{ij}^B(1-p_{ij}^B)\Big) \text{ for } i \neq j.$$

These binomial to normal approximations hold when $n_A(m-1)p_{ij}^A(1-p_{ij}^A)$, $n_B(m-1)p_{ij}^B(1-p_{ij}^B)$, $n_Ap_{ij}^A(1-p_{ij}^A)$ and $n_Bp_{ij}^B(1-p_{ij}^B)$ tend to infinity, which they do unless p_{ij}^A or p_{ij}^B equals 0 (Molenaar (1970)). As

$$\boldsymbol{x} = \left(\frac{n_An_B}{n_A+n_B}\right)^{\frac{1}{2}}\left(\hat{a}_{12} - \hat{b}_{12}, \ldots, \hat{a}_{m-1\,m} - \hat{b}_{m-1\,m}\right),$$

it is clear

$$\boldsymbol{x} \dot{\sim} \mathcal{N}\left(\left(\frac{n_A n_B}{n_A + n_B}\right)^{\frac{1}{2}} \boldsymbol{\mu}', \boldsymbol{\Sigma}\right)$$

where $\boldsymbol{\mu}' = (p^B_{12} - p^A_{12}, \ldots, p^B_{m-1\,m} - p^A_{m-1\,m})^T$,

and $\boldsymbol{\Sigma} = \text{diag}\left(\frac{n_B\, p^A_{12}\,(1 - p^A_{12}) + n_A\, p^B_{12}\,(1 - p^B_{12})}{n_A + n_B}, \ldots \right.$

$$\left., \frac{n_B\, p^A_{m-1\,m}\,(1 - p^A_{m-1\,m}) + n_A\, p^B_{m-1\,m}\,(1 - p^B_{m-1\,m})}{n_A + n_B}\right).$$

□

4.8.5 Distribution of T_E under H_1 for Erdös-Renyi model network samples

Under the alternative hypothesis $p_A \neq p_B$. We already know

$$\boldsymbol{x} \dot{\sim} \mathcal{N}\left(\left(\frac{n_A n_B}{n_A + n_B}\right)^{\frac{1}{2}} \boldsymbol{\mu}', \boldsymbol{\Sigma}\right),$$

were $\boldsymbol{\mu}'$ and $\boldsymbol{\Sigma}$ are defined in Equations (4.2.14) and (4.2.15). We shall now define $\boldsymbol{\mu}^* = \left(\frac{n_A n_B}{n_A + n_B}\right)^{\frac{1}{2}} \boldsymbol{\mu}'$. We can now prove the distribution of the Euclidean distance squared under the alternative hypothesis

First it should be noted from Imhof (1961) we are expecting this distribution to be the sum of non-central chi-square random variables. We have seen previously $d_1(\hat{\boldsymbol{A}}_E, \hat{\boldsymbol{B}}_E)^2 = \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x}$, as $\boldsymbol{Q}$ is symmetric it can be decomposed giving,

$$\boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} = \boldsymbol{x}^T \boldsymbol{U}^T \boldsymbol{\Lambda} \boldsymbol{U} \boldsymbol{x},$$

where $\boldsymbol{U}$ has rows, $\boldsymbol{u}_i^T$, as orthonormal eigenvectors of $\boldsymbol{Q}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix with the eigenvalues of $\boldsymbol{Q}$, λ_i^q, on the diagonal. We look at the distributions of $\boldsymbol{U}\boldsymbol{x} = \left(\boldsymbol{u}_1^T \boldsymbol{x}, \boldsymbol{u}_2^T \boldsymbol{x}, \ldots\right)^T$,

$$\begin{aligned}\boldsymbol{U}\boldsymbol{x} &\dot{\sim} \mathcal{N}\left(\left(\boldsymbol{u}_1^T \boldsymbol{\mu}^*, \boldsymbol{u}_2^T \boldsymbol{\mu}^* \ldots\right)^T, \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^T\right), \\ &\dot{\sim} \mathcal{N}\left(\left(\boldsymbol{u}_1^T \boldsymbol{\mu}^*, \boldsymbol{u}_2^T \boldsymbol{\mu}^* \ldots\right)^T, \boldsymbol{\Sigma}\right),\end{aligned}$$

the covariance becomes $\boldsymbol{\Sigma}$ due to being a scaled identity matrix, $\boldsymbol{\Sigma} = \sigma_1^2 \boldsymbol{I}_{\frac{m(m-1)}{2}}$, where $\sigma_1^2 = \frac{n_B\, p_A\,(1-p_A)+n_A\, p_B\,(1-p_B)}{n_A+n_B}$, and the fact $\boldsymbol{U}\boldsymbol{U}^T$. This gives the elements the distribution

$$\begin{aligned}\boldsymbol{u}_i^T\boldsymbol{x} &\dot{\sim} \mathcal{N}\left(\boldsymbol{u}_i^T\boldsymbol{\mu}^*, \sigma_1^2\right),\\ \frac{\boldsymbol{u}_i^T\boldsymbol{x}}{\sigma_1} &\dot{\sim} \mathcal{N}\left(\frac{\boldsymbol{u}_i^T\boldsymbol{\mu}^*}{\sigma_1}, 1\right).\end{aligned}$$

We can now write

$$\begin{aligned} d_1(\hat{\boldsymbol{A}}_E, \hat{\boldsymbol{B}}_E)^2 &= \left(\frac{n_A n_B}{n_A+n_B}\right)^{-1} \boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} && (4.8.15)\\ &= \left(\frac{n_A n_B}{n_A+n_B}\right)^{-1} \sum_{i=1}^{\frac{m(m-1)}{2}} \lambda_i^q (\boldsymbol{u}_i^T\boldsymbol{x})^2,\\ &= \left(\frac{n_A n_B}{n_A+n_B}\right)^{-1} \sum_{i=1}^{\frac{m(m-1)}{2}} \lambda_i^q\, \sigma_1^2\, \chi_1^2\left(\left(\frac{\boldsymbol{u}_i^T\boldsymbol{\mu}^*}{\sigma_1}\right)^2\right), && (4.8.16)\end{aligned}$$

using the definition for the non-central chi-square distribution on page 412 of Scheffe (1999). The terms in this sum are all independent. We already know the eigenvalues λ_i^q from Result 4.2.3, and saw the first eigenvector corresponding with eigenvalue $2m$ was $(1,\ldots,1)^T$. We want an orthonormal set of eigenvectors and so $\boldsymbol{u}_1^T = \sqrt{\frac{2}{m(m-1)}}(1,\ldots,1)^T$. As all the eigenvectors in the set must be orthogonal we have $\boldsymbol{u}_1^T\boldsymbol{u}_i^T = 0$ for $i \neq 1$. Now we note $\boldsymbol{\mu}^* = \left(\frac{n_A n_B}{n_A+n_B}\right)^{\frac{1}{2}}(p_B - p_A)\sqrt{\frac{m(m-1)}{2}}\boldsymbol{u}_1$ and so

$$\begin{aligned}\boldsymbol{u}_i^T\boldsymbol{\mu}^* &= \left(\frac{n_A n_B}{n_A+n_B}\right)^{\frac{1}{2}}(p_B - p_A)\sqrt{\frac{m(m-1)}{2}}\boldsymbol{u}_i^T\boldsymbol{u}_1,\\ \left(\boldsymbol{u}_i^T\boldsymbol{\mu}'\right)^2 &= \begin{cases}\left(\frac{n_A n_B}{n_A+n_B}\right)(p_B - p_A)^2\left(\frac{m(m-1)}{2}\right) & \text{if } i = 1,\\ 0 & \text{if } i \neq 1.\end{cases}\end{aligned}$$

Substituting these results into Equation (4.8.15) gives

$$\begin{aligned} d_1(\hat{\boldsymbol{A}}_E, \hat{\boldsymbol{B}}_E)^2 \dot{\sim} & 2m \left(\frac{n_A n_B}{n_A + n_B}\right)^{-1} \sigma_1^2 \chi_1^2 \left(\left(\frac{n_A n_B}{n_A + n_B}\right)(p_B - p_A)^2 \frac{m(m-1)}{2\sigma_1^2}\right) \\ & + \underbrace{m \left(\frac{n_A n_B}{n_A + n_B}\right)^{-1} \sigma_1^2 \chi_1^2 + \cdots + m \left(\frac{n_A n_B}{n_A + n_B}\right)^{-1} \sigma_1^2 \chi_1^2}_{m-1 \text{times}} \\ & + \underbrace{2 \left(\frac{n_A n_B}{n_A + n_B}\right)^{-1} \sigma_1^2 \chi_1^2 + .. + 2 \left(\frac{n_A n_B}{n_A + n_B}\right)^{-1} \sigma_1^2 \chi_1^2}_{\frac{m(m-3)}{2} \text{ times}}, \end{aligned}$$

the chi-square random variables are independent and so this is simplified to Result 4.2.6.

CHAPTER 5

Classification and anomaly detection

5.1 Classification

In the previous chapter we have seen many examples where networks can belong to different classes, such as the dataset of networks representing novels belonging to the class of novels written by Austen or by Dickens. The class a network belongs to may be unknown and in these cases it is useful to be able to classify the graph Laplacian to determine which class it belongs to. In this chapter we will provide two novel methods of classifying graph Laplacians, one will be performed in the embedding space and the other in the space of PC scores.

For classification, graph Laplacians representing networks must belong to a class out of C possible classes. We will only consider a binary classification problem, meaning $C = 2$ and we will refer to the classes as '1' and '0'. The classification methods are supervised methods so require a training set of graph Laplacians where the classes are already known, described in Section 1.2.4. The training set can be thought of as two sets, one from each class denoted $\mathcal{A} = \{\boldsymbol{L}_1^1, ..., \boldsymbol{L}_{n_1}^1\}$ and $\mathcal{B} = \{\boldsymbol{L}_1^0, ..., \boldsymbol{L}_{n_0}^0\}$. Each classification method will output probabilities a graph Laplacian belongs to each class. For the binary classes, '0' and '1', we will choose to predict p_i^1, which is the probability the graph Laplacian $\boldsymbol{L}_i$ belong to class 1, we shall write this as $p_i^1 = p^1(\boldsymbol{L}_i)$. As there are only two classes then the probability of being in class '0' can be found easily as $p_i^0 = 1 - p_i^1$. These probabilities can be converted to a classification rule, a natural one being to classify a graph Laplacian as belonging to a certain class if the probability it belongs to this class is over 0.5.

At the end of this chapter we also provide a method of classifying if a graph Laplacian

represents an anomaly in a dataset. This is an unsupervised method as we want to be able to detect an anomaly with no prior knowledge of a dataset.

5.1.1 Method 1: Classification in the manifold

The classification method we describe now takes place in the embedding space. This method is very similar to the regression method we described in Section 3.3 with it too using the Nadaraya-Watson model, described in Section 1.2.4, however instead of predicting a Euclidean response based on graph Laplacian predictors we now are predicting the probability a graph Laplacian belongs to a certain class.

To use the Nadaraya-Watson model for classification we need the probabilities each graph Laplacian in our training sample is in each class. As we know the class each of these graph Laplacians are in then we will set the probability as

$$p^1(\boldsymbol{L}_i) \begin{cases} 1 & \text{if } \boldsymbol{L}_i \text{ is in class } 1 \\ 0 & \text{if } \boldsymbol{L}_i \text{ is in class } 0. \end{cases}$$

When setting the probabilities we are assuming that if two graph Laplacians are identical they will belong to the same class, i.e. if $L_i = L_j$ and L_i belongs to class '0' then L_j must belong to class '0' too, this seems like a reasonable assumption. Just like the Nadaraya-Watson estimate for regression in Equation (3.3.1), our estimate for probability is a linear combination of the training set probabilities. The Nadaraya-Watson estimator for the probability $\boldsymbol{L} \in \mathcal{L}_m$ belongs to class 1 becomes

$$\hat{p}^1(\boldsymbol{L}) = \frac{\sum_{i=1}^{n_0} K_h(d(\boldsymbol{L}, \boldsymbol{L}_i^0)) \times 0 + \sum_{i=1}^{n_1} K_h(d(\boldsymbol{L}, \boldsymbol{L}_i^1)) \times 1}{\sum_{i=1}^{n_0} K_h(d(\boldsymbol{L}, \boldsymbol{L}_i^0)) + \sum_{i=1}^{n1} K_h(d(\boldsymbol{L}, \boldsymbol{L}_i^1))},$$

where d can be any metric between two graph Laplacians, including the Euclidean power and Procrustes power metrics and K_h is the kernel function with bandwidth h. Just as in Section 3.2 a common kernel and the one we shall choose in the Gaussian kernel defined in (3.2.2). We see $\hat{p}^1$ is guaranteed to represent a probability itself in Result 5.1.1.

Result 5.1.1. $0 \leq \hat{p}^1 \leq 1$.

Proof. As $K_h(\cdot) \geq 0$ clearly $\hat{p}^1 \geq 0$ and

$$\begin{aligned}\sum_{i=1}^{n_0} K_h(d(\boldsymbol{L}, \boldsymbol{L}_i^0)) \times 0 + \sum_{i=1}^{n_1} K_h(d(\boldsymbol{L}, \boldsymbol{L}_i^1)) \times 1 &= \sum_{i=1}^{n_1} K_h(d(\boldsymbol{L}, \boldsymbol{L}_i^1)) \\ &\leq \sum_{i=1}^{n_0} K_h(d(\boldsymbol{L}, \boldsymbol{L}_i^0)) + \sum_{i=1}^{n_1} K_h(d(\boldsymbol{L}, \boldsymbol{L}_i^1)) \\ \Rightarrow \hat{p}^1(\boldsymbol{L}) &\leq 1.\end{aligned}$$

□

In this method of classification the bandwidth, h is a parameter that needs to be chosen. This parameter can be optimised, by repeating the classification for different h and choosing the h that performed best, often chosen by the accuracy produced defined in (1.2.10). This optimisation of h is similar to the optimisation of the bandwidth when the Nadaraya-Watson model was used for regression in Example 3.3.1. We shall not look into optimising h in our examples and instead choose the bandwidth as a quarter of the mean distance between every graph Laplacian in the sample as we feel this is a sensible bandwidth that gives good results.

5.1.2 Method 2: Classification in the space of PC scores

An alternative method of classification of graph Laplacians we consider is classification within a linear space. One option of a linear space to use is the graph Laplacian's tangent space for a specific metric. We choose not to perform the classification in the tangent space as this has a very large number of dimensions. Instead we perform classification in a reduced dimensional space offered by the PC scores, defined in Section 2.5. This method of classification is similar to Wang et al. (2017) which proposes a joint embedding of multiple undirected graphs for this purpose.

With the PC scores we can then use standard supervised classification methods to classify the graph Laplacians. The three different classification methods we will consider when using the PC scores are linear discriminant analysis (LDA), random forests and support vector machines (SVM) described in Section 1.2.4. Whilst we will only consider these three classification methods in the PCA space the general method we have described will hold to numerous other standard classification methods such as gradient boosting machines and Naive Bayes.

5.1.3 Application of classification methods to network data

As we explained in Section 1.2.4 we shall use cross validation to evaluate the classification methods and we shall use leave one out cross validation where the algorithm is trained on all the data except one graph Laplacian. The trained algorithm will then predict a probability for the left out graph Laplacian, this is repeated until every graph Laplacian has a prediction for it. For a sample of n graph Laplacians the leave one out method requires the algorithm to be trained and tested for each of the graph Laplacians, so the algorithm will be trained and tested a total of n times.

For the classification of PC scores the leave one out method requires the PC scores to be calculated for each training set and so for a sample of n graph Laplacians PCA will be performed n times. Therefore we only use the Euclidean power metrics and specifically $\alpha = 1$ and $\frac{1}{2}$, as the Procrustes power metric is very time consuming to get the PC scores for each training set and then project other graph Laplacians into this space.

Example 5.1.1: Classification methods applied to the novel data
We shall compare our methods on the 19th century authors dataset to demonstrate how we can classify text by author. This is a useful task explained in Coulthard (2004) and could be used for example in plagiarism detection and even for author identification in crimes involving text evidence. We shall classify Dickens' novels with other 19th century authors. We will think of a graph Laplacian belonging to the class '1' as representing a novel written by Dickens. To begin we just classify Dickens and Austen novels and so the '0' class are graph Laplacians representing Austen's novels.

For the Austen and Dickens dataset, Figure 5.1 shows the probability of classifying a network as corresponding to a Dickens novel using Method 1 and Method 2 using the linear discriminant analysis classifier in the tangent space. The LDA is performed firstly using just the first PC for the Euclidean and square root Euclidean metrics. We can see for the Euclidean metric this does a good job only incorrectly classifying *David Copperfield*. For the square root metric all the novels are correctly classified. When LDA is performed using the first 2 PCs then for both metrics all novels are correctly classified. For Method 1 all novels are correctly classified for both metrics.

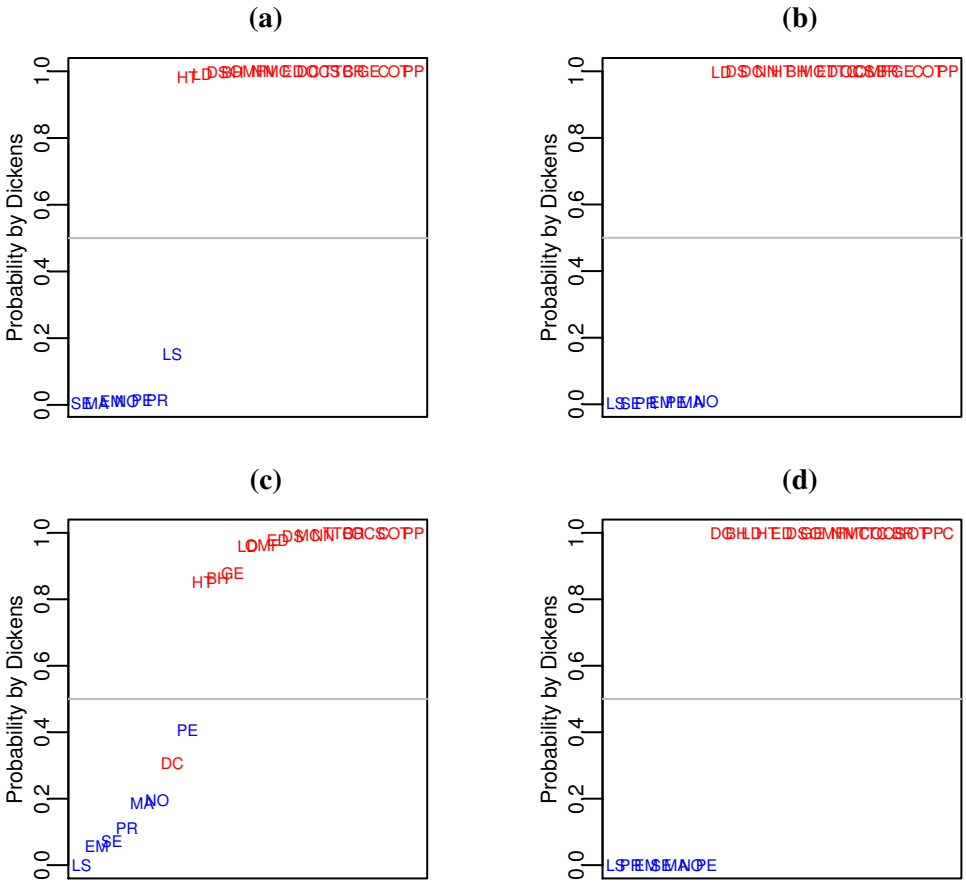


Figure 5.1: *The probability of classifing if a novel was written by Dickens, ordered along the x axis by the magnitude of this probability. Coloured red if Dickens novel or blue if Austen novel. The top row is when classifying using Method 1 using the a) Euclidean and b) square root Euclidean metric. The bottom row is classifying using Method 2 with the LDA classifier with 1 PCA coordinate for the c) Euclidean and b) square root Euclidean metric. The abbreviations for novels are found in Table 1.2.*

We saw in Example 4.6.3 that the Austen and Dickens novels have significantly different means and the two novelists are very well separated on 1st and 2nd PC plots in Example 2.5.1, therefore it is not surprising classifying them is almost trivial. To demonstrate our method for a more interesting example we look at the full 19th century novel data. The 1st and 2nd PC scores for all these novels are plotted in Figure 5.2 for the Euclidean and square root Euclidean metrics, there is far more overlapping of Dickens novels with the extra 19th century author's novels, so classifying Dickens novel with the addition of these extra authors is a less trivial example. For this example the class '1' now is still the graph Laplacians representing Dickens novels however the '0' class represents any non-Dickens 19th century author's novel. From the PC plots in Figure 5.2 the Dickens novels form a more distinct cluster when using the square root Euclidean metric compared to

using the Euclidean metric, and so it seems likely we will get better results for the classification when using the square root Euclidean metric.

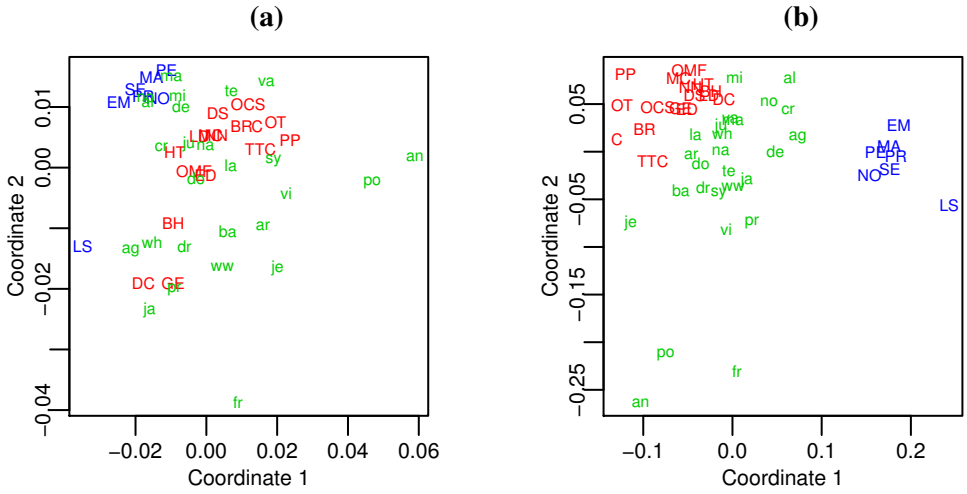


Figure 5.2: *Plot of PC1 and PC2 scores for all 19th century novels using the (a) Euclidean metric and (b) square root Euclidean metric. Red - Dickens, blue - Austen and green - other. The abbreviations for novels are found in Tables 1.2 and 1.3.*

Figure 5.3 shows the classification probabilities when using Method 1, for the Euclidean and square root Euclidean metrics, the Procrustes size-and-shape is not included as results are nearly identical to results for the square root Euclidean metric. The bandwidth chosen was 0.010, 0.073 and 0.072 for the Euclidean, square root Euclidean and Procrustes size-and-shape respectively. For the square root metric the novels with the highest probability of being written by Dickens are his novels and hence if we classified a novel as being written by Dickens if it has over 0.8 probability as being written by him we would get a classification accuracy of 100%. For the Euclidean metric the novels with highest probability of being written by Dickens are all his own novel with the exception of *Vanity Fair* and *Dracula*, and so no classification rule exists that could give 100% accuracy.

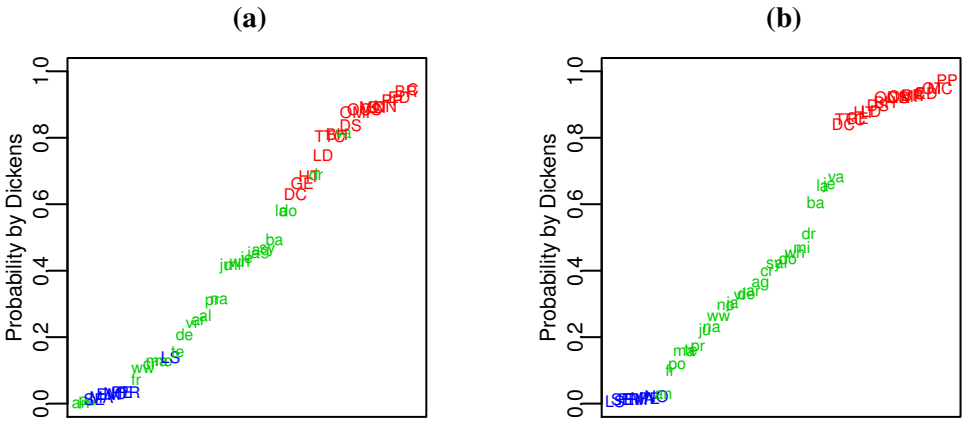


Figure 5.3: *The probability of classifing if a novel was written by Dickens using Method 1 using the a) Euclidean and b) square root Euclidean metric. Coloured red if Dickens novel, blue if Austen novel and green if another author. Ordered along the x axis by the magnitude of this probability. The abbreviations for novels are found in Tables 1.2 and 1.3.*

Figure 5.4, 5.5 and 5.6 show the classification probabilities for the full 19th century novel dataset when using Method 2 using LDA, random forests and SVM respectively. Performing LDA on the square root Euclidean PC scores gives very good results, for just two PCs the novels with the largest probability of being written by Dickens are almost all his novels, and when using 8 PCs all the novels are correctly classified if the classification probability was chosen to be 0.8. Performing LDA on the Euclidean PC scores does not give as good results. Keeping the first 2 PCs leads to poor prediction of whether a novel was written by Dickens. The prediction is improved when more PCs are included. When using the first 8 PCs the novels with a high probability of being Dickens are all his novels with the exception of *Vanity Fair*. The random forest and SVM both perform worse than LDA for both metrics as even when 8 PCs are used for both there is no classification rule that could give 100% classification accuracy. We see repeatedly Charles Dickens' own novel *David Copperfield* has quite a low fitted probability of being written by Dickens in our classifications, this novel is thought to be semi-autobiographical and perhaps this explains why it would be misclassified (LaFarge, 2009). As well often Robert Louis Stevenson's *Jekyll and Hyde*, William Thackeray's *Vanity Fair* and Emily Brontë's *Whuthering Heights* have high fitted probabilities of being written by Dickens. William Thackeray knew Charles Dickens and were described as literary rivals, so perhaps it is unsurprising that they may write similarly (Maggie Kopp, 2011).

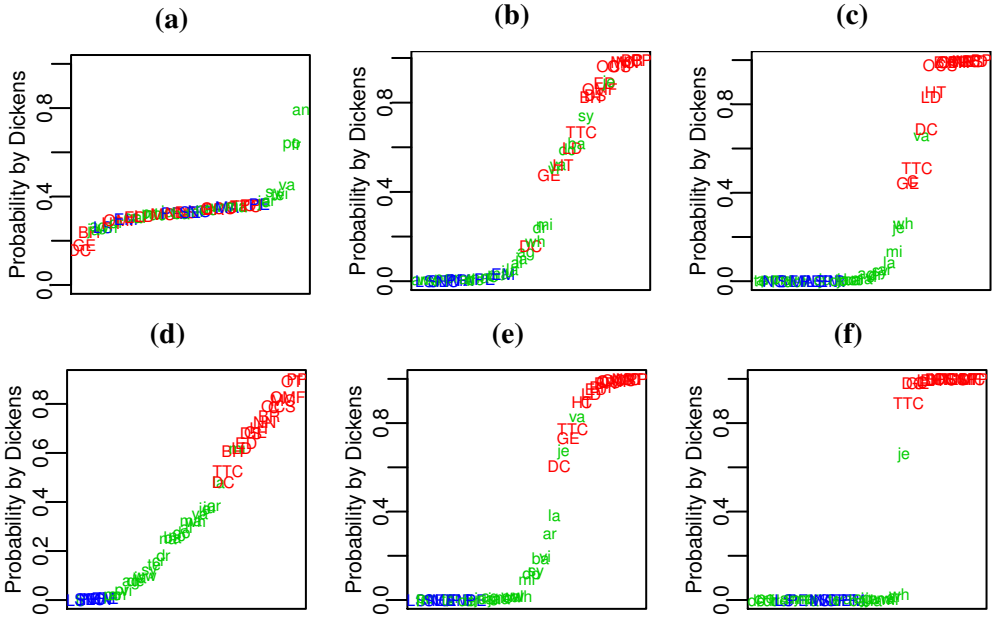


Figure 5.4: *The probability of classifing if a novel was written by Dickens using LDA, ordered along the x axis by the magnitude of this probability. Coloured red if Dickens novel, blue if Austen novel and Green if other. Using (left to right) 2, 5 or 8 PCA coordinates, from the (top) Euclidean and (bottom) square root Euclidean metrics, for the classification. The abbreviations for novels are found in Tables 1.2 and 1.3.*

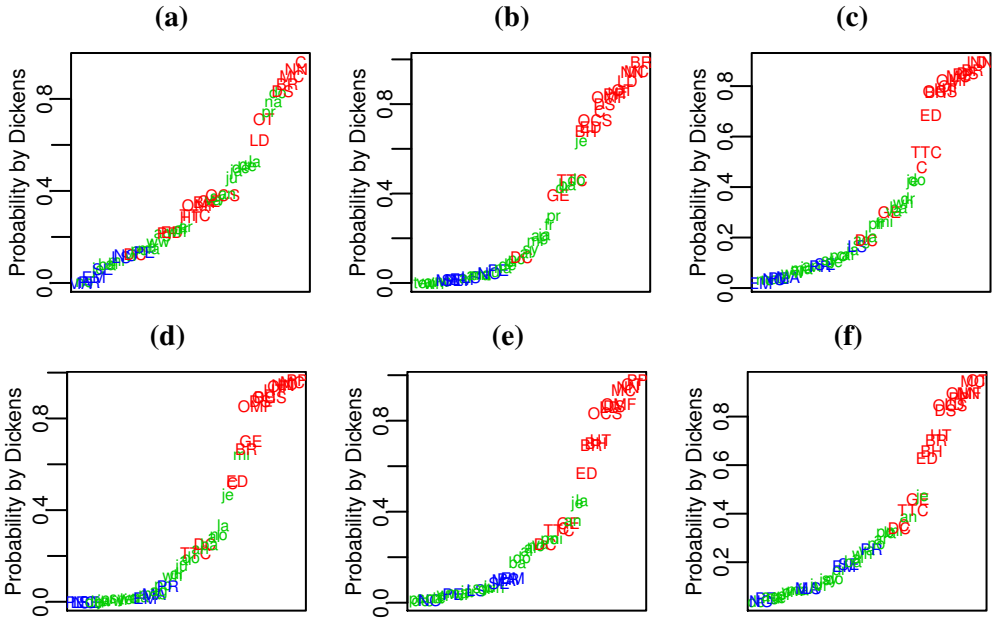


Figure 5.5: *The probability of classifing if a novel was written by Dickens using Random forests, ordered along the x axis by the magnitude of this probability. Coloured red if Dickens novel, blue if Austen novel and Green if other. Using (left to right) 2, 5 or 8 PCA coordinates, from the (top) Euclidean and (bottom) square root Euclidean metrics, for the classification. The abbreviations for novels are found in Tables 1.2 and 1.3.*

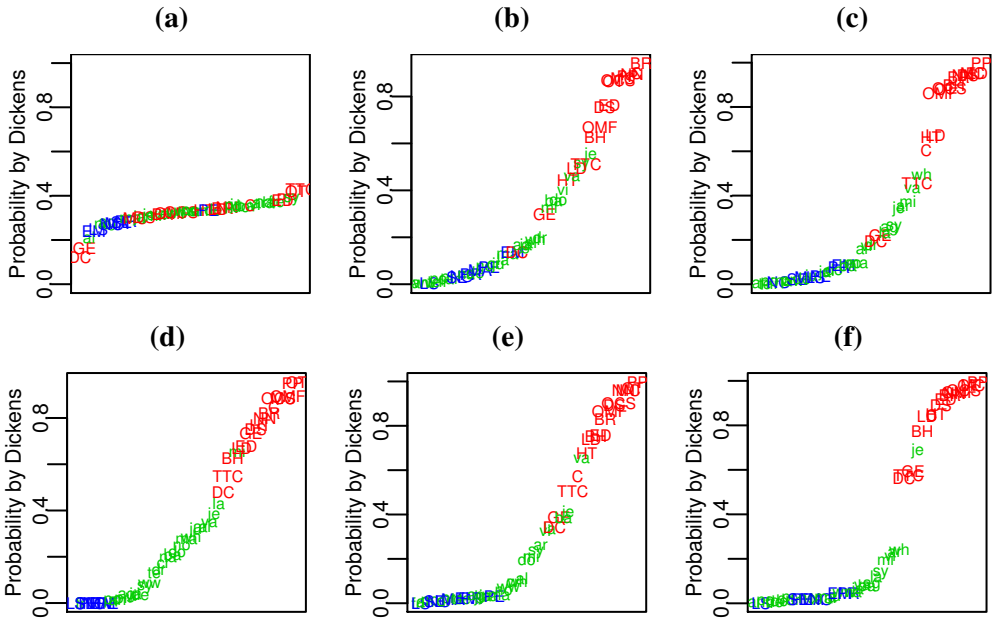


Figure 5.6: *The probability of classifing if a novel was written by Dickens using SVM, ordered along the x axis by the magnitude of this probability. Coloured red if Dickens novel, blue if Austen novel and Green if other. Using (left to right) 2, 5 or 8 PCA coordinates, from the (top) Euclidean and (bottom) square root Euclidean metrics, for the classification. The abbreviations for novels are found in Tables 1.2 and 1.3.*

Example 5.1.2: Classification methods applied to the M-money transaction data

We have seen for the M-money transaction data in Example 4.6.2 that the mean network for a weekday is significantly different to the mean of a weekend. We shall use our different classification methods to try and classify the M-money networks into weekdays and weekend days. In Example 2.5.2 we saw from the plot of first and second PC coordinates that the weekend days, especially Saturdays overlapped a lot with weekdays, we therefore do not expect to get 100% accuracies in this classification.

To compare the different classification methods we shall compare their maximum accuracies, where accuracy is defined in (1.2.10). We choose to look at a balanced dataset so we know the accuracy should be above 50% as we can achieve 50% accuracy by just classifying randomly. The balanced dataset consists of the graph Laplacians for the 104 weekend days of the year, class '1', and then 104 randomly selected graph Laplacians corresponding to weekdays, class '0'. The classification were again run using a leave one out cross validation strategy. The threshold probability to classify a graph Laplacian as a weekend was then chosen to give the highest accuracy. The results are found in Table 5.1. The largest accuracy achieved, 69.2% correspond to using a random forest with the top 8 Euclidean PC scores. Using the Nadaraya Watson approach also

gives good results of around 63/64% for any of the Euclidean, square root Euclidean and Procrustes size-and-shape metric.

Method used	**Pcs used**	**Threshold probability**	**Maximum Accuracy** (%)
Euclidean NW	NA	0.47	64.4
Square root Euclidean NW	NA	0.47	63.5
Procrustes size-and-shape NW	NA	0.47	63.5
Euclidean LDA	2	0.59	57.2
Euclidean LDA	5	0.52	57.2
Euclidean LDA	8	0.50	66.8
Square root Euclidean LDA	2	0.56	58.2
Square root Euclidean LDA	5	0.55	62.0
Square root Euclidean LDA	8	0.50	65.4
Euclidean RF	2	0.62	54.8
Euclidean RF	5	0.50	61.5
Euclidean RF	8	0.43	69.2
Square root Euclidean RF	2	0.89	52.9
Square root Euclidean RF	5	0.46	62.0
Square root Euclidean RF	8	0.43	63.0
Euclidean SVM	2	0.56	51.4
Euclidean SVM	5	0.54	51.0
Euclidean SVM	8	0.18	50.0
Square root Euclidean SVM	2	0.58	53.4
Square root Euclidean SVM	5	0.26	50.0
Square root Euclidean SVM	8	0.33	50.0

Table 5.1: *Maximum classification accuracies for the M-money networks, classifying if a network corresponds to a weekday or weekend. The threshold probability is the threshold to classify a graph Laplacian as a weekend to give the corresponding accuracy.*

5.2 Anomaly detection

Detecting anomalies in data is a task of interest in statistics, and this remains true when the data consists of networks (Akoglu et al., 2015). Anomaly detection is similar to a classification problem, where one class is the majority of the data, 'inliers', and a second class are the outliers or anomalies. However anomaly detection differs from standard classification problems as it is an unsupervised problem as we have no training data in which the class each graph Laplacian belongs to is known. A simple way to investigate anomalies of networks is by looking at a 2D representation of graph Laplacians, such as MDS or PCA plots where anomalies can be detected visually (Bunke et al., 2007) which we saw in Section 2.5. However this relies on our own judgements and therefore is subjective, so we propose a classification rule for outliers using our graph Laplacian framework.

A simple and intuitive way of detecting an anomaly is by considering the distance between each graph Laplacian and the unprojected sample mean, where the distances are

$$d_\alpha(\boldsymbol{L}_k, \hat{\boldsymbol{\eta}})$$
$$d_{\alpha,S}(\boldsymbol{L}_k, \hat{\boldsymbol{\eta}}),$$

and $\hat{\boldsymbol{\eta}}$ is the sample of graph Laplacians unprojected sample mean defined in Equation (2.3.1). If a graph Laplacian has a much greater distance from the mean than other graph Laplacians within the sample this could be an indication that it is an anomaly.

To classify a graph Laplacian as an anomaly requires defining a threshold such that a distance to the sample mean greater than this threshold indicates an anomaly. This threshold can be found when we choose the distance as the Euclidean power distance. To calculate this threshold we will work with the distance squares, as the distribution and hence threshold of these can be found and then square rooted back to distances. The test statistic that we are therefore using is $Z = d_\alpha(\boldsymbol{L}, \hat{\boldsymbol{\eta}})^2$. One way to calculate the threshold is to assume a model for the graph Laplacians and we will assume

$$\phi(\mathrm{F}_\alpha(\boldsymbol{L})) \sim \mathcal{N}_{\frac{m(m-1)}{2}}(\phi(\mathrm{F}_\alpha(\boldsymbol{\eta})), \boldsymbol{\Sigma}), \tag{5.2.1}$$

where ϕ is defined in (0.0.4), hence we are working in the off diagonal space of the embedded graph Laplacians, similarly to Section 4. Due to the consistency of the sample means in Result 2.3.1, as $n \to \infty$ we have $\phi(\mathrm{F}_\alpha(\hat{\boldsymbol{\eta}})) \to \phi(\mathrm{F}_\alpha(\boldsymbol{\eta}))$. Therefore using the

central limit theorem in Result 4.2.1, as $n \to \infty$

$$\phi(\mathrm{F}_\alpha(\boldsymbol{L})) - \phi(\mathrm{F}_\alpha(\hat{\boldsymbol{\eta}})) \sim \mathcal{N}_{\frac{m(m-1)}{2}}(\mathbf{0}, \boldsymbol{\Sigma}).$$

Using the same logic as used to prove Result 4.2.2 we can write the distance squared as a quadratic form of normals,

$$Z = d_\alpha(\boldsymbol{L}, \hat{\boldsymbol{\eta}})^2 = \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x},$$

where $\boldsymbol{x} = \phi(\mathrm{F}_\alpha(\boldsymbol{L})) - \phi(\mathrm{F}_\alpha(\hat{\boldsymbol{\eta}}))$ and $\boldsymbol{Q}$ is defined in (4.2.8). Therefore similarly to Result 4.2.2 when $n \to \infty$

$$Z = d_\alpha(\boldsymbol{L}, \hat{\boldsymbol{\eta}})^2 \xrightarrow{D} \sum_{i=1}^{m(m-1)/2} \lambda_i \chi_1^2, \tag{5.2.2}$$

in which each χ_1^2 is independent and λ_i are the $m(m-1)/2$ non-zero eigenvalues of $\boldsymbol{\Sigma Q}$. In general the value of $\boldsymbol{\Sigma}$ and hence the λ_is are unknown and we will have to estimate $\boldsymbol{\Sigma}$ and to do this we use the shrinkage estimator from Schäfer and Strimmer (2005), that we also used in our two-sample test in Section 4.2.

Once the λ_is are found, then through large simulations the quantiles of the distribution for the distance squared can be found which can then be used as thresholds, for example at a significance value of $100a\%$, we can find by simulation a threshold, c, as the $100(1-a)$th quantile of $\sum_{i=1}^{m(m-1)/2} \lambda_i \chi_1^2$, we therefore classify a graph Laplacian as an outlier if

$$Z = d_\alpha(\boldsymbol{L}, \hat{\boldsymbol{\eta}})^2 > c.$$

Often it is not appropriate to estimate $\boldsymbol{\Sigma}$, especially when m is large, as we have already seen in Chapter 4, and in these cases it is better to approximate the distribution in (5.2.2) to avoid estimating $\boldsymbol{\Sigma}$. The distribution can be approximated using results from Box (1954), as we have seen previously in (4.2.10) for the approximation of the distribution of the two-sample test. This approximation is

$$Z = d_\alpha(\boldsymbol{L}, \hat{\boldsymbol{\eta}})^2 \dot{\sim} g\chi_h^2. \tag{5.2.3}$$

As $\boldsymbol{\Sigma}$ is inappropriate to estimate we shall not use the λ values to estimate g and h and

instead estimate them using the median and upper and lower quartiles of the distribution, denoted for this distribution as $Z_{0.5}(h, g)$, $Z_{0.75}(h, g)$ and $Z_{0.25}(h, g)$. We use the quartiles to approximate g and h, as the quartiles are robust to anomalies and as we are using data which may contain anomalies we want a method that is robust to the anomalies. The quartiles for the distance squared to the mean for the observed data are easily calculated and denoted $\hat{Z}_{0.5}$, $\hat{Z}_{0.75}$ and $\hat{Z}_{0.25}$. To calculate g and h we set the median of the approximate distribution to the known median to give

$$\begin{aligned} \hat{Z}_{0.5} &\approx \hat{g}\hat{h}(1 - \frac{2}{9\hat{h}})^2 \\ \hat{g} &\approx \frac{\hat{Z}_{0.5}}{\hat{h}(1 - \frac{2}{9\hat{h}})^2}. \end{aligned} \tag{5.2.4}$$

We can then find the $\hat{h}$ that minimises the sum of the squares between the quartiles, given as

$$\hat{h} = \arg_h \min((Z_{0.5}(h) - \hat{Z}_{0.5})^2 + (Z_{0.75}(h) - \hat{Z}_{0.75})^2 + (Z_{0.25}(h) - \hat{Z}_{0.25})^2).$$

The value of $\hat{h}$ can be found using the `optimize` function in R (R Core Team, 2018). We now have estimates $\hat{h}$ and $\hat{g}$ for the unknown parameters, g and h, of (5.2.3) and hence an approximated distribution of Z, the test statistic, is now known and the threshold for this at a $100a\%$ significance level is $g\chi^2_{h,1-a}$, which will be known.

Example 5.2.1: Anomaly detection applied to Austen and Dickens novels

It is interesting to determine if certain Austen and Dickens novels are outliers with the rest of their respective writing. We will use the method we have described for both authors using $\alpha = 1$ to classify anomalies. A limitation of our method is when calculating the threshold we assumed $n \to \infty$, for both authors n is not at all large and so this assumption is violated. We also are assuming the graph Laplacians follow a normal distribution as stated in Equation (5.2.1). We will compare the actual distribution of distance squared with the theoretical and approximated theoretical distribution to see the effect of these assumptions.

To use the theoretical distribution in (5.2.2) we must estimate a covariance matrix, and when $m = 1000$ this covariance is far too large to estimate hence instead we just look at the top 50 words for all novels so that our graph Laplacians have $m = 50$. Figure 5.7 includes plots for both Authors of the distance each novel is from the mean novel

using the Euclidean metric. The threshold to classify a novel as an anomaly at a 0.1% significance level is included for both authors too. The threshold was calculated by a million simulations of the distribution which was deemed large enough to give sensible results. From these plots we can see clearly that the novel Lady Susan is an anomaly of Jane Austen's novels. *Lady Susan* is actually a novella of Austen's and is sensible to be chosen as an outlier with it often being referred to as atypical for Austen's work as it comprises of letters (Gaston, 2016). Also *Persuasion* is above the threshold and so is also suggested to be an outlier. For the Dickens novels the novels, *The Pickwick Papers*, *Oliver Twist*, *The Old Curiosity Shop*, *A Christmas Carol*, *David Copperfield*, *Bleak House* and *Great Expectations* are all above the threshold and hence anomalies. Our method is picking out over a third of Dickens' novels as anomalies and so this is indicating the theoretical threshold is not sensible.

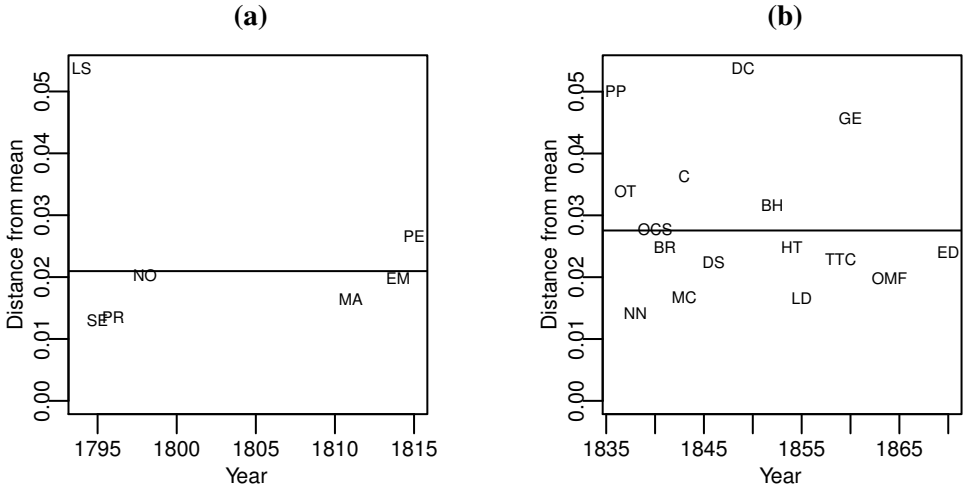


Figure 5.7: *Distance a novel's graph Laplacian is from their authors mean for the Euclidean metric, for Austen (left) and Dickens (right), with the threshold line for an anomaly included for $m = 50$. The abbreviations for novels are found in Table 1.2.*

In Figure 5.8 the distribution of the distance squared, found from the data, for $\alpha = 1$ is plotted against the theoretical distribution given in (5.2.2) when $m = 50$, it is clear these distributions do not match up at all, and so using this distribution to calculate a threshold is not sensible. When using the approximated distribution, now for $m = 1000$, we can see from Figure 5.8 that this distribution does match up well and so is far more sensible to use to estimate a threshold. This is most likely as when we use the approximated distribution we no longer need to estimate a large covariance matrix. We choose $m = 1000$ as for the approximated distribution we do not need to estimate a

large covariance matrix so having a larger m is OK.

Figure 5.9 has the distances for $m = 1000$ and the approximated threshold line. This gives far more sensible results than before with none of Dickens novels picked out as anomalies and just *Lady Susan* as an anomaly for Austen.

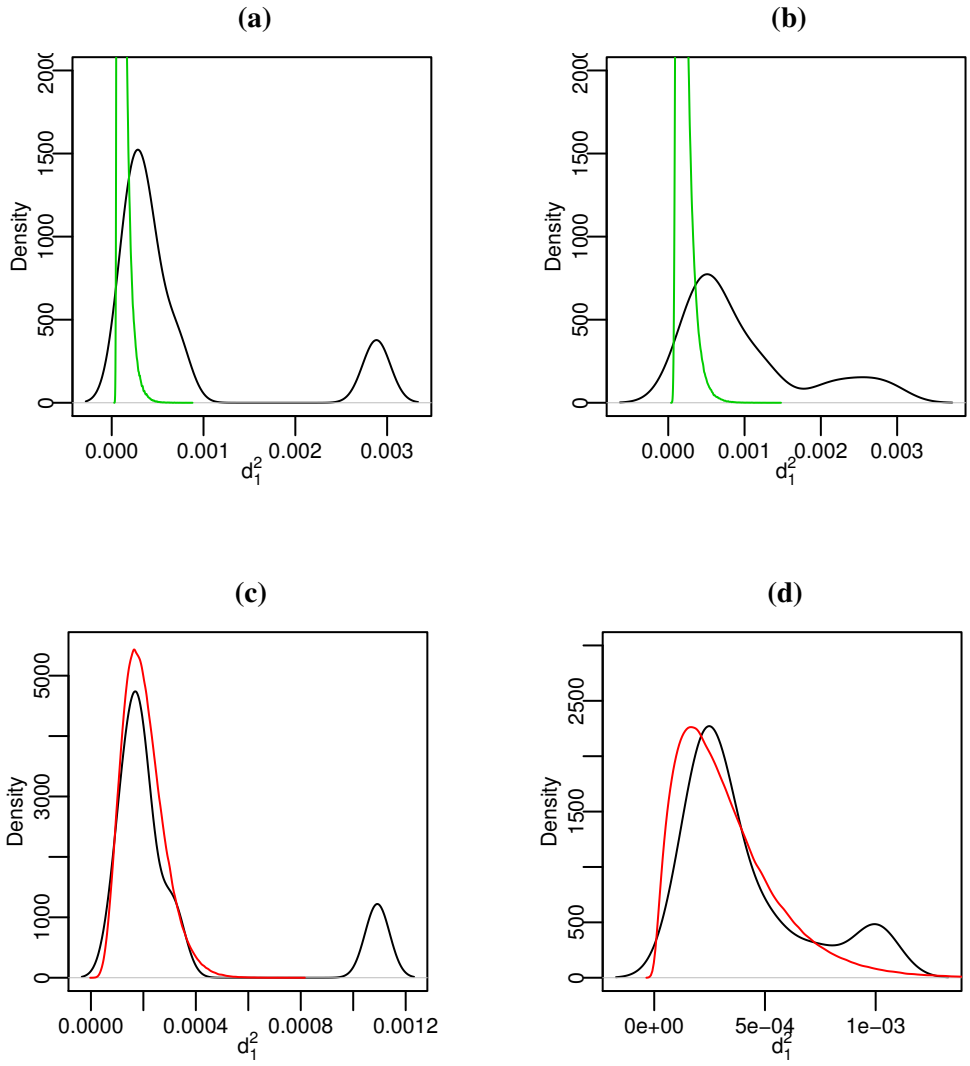


Figure 5.8: *Distribution of d_1^2 for Austen's novels (left) and Dickens' novels (right) for $m = 50$ (top) and $m = 1000$ (bottom). Black- true distribution, green - theoretical distribution in (5.2.2) and red- approximated theoretical distribution given in (5.2.3) using g and h approximated using the quartiles.*

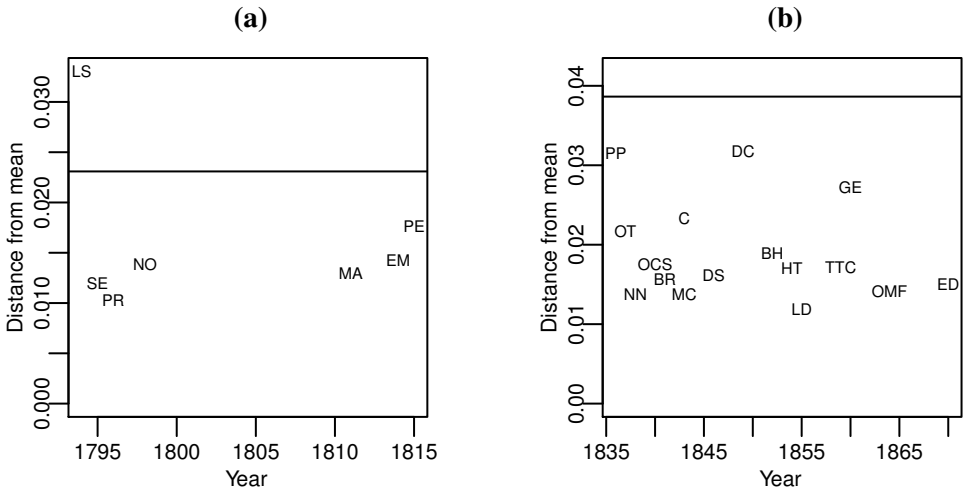


Figure 5.9: *Distance a novel's graph Laplacian is from their authors mean for the Euclidean metric, for Austen (left) and Dickens (right), with the threshold line for an the approxmiated threshold for* $m = 1000$*. The abbreviations for novels are found in Table 1.2.*

5.3 Summary

In this chapter we have proposed two methods for the classification of graph Laplacians into binary classes. Both methods output a probability the graph Laplacian is in a certain class and this probability can be thresholded to assign the graph Laplacian to a class. The first method used Nadaraya-Watson regression on the manifold, very similar to the regression performed in Section 3.3 however now to predict probabilities. The second method used standard classification methods, like LDA, random forests and SVMs on the PC score space defined in our graph Laplacian framework. These methods were compared for the 19th century author data where classes were chosen as graph Laplacians representing novels by certain authors. The first method seemed favourable especially when using the square root Euclidean metric, however the second method still performed well, especially when LDA was used. Many other classification methods could also be considered for the second case, like logistic regression, which may out perform those we currently have considered.

We also looked at anomaly detection, where a graph Laplacian was considered an anomaly if its distance to the unprojected mean was above a threshold. This threshold could be found asymptotically however for the Dickens and Austen novels, which we applied it on, this threshold did not give sensible results. We described a method of estimating the threshold, which gave far better results for Dickens and Austen novels.

CHAPTER 6

Comparing tangent coordinates

For the previous chapters when we have used ideas from shape analysis we have stuck to the size-and-shape space, defined in (1.1.6), where objects are not restricted to be invariant to scale but had invariance to reflection. However in this chapter we will now define and consider the specific case in shape analysis where the effect of scale is removed but the effect of reflection remains; this space that we work on is called the shape space or before the effect of rotation is removed the pre-shape space.

There are many instances in shape analysis where scale is removed, an obvious one being if the objects are not recorded on the same scale so scale must be removed to make the objects comparable. Also in many applications where scale information is available instead of working in the size-and-shape space it can be beneficial to work in the shape space and consider the size variable separately (Dryden and Mardia, 2016). In shape analysis it is common for information on reflection of the shape to remain, as often we would consider shapes to be different if they were reflections of one another.

Just like our statistical analysis of networks, in shape analysis the use of a linear space such as a tangent space to the shape space or pre-shape space is of interest as standard multivariate analysis can be applied here, such as shape PCA (Kent, 1994). Unlike the size-and-shape space we used in previous chapters where there is only one way commonly used to project to the tangent space, for the shape space there are several common ways of projecting on to a tangent space; the three we consider here are residual tangent coordinates, partial tangent coordinates and inverse exponential map tangent coordinates. Partial and inverse exponential map tangent coordinates are projections of a configuration onto the tangent space, whereas residual tangent coordinates are an approximation onto the tangent space and this approximation is only good for low vari-

ability data.

In this chapter we investigate empirically the characteristics of the three different tangent coordinates in the context of different datasets. We explore why residual tangent coordinates are not appropriate for large variability datasets specifically when applying shape PCA and demonstrate this idea using several datasets. Finally we will conclude and provide guidance on which tangent coordinates are most suitable to use.

6.1 Shape analysis

We briefly introduced the idea of shape analysis in Section 1.1.1. As we mentioned, in the present chapter we will consider a space where we have invariance to scaling but not reflection, named the shape space. For a $k \times m$ configuration matrix, $\boldsymbol{X}_i$, where k is the number of landmarks and m is the number of dimensions the shape space for it is defined

$$[\boldsymbol{X}]_S = \{\boldsymbol{Z}\boldsymbol{R} : \boldsymbol{R} \in \mathcal{SO}_m\},$$

where

$$\boldsymbol{Z}_i = \frac{\boldsymbol{H}\boldsymbol{X}_i}{\|\boldsymbol{H}\boldsymbol{X}_i\|},$$

and $\boldsymbol{H}$, the Helmert sub-matrix, is defined in (1.1.5). The rotation term $\boldsymbol{R}$ belongs now to $\mathcal{SO}_m$, the set of orthogonal matrices with determinant 1, and not $\mathcal{O}_m$ like in previous chapters as we now do not have invariance to reflection. From the definition of $\boldsymbol{H}$ the denominator is equivalent to the centroid size of $\boldsymbol{X}_i$ (Dryden and Mardia, 2016, Section 3.2.5), where centroid size is defined

$$S(\boldsymbol{X}) = \|\boldsymbol{X} - \mathbf{1}\hat{\boldsymbol{x}}^T\|, \tag{6.1.1}$$

where $\hat{\boldsymbol{x}}$ is the centroid which is a column vector with ith element $(\frac{1}{k}\sum_{i=1}^{k} X_{ij})$. The $\boldsymbol{Z}_i$ matrix is a $(k-1) \times m$ matrix which lies on the pre-shape sphere, as it has had location and scale removed. For a matrix to lie on the pre-shape sphere it must satisfy $\|\boldsymbol{Z}_i\| = 1$.

There are three common distance metrics used on the pre-shape sphere. These three

distances are the full Procrustes distance, d_F, the partial Procrustes distance, d_P, and the Riemannian distance, ρ, defined between the configuration matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, with pre-shape coordinates $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ respectively, as,

$$d_P(\boldsymbol{X}_1, \boldsymbol{X}_2) = \inf_{\boldsymbol{\Gamma} \in SO(m)} \|\boldsymbol{Z}_2 - \boldsymbol{Z}_1 \hat{\boldsymbol{\Gamma}}\|, \tag{6.1.2}$$

$$d_F(\boldsymbol{X}_1, \boldsymbol{X}_2) = \inf_{\hat{\boldsymbol{\Gamma}} \in SO(m), \beta \in \mathbb{R}^+} \|\boldsymbol{Z}_2 - \beta \boldsymbol{Z}_1 \hat{\boldsymbol{\Gamma}}\|, \tag{6.1.3}$$

$$\rho(\boldsymbol{X}_1, \boldsymbol{X}_2) = \inf_{\boldsymbol{\Gamma} \in SO(m)} \arccos(\boldsymbol{Z}_1^T \boldsymbol{Z}_2 \hat{\boldsymbol{\Gamma}}). \tag{6.1.4}$$

The Riemannian distance, ρ, is the minimised great circle distance, i.e. the minimal geodesic path, defined in Section 1.1, carried out over rotations between $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$. $\boldsymbol{\Gamma}$ is the optimal rotation defined in (1.1.7), but restricted to now belong to $\mathcal{SO}_m$, this can be written explicitly as

$$\begin{aligned} \hat{\boldsymbol{\Gamma}} &= \boldsymbol{U}\boldsymbol{V}^T \\ \text{where } \boldsymbol{Z}_2^T \boldsymbol{Z}_1 &= \|\boldsymbol{Z}_1\| \|\boldsymbol{Z}_2\| \boldsymbol{V} \boldsymbol{\Lambda} \boldsymbol{U}^T, \boldsymbol{U}, \boldsymbol{V} \in \mathcal{SO}(m). \end{aligned} \tag{6.1.5}$$

The scaling parameter β can be written explicitly as

$$\beta = \frac{\text{trace}(\boldsymbol{Z}_2^T \boldsymbol{Z}_1 \hat{\boldsymbol{\Gamma}})}{\text{trace}(\boldsymbol{X}_1^T \boldsymbol{X}_1)}. \tag{6.1.6}$$

The three distances are all related hence the partial and full Procrustes distance can be written in terms of the Riemannian distance, ρ, by

$$\begin{aligned} d_P &= 2 \sin(\frac{\rho}{2}) \\ d_F &= \sin(\rho), \end{aligned} \tag{6.1.7}$$

see Dryden and Mardia (2016).

6.2 Shape tangent coordinates

Just like the analysis of graph Laplacians, in shape analysis we work in a tangent space where we can perform standard statistical methods. For the shape analysis we consider, where shapes are invariant to scale but not reflection, the tangent space is a linearized space tangent to a pole. We will normally choose the pole as the sample full Procrustes mean shape on the pre-shape sphere. Using (1.1.3) the sample full Procrustes mean shape on the pre-shape sphere, $\hat{\boldsymbol{\mu}}$ ($(k-1)\times m$), is defined as

$$\hat{\boldsymbol{\mu}} = \frac{\boldsymbol{H}\left(\arg\inf_{\mu'\in\mathbb{R}^{k\times m}} \frac{1}{n}\sum_{k=1}^{n} d_F^2(\mu', X_k)\right)}{\|\boldsymbol{H}\left(\arg\inf_{\mu'\in\mathbb{R}^{k\times m}} \frac{1}{n}\sum_{k=1}^{n} d_F^2(\mu', X_k)\right)\|}. \tag{6.2.1}$$

The dimension, q, of the tangent space is

$$q = km - m - \frac{m(m-1)}{2} - 1. \tag{6.2.2}$$

This is as the original space is km dimensions then m are removed by translation constraints, $m(m-1)/2$ by rotational constraints and 1 due to size constraints. The rotational constraint comes from the fact the tangent space we consider does not depend on rotation (Dryden and Mardia, 2016, Page 65).

There are multiple ways of projecting on to the tangent space, we shall now define and study the same three as defined in Section 4.4 Dryden and Mardia (2016) which are the (i) residual tangent coordinates, (ii) partial tangent coordinates and (iii) inverse exponential map tangent coordinates, schematics for them are found in Figure 6.1.

6.2.1 Residual tangent coordinates

The residual tangent coordinates appear to be favoured by practitioners as they are formulated in a more straightforward way than the other tangent coordinates (Dryden and Mardia, 2016). However the residual tangent coordinates only give an approximation to the tangent space as configurations are not projected onto it, seen for the pre-shape residuals in the schematic in Figure 6.1. Also the term 'residual tangent coordinates' is used by different authors to mean different things. We review two which we term pre-shape residuals and denote by $\boldsymbol{v}_R$ (used in Dryden and Mardia (2016)) and Procrustes residuals denoted by $\boldsymbol{v}_{Rproc}$ (used in the implementation of the generalised Procrustes

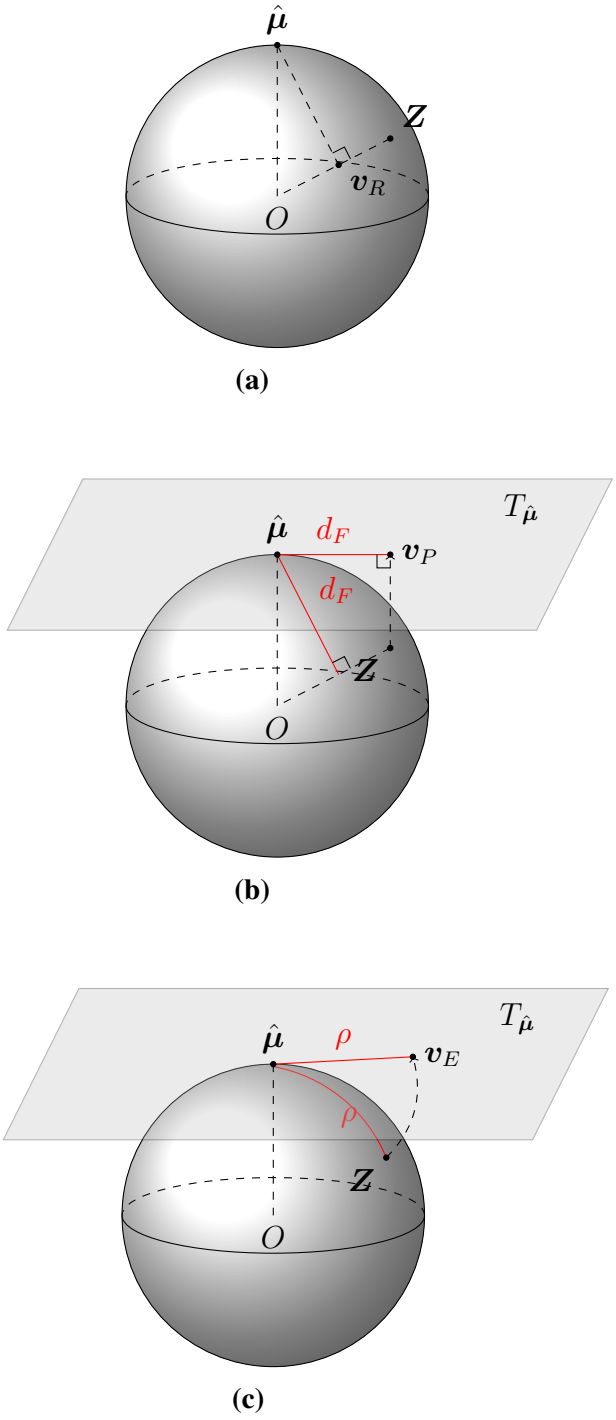


Figure 6.1: *The different tangent coordinates, (a) the residual, (b) the parital and (c) the inverse exponential.*

algorithm found in the `shapes` package in R (Dryden, 2018) based on work from Goodall (1991), Ten Berge (1977) and Gower (1975)) We shall define both types of residual tangent coordinates and explain why it is appropriate for us to only consider the pre-shape residuals.

The pre-shape residual tangent coordinates are defined for a configuration on the pre-shape sphere $\boldsymbol{Z}_i$, with the sample full Procrustes mean shape $\hat{\boldsymbol{\mu}}$ on the pre-shape sphere, defined in (6.2.1), chosen as the pole, as

$$\boldsymbol{v}_R^{(i)} = \text{vec}(\hat{\beta}^{(i)}\boldsymbol{Z}_i\hat{\boldsymbol{\Gamma}}^{(i)}) - \text{vec}(\hat{\boldsymbol{\mu}}),$$

where the optimal rotation, $\hat{\boldsymbol{\Gamma}}^{(i)}$, and the optimal scaling, $\hat{\beta}^{(i)}$, are given in (6.1.5) and (6.1.6) to minimize the Procrustes distance between $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{Z}_i$, and where vec is defined in (0.0.1).

The Procrustes residuals are similar and defined as

$$\boldsymbol{v}_{Rproc}^{(i)} = c\left(\text{vec}(\hat{\beta}^{(i)}\boldsymbol{Z}_i\hat{\boldsymbol{\Gamma}}^{(i)}) - \frac{1}{n}\sum_{j=1}^{n}\text{vec}(\hat{\beta}^{(j)}\boldsymbol{Z}_j\hat{\boldsymbol{\Gamma}}^{(j)})\right), \tag{6.2.3}$$

where

$$c = \frac{\sqrt{\sum_{j=1}^{n} S(\boldsymbol{X}_j)^2}}{\sqrt{\sum_{j=1}^{n} S(\hat{\beta}^{(j)}\boldsymbol{Z}_j\hat{\boldsymbol{\Gamma}})^2}},$$

where $S(\boldsymbol{X})$ is defined in (6.1.1). These differ to the pre-shape residuals as now the arithmetic mean of the configurations registered to the mean is subtracted, which is not on the pre-shape sphere and hence does not have unit size as $\hat{\boldsymbol{\mu}}$ does for the pre-shape residuals. The Procrustes residuals also use scaling, c, so the sum of the squared centroid sizes for the original configurations is equal to that for the registered configurations.

We can write the Procrustes residuals in terms of the pre-shape residuals as

$$\begin{aligned}\boldsymbol{v}_{Rproc} &= c(\boldsymbol{v}_R + \boldsymbol{a})\\ \text{where } \boldsymbol{a} &= \text{vec}(\hat{\boldsymbol{\mu}}) - \frac{1}{n}\sum_{j=1}^{n}\text{vec}(\hat{\beta}^{(j)}\boldsymbol{Z}_j\hat{\boldsymbol{\Gamma}}^{(j)}).\end{aligned}$$

As $\boldsymbol{v}_{Rproc}$ is just a translation followed by a scaling of $\boldsymbol{v}_R$ we do not consider the differ-

ences of $\boldsymbol{v}_{Rproc}$ and $\boldsymbol{v}_R$ further as both residuals will lead to an identical arrangement of PCA coordinates just on a different scale. From now on we think of the residual tangent coordinate as the pre-shape residual tangent coordinate, we choose these as their scale is comparable to that of the partial and inverse exponential map tangent coordinates.

As the residual tangent coordinates, $\boldsymbol{v}_R$, are an approximation to the tangent space coordinates these coordinates are not actually orthogonal to the chosen pole on the pre-shape sphere, meaning in general trace$(\boldsymbol{v}_R^T \text{vec}(\hat{\boldsymbol{\mu}})) \neq 0$. Further the space of residual tangent coordinates is $q+1$ dimensional, instead of q dimensional, defined in (6.2.2), this is one more than the pre-shape sphere as the size constraint is lost for this approximation. The residual approximation to a tangent space is good when configurations are close to the pole of the projection (Dryden and Mardia, 2016), however for datasets with high variability the approximation is unsuitable. We will compare $\boldsymbol{v}_R$ with the other two tangent coordinates to determine if the current wide use of $\boldsymbol{v}_R$ by practitioners is suitable.

6.2.2 Partial tangent coordinates

The partial tangent coordinates, $\boldsymbol{v}_P$, are formed by projecting a configuration up from the pre-shape sphere to the tangent space seen in Figure 6.1. For the configuration on the pre-shape sphere $\boldsymbol{Z}_i$ the partial tangent coordinates with $\boldsymbol{\gamma}$ ($(k-1) \times m$) chosen as the pole are

$$\boldsymbol{v}_P^{(i)} = [\boldsymbol{I}_{km-m} - \text{vec}(\boldsymbol{\gamma})\text{vec}(\boldsymbol{\gamma})^T]\text{vec}(\boldsymbol{Z}_i\hat{\boldsymbol{\Gamma}}^{(i)}), \tag{6.2.4}$$

where $\hat{\boldsymbol{\Gamma}}^{(i)}$ is defined in Equation 6.1.5 for $\boldsymbol{Z}_i$ and $\boldsymbol{\gamma}$. This type of tangent coordinates preserves the full Procrustes distance between points on the pre-shape sphere and the pole, so $\|\boldsymbol{v}_P^{(i)}\| = d_F^{(i)}$, where $d_F^{(i)} = d_F(\boldsymbol{X}_i, \boldsymbol{\gamma})$. We use the sample full Procrustes mean shape, $\hat{\boldsymbol{\mu}}$, defined in (6.2.1), as the pole, as this makes the tangent coordinate more comparable to the residual tangent coordinates which use the full Procrustes mean.

6.2.3 Inverse exponential map tangent coordinates

Inverse exponential map tangent coordinates are another projection of a configuration onto the tangent space seen in Figure 6.1. They have the property that the Riemannian distance between a point on the pre-shape sphere and the pole are preserved, so $\|\boldsymbol{v}_E^{(i)}\| = \rho^{(i)}$, where $\rho^{(i)} = \rho(\boldsymbol{X}_i, \boldsymbol{\gamma})$. A configuration on the pre-shape sphere, $\boldsymbol{Z}_i$, has inverse

exponential map tangent coordinates with $\boldsymbol{\gamma}$ ($(k-1)\times m$) as the pole given by,

$$\boldsymbol{v}_E^{(i)} = \frac{\rho^{(i)}}{\sin(\rho^{(i)})}[\boldsymbol{I}_{km-m} - \text{vec}(\boldsymbol{\gamma})\text{vec}(\boldsymbol{\gamma})^T]\text{vec}(\boldsymbol{Z}_i\hat{\boldsymbol{\Gamma}}^{(i)}), \tag{6.2.5}$$

where $\hat{\boldsymbol{\Gamma}}^{(i)}$ is defined in Equation 6.1.5 for $\boldsymbol{Z}_i$ and $\boldsymbol{\gamma}$ and $\rho^{(i)} = \rho(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}})$ defined in Equation 6.1.4. We again use the sample full Procrustes mean shape, $\hat{\boldsymbol{\mu}}$, defined in (6.2.1), as the pole.

6.2.4 Criteria for comparing tangent coordinates

To compare the tangent coordinates we will consider the difference between them for low variability data and higher variability data. By low variability we mean for each configuration's Riemannian distance to the sample full Procrustes mean is 'small' and by using (6.1.7) if ρ is small so are d_F and d_P.

A use of tangent coordinates is for performing shape PCA (Dryden and Mardia, 1993), where configurations are projected onto a tangent space and then standard PCA is performed on this. We expect for high variability data the choice of tangent coordinate is important in shape PCA and it is this effect we will study. However for low variability data the three different tangent coordinates are close to being equal and so the choice is not important.

Relation between tangent coordinates for data with low variability

To show the three tangent coordinates are close to being equal when there is low variability first we will show when ρ is small $\boldsymbol{v}_P$ and $\boldsymbol{v}_E$ are approximately equal. From the schematics in Figure 6.1 and the formulas for the tangent coordinates, (6.2.4) and (6.2.5), it is clear the difference between using $\boldsymbol{v}_P$ and $\boldsymbol{v}_E$ is just from their lengths. Hence $\boldsymbol{v}_E^{(i)}$ can be written in terms of $\boldsymbol{v}_P^{(i)}$ by

$$\boldsymbol{v}_E^{(i)} = \frac{\rho^{(i)}}{\sin(\rho^{(i)})}\boldsymbol{v}_P^{(i)}. \tag{6.2.6}$$

To show when ρ is small $\boldsymbol{v}_P$ and $\boldsymbol{v}_E$ are always approximately equal we use a Taylor expansion on (6.2.6), giving

$$\begin{aligned}\boldsymbol{v}_E &= \rho(\rho - \frac{\rho^3}{3!} + ...)^{-1}\boldsymbol{v}_P \\ &= \left(1 - \frac{\rho^2}{3!} + ...\right)^{-1}\boldsymbol{v}_P,\end{aligned}$$

and then we use a Taylor expansion again to give

$$\begin{aligned}&= (1 + \frac{\rho^2}{3!} - ...)\boldsymbol{v}_P \\ &= \boldsymbol{v}_P\left(1 + O(\rho^2)\right),\end{aligned}$$

and so we have shown $\boldsymbol{v}_P$ and $\boldsymbol{v}_E$ are approximately equal for low variability data.

To show $\boldsymbol{v}_R$ and $\boldsymbol{v}_P$ are approximately equal for small ρ we note for $\boldsymbol{v}_R$ when there is low variability it is clear, from the schematics in Figure 6.1, that very little scaling is needed to minimise the Procrustes distance between a point on the pre-shape sphere and the pole and hence the optimal scaling will be

$$\hat{\beta}^{(i)} = 1 - \epsilon_1,$$

where ϵ_1 is small and non-negative. Suppose that the partial Procrustes distance between the configuration and the pole, $\hat{\boldsymbol{\mu}}$, is small, then

$$\begin{aligned}\text{vec}(\boldsymbol{Z}\hat{\boldsymbol{\Gamma}}) &= \text{vec}(\hat{\boldsymbol{\mu}}) + \boldsymbol{\epsilon}_2 \\ \text{vec}(\hat{\boldsymbol{\mu}})^T\text{vec}(\boldsymbol{Z}\hat{\boldsymbol{\Gamma}}) &= 1 + \text{vec}(\hat{\boldsymbol{\mu}}^T)\boldsymbol{\epsilon}_2,\end{aligned}$$

where $\boldsymbol{\epsilon}_2$ is small, meaning $\|\boldsymbol{\epsilon}_2\| \ll 1$. Therefore

$$\begin{aligned}\boldsymbol{v}_R &= \text{vec}(\hat{\beta}\boldsymbol{Z}\hat{\boldsymbol{\Gamma}}) - \text{vec}(\hat{\boldsymbol{\mu}}) \\ &= \text{vec}((1 + \epsilon_1)\boldsymbol{Z}\hat{\boldsymbol{\Gamma}}) - \text{vec}(\hat{\boldsymbol{\mu}}) \\ &= \text{vec}(\boldsymbol{Z}\hat{\boldsymbol{\Gamma}}) - \text{vec}(\hat{\boldsymbol{\mu}}) + O(\text{vec}(\epsilon_1\boldsymbol{Z}\hat{\boldsymbol{\Gamma}})),\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{v}_P &= (\boldsymbol{I}_{km-m} - \text{vec}(\hat{\boldsymbol{\mu}})\text{vec}(\hat{\boldsymbol{\mu}})^T)\text{vec}(\boldsymbol{Z}\hat{\boldsymbol{\Gamma}}) \\ &= \text{vec}(\boldsymbol{Z}\hat{\boldsymbol{\Gamma}}) - \text{vec}(\hat{\boldsymbol{\mu}})(1 + \text{vec}(\hat{\boldsymbol{\mu}}^T)\boldsymbol{\epsilon}_2) \\ &= \text{vec}(\boldsymbol{Z}\hat{\boldsymbol{\Gamma}}) - \text{vec}(\hat{\boldsymbol{\mu}}) + O(\text{vec}(\hat{\boldsymbol{\mu}}^T)\boldsymbol{\epsilon}_2),\end{aligned}$$

so clearly as ϵ_1 and $\boldsymbol{\epsilon}_2$ are small then $\boldsymbol{v}_R$ and $\boldsymbol{v}_P$ are approximately equal. And so we can see all three tangent coordinates are just as appropriate to use in the low variability case as one another.

Relation between tangent coordinates for data with high variability

To the best of our knowledge there is no literature comparing the use of the three tangent coordinates for higher variability data, perhaps because in practice shape data often have low variability, and so any choice is suitable, however the motivating enzyme data, described in Example 6.3.1, showed higher variability. It turns out for higher variability data the choice of tangent coordinates is important when performing shape PCA, as $\boldsymbol{v}_R$ are not suitable to use for higher variability data. The process of shape PCA can be found in detail in Section 7.7 of Dryden and Mardia (2016). Just like the PCA we have defined for graph Laplacians in Section 2.5, in shape PCA shapes are projected onto the tangent plane, then standard PCA is performed on these. The results of this are projected back into the pre-shape space then results can easily be visualised back in configuration space. In shape PCA there is a question of what tangent coordinates to use in the projection and this does not appear to have an answer, with Kent (1994) using $\boldsymbol{v}_P$ and Cootes et al. (1992) using $\boldsymbol{v}_R$.

When performing shape PCA the number of PCs with non-zero eigenvalues when using $\boldsymbol{v}_P$ and $\boldsymbol{v}_E$ will be at most q if $q \geq n-1$, defined in Equation 6.2.2, or otherwise there will be $n-1$. When using $\boldsymbol{v}_R$ there are at most $q+1$ if $q+1 \geq n-1$ or otherwise $n-1$. Using $\boldsymbol{v}_R$ gives one more non-zero eigenvalue than the other tangent coordinates due to their extra dimension.

Figure 6.2 shows a schematic for finding $\boldsymbol{v}_R$ for data with quite a high variability. The $\boldsymbol{v}_R$s are pulled in close to the sample mean vector and hence for high variability data it is this vector that will dominate the first PCs. So we expect that for higher variability data when $\boldsymbol{v}_R$ are used shape PCA will give first PCs that are just indications of the

Procrustes distance from the sample mean to the shape, as the first PC becomes the mean vector coloured red in Figure 6.2. Obviously this is undesirable because shape PCA is being dominated by the sample mean when we want to remove its effect. The PCA is not taking into account as much information on landmark configurations so information from this is being lost. We use the `shapes` package in R (Dryden, 2018) throughout to perform the shape PCA.

To investigate the effect of using $\boldsymbol{v}_R$ for high variability data the plots of PC scores, found in Section 6.3, are coloured by a configuration's Riemannian distance, ρ, to the mean shape, defined in Equation 6.1.4. This colouring is equivalent to colouring by the full Procrustes distance to the mean shape, d_F, for $\rho \leq \frac{\pi}{2}$, seen using (6.1.7). The condition $\rho \leq \frac{\pi}{2}$ has been checked and met for each dataset. These plots then show how the full Procrustes distance affects the PC scores, for each type of tangent coordinate, so we can see if the distances from the mean is having an effect when $\boldsymbol{v}_R$ are used.

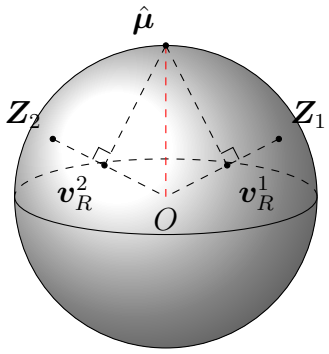


Figure 6.2: *Residual tangent coordinates for data with a large variability.*

It should be noted that shape PCA may not always be the best method for very high variation using any of the tangent coordinates and methods such as Geodesic PCA, found in Huckemann et al. (2010), principal Geodesic analysis (PGA) found in Fletcher and Joshi (2004) or Barycentric subspace analysis (BSA) found in Pennec et al. (2018) may be more appropriate. These methods adapt PCA onto a manifold, however they are often more computer intensive and so we do not consider them now.

6.3 Comparison of shape tangent coordinates for shape data

The motivating data to consider the different tangent coordinates for high variability data was the enzyme data introduced in Section 1.3.5, with $k = 88$, $m = 3$ and $n = 4216$ different times, which we use in Example 6.3.1. Just from exploring the data it is clear this data displays high variability, for example in Figure 1.4 which shows some example landmark configurations for the enzyme data that seen very varied.

To check the choice of tangent coordinate is not important for lower variability data we use three other datasets in Example 6.3.2. These are the Ape skull data, DNA data and sand grain data, introduced in Section 1.3.5. These three datasets all have low variability. Finally we consider a simulation study were we can control the variability of the shapes and see the effect this has on shape PCA for the three different tangent coordinates.

Example 6.3.1: Comparing tangent coordinates for the enzyme data
For the enzyme data we perform shape PCA. Plots for the PC 1 and PC 2 raw scores for each set of tangent coordinates are shown in Figure 6.3. The graphs show a clear difference between using $\boldsymbol{v}_R$ compared to $\boldsymbol{v}_P$ or $\boldsymbol{v}_E$. There is a very well defined convex hull for the PC scores when $\boldsymbol{v}_R$ is used but a far less defined one for both $\boldsymbol{v}_P$ and $\boldsymbol{v}_E$. For all the different tangent coordinates the variance explained by the PCs were very similar and for all tangent coordinates only a small fraction of the PCs are needed to explain a large amount of the variance. Only 9 PCs are needed to explain around 80% of the variance when using each tangent coordinate.

As stated in Section 6.2 the PC plots in Figure 6.3 are coloured to compare the effect a configuration's Procrustes distance from the mean has on its PC score for different tangent coordinates, additional plots for further PCs using $\boldsymbol{v}_R$ are also included. From Figure 6.3 it can be seen, by the red points, that when $\boldsymbol{v}_P$ and $\boldsymbol{v}_E$ are used configurations closer to the mean shape are located near the origin hence have PC scores near to 0. This is not surprising and can be interpreted as the sample Procrustes mean of the data is located near the arithmetic mean of the tangent coordinates; this is partly explained by Chapter 7 of Dryden and Mardia (2016) which states that after optimal full Procrustes positioning has been carried out the full Procrustes mean is equal to the arithmetic mean of each coordinate. However when $\boldsymbol{v}_R$ are used configurations closest to the mean have

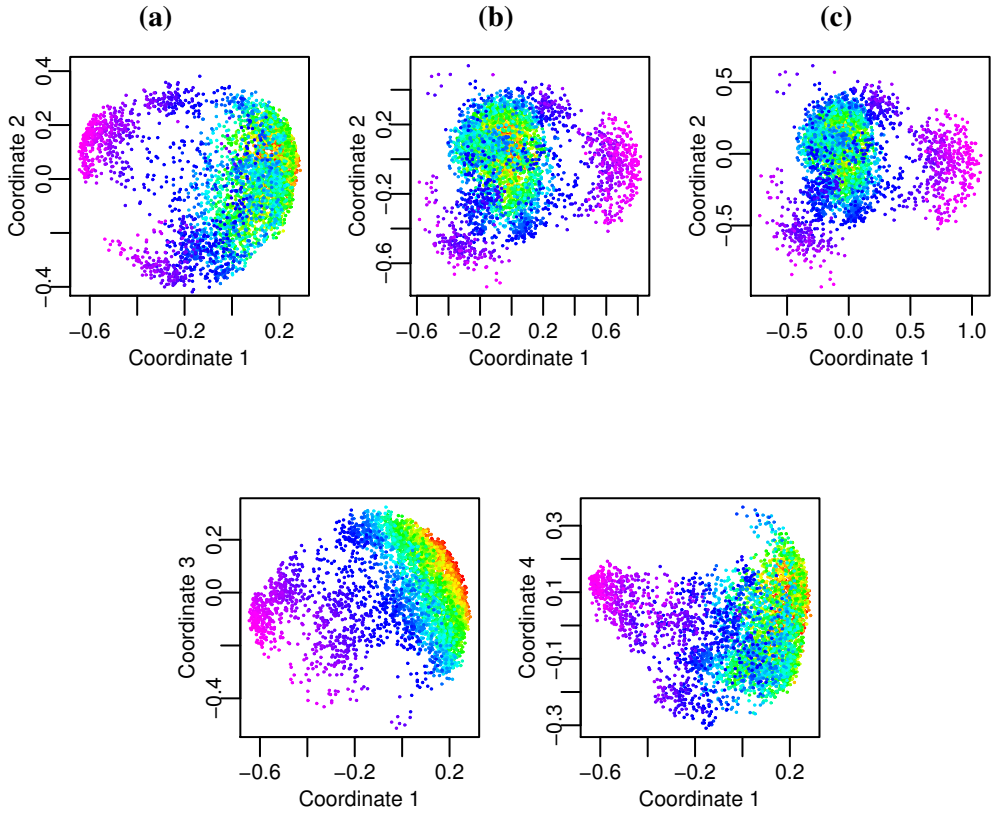


Figure 6.3: *(top) Graphs of PC 1 vs PC 2 raw scores for the enzyme data using (left to right) $\boldsymbol{v}_R$, $\boldsymbol{v}_P$ and $\boldsymbol{v}_E$. (bottom) Graphs of PC 1 scores against d) PC 3 scores and e) PC 4 scores when using $\boldsymbol{v}_R$. Coloured by the Riemannian distance to the mean shape, Equation 6.1.4, with red showing closest to mean shape.*

a higher PC 1 score. There also is a correlation between the Procrustes distance from the mean and PC 1 and 3 scores, seen in Figure 6.3, shown by the band of red.

We look at the modulus of the correlation coefficient between the full Procrustes distance and the first PC score to see how correlated these are for each tangent coordinate. The correlation coefficients are 0.920, 0.531 and 0.547 for $\boldsymbol{v}_R$, $\boldsymbol{v}_P$ and $\boldsymbol{v}_E$ respectively. The correlation coefficient is very near 1 for $\boldsymbol{v}_R$ and not that near it in the other two cases confirming the full Procrustes distance is highly correlated to the first PC score when $\boldsymbol{v}_R$ are used. This supports the reasoning in Section 6.2 of why $\boldsymbol{v}_R$ is not appropriate for higher variability datasets, which is that some of the PCs are highly linked to the Procrustes distance to the mean. Hence the information gained from shape PCA is less relevant and the use of $\boldsymbol{v}_R$ is not suitable for this data.

Example 6.3.2: Comparing tangent coordinates for the low variation data
We now look at the three tangent coordinates for lower variability datasets to confirm they are all close to being equivalent in this case, as reasoned for low variability data in general in Section 6.2. Using the residual tangent coordinates, $\boldsymbol{v}_R$ should give a good approximation to a tangent space when there is little variation from configurations to the mean shape like in this data. Figure 6.4 shows results of shape PCA for the apes, DNA and sand data. These show little difference between all three sets of tangent coordinates, and all have configurations that are close to the mean shape having PC 1 and 2 scores close to the zero. For the ape data there is a difference in the PC score plots between $\boldsymbol{v}_R$ and the other two tangent coordinates, however they are just a mirror images of each other and therefore provide identical information. All three tangent coordinates appear approximately equivalent and $\boldsymbol{v}_R$ are indeed suitable to use for lower variability datasets.

6.3.1 Simulation study

As seen for the enzyme data, which has high variability, the use of residual tangent coordinates was not suitable, we want to see now if this is the case for other cases of high variability data. For synthetic data we can precisely control the shape variability and see the effect this has when using the three different tangent coordinates. We use three models, defined below, for simulating data, in each case with $k = 8$, $m = 3$ and $n = 4000$, with $\boldsymbol{X}_i$ the i^{th} configuration.

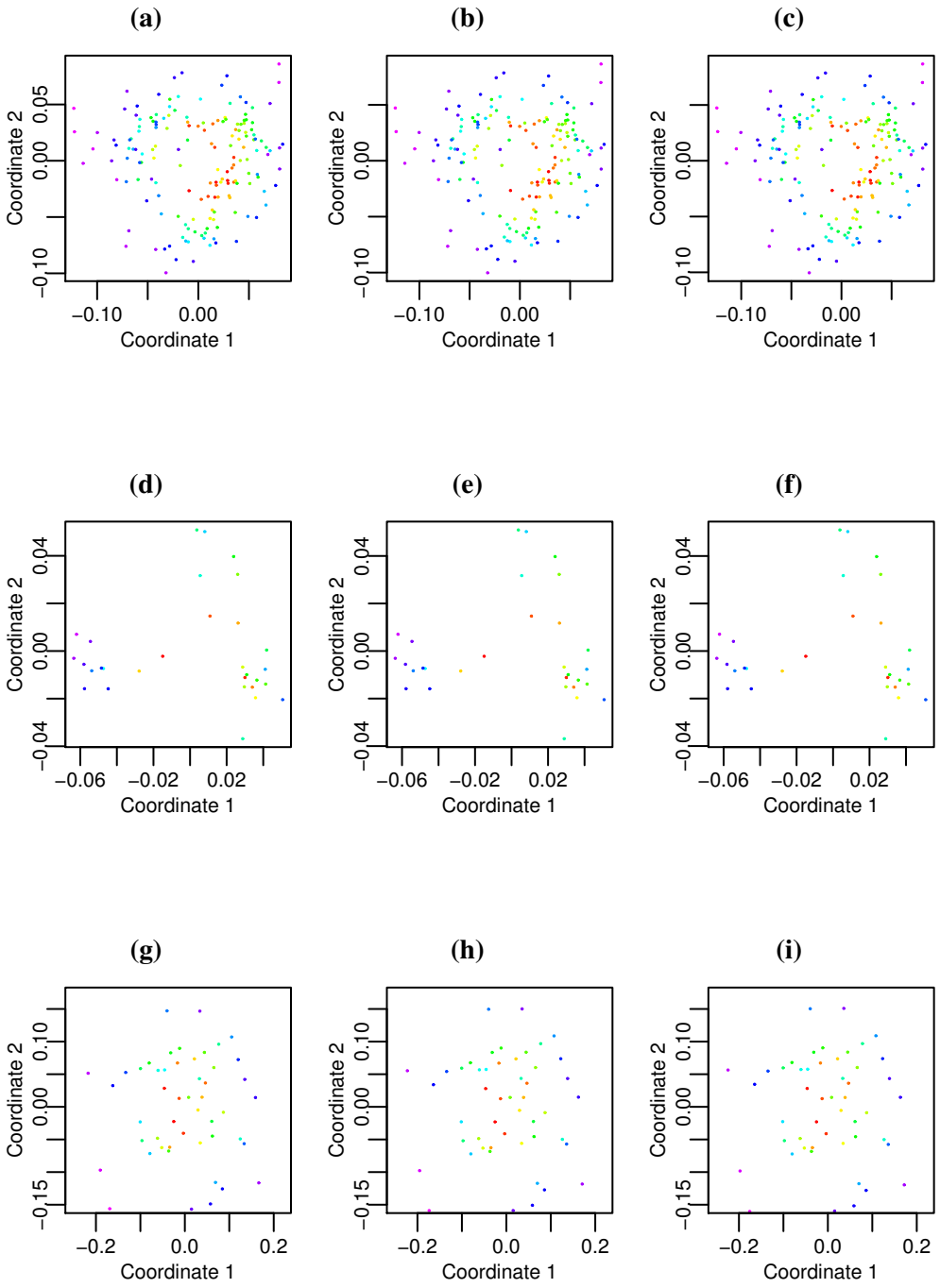


Figure 6.4: *Graphs of PC 1 and 2 scores for the (top to bottom) ape data, DNA data and sand data using (left to right) $\boldsymbol{v}_R$, $\boldsymbol{v}_P$ and $\boldsymbol{v}_E$. Coloured by the Riemannian distance to the mean shape, Equation 6.1.4, with red showing closest to mean shape.*

Model 1 :

$$\text{vec}(\boldsymbol{X}_i) \sim N(\text{vec}(\boldsymbol{c}), \sigma^2 \boldsymbol{I}_{mk})$$
$$\text{where } \boldsymbol{c} = \begin{pmatrix} (0, & 0, & 0) \\ (1, & 0, & 0) \\ & \ldots & \\ (0, & 1, & 1) \end{pmatrix},$$

$\boldsymbol{c}$ is the matrix of the coordinates of the vertices of a unit cube.

Model 2 :

$$\text{vec}(\boldsymbol{X}_i) \sim N(\text{vec}(\boldsymbol{c}), \boldsymbol{\Sigma})$$
$$\text{where } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 \boldsymbol{I}_m & \boldsymbol{0} \\ \boldsymbol{0} & \alpha^2 \boldsymbol{I}_{k(m-1)} \end{pmatrix},$$

with $\boldsymbol{c}$ defined as above.

Model 3 :

$$\boldsymbol{X}_1 = \boldsymbol{c}$$
$$\boldsymbol{X}_i = p\boldsymbol{X}_{i-1} + \boldsymbol{\epsilon}$$
$$\text{vec}(\boldsymbol{\epsilon}) \overset{iid}{\sim} N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{mk}),$$

this an autoregressive model of order 1, denoted AR(1).

Simulated data plots of PC 1 and PC 2 raw scores are coloured by the Riemannian distance from the mean for each tangent coordinate, to see the relationship between these, found in Figure 6.5. Model 1 has $\sigma^2 = 1$, Model 2 with $\alpha^2 = 0.1$ and $\sigma^2 = 5$ and Model 3 where $p = 0.999$ and $\sigma^2 = 1$. The cumulative variance plots are not included but in each dataset only a small fraction of PCs explain a large amount of the variance. The PC score plots show a similar effect as the enzyme data, when using $\boldsymbol{v}_P$ and $\boldsymbol{v}_E$, configuration closer to the mean are near the origin. When $\boldsymbol{v}_R$ are used the PC 1 scores seems to be linked to the configurations Riemannian distance from the mean shape. This effect is very extreme for Model 1 and 2, where the first PC seems to actually just be an indication of the Procrustes distance between a configuration and the mean shape. For Model 3 this effect from using $\boldsymbol{v}_R$ is less but it is still clear the shapes closer to the

mean are not centred around the origin; they tend to have a negative PC 1 and 2 score, showing the first PCs for a configuration are linked with the Procrustes distance from the mean. Whereas the plots for $\boldsymbol{v}_P$ and $\boldsymbol{v}_E$ have red points more centred around the origin.

Just as with the enzyme data we look at the modulus of the correlation coefficients between full Procrustes distance and the first PC score for $\boldsymbol{v}_R$, $\boldsymbol{v}_P$ and $\boldsymbol{v}_E$. For Model 1 these are 0.989, 0.017 and 0.016 and for Model 2 they are 0.948, 0.003 and 0,002. This confirms the correlation is only when using $\boldsymbol{v}_R$ for Model 1 and 2. It is clear from Figure 6.5g that for Model 3 the relationship between the full Procrustes distance and PC 1 score is not linear when using $\boldsymbol{v}_R$ and so the correlation coefficient will not tell us anything meaningful in this case.

Further investigation into the relation between PC1 and the mean shape

We have seen when using $\boldsymbol{v}_R$ that increasing variability leads to the first PC being just an indication of a configurations Procrustes distance from the mean, and so the first PC vector becomes the mean vector, seen as the red line in Figure 6.2. To look at this effect in more detail we look at the Cosine similarity between the Procrustes mean shape and first PC vector for the simulated data in Model 2 when using $\boldsymbol{v}_R$, for 4000 configurations. The cosine similarity between the sample Procrustes mean, $\hat{\boldsymbol{\mu}}$ and a vector $\boldsymbol{v}$ is defined as

$$\cos(\theta) = \frac{\hat{\boldsymbol{\mu}} \cdot \boldsymbol{v}}{\|\hat{\boldsymbol{\mu}}\|\|\boldsymbol{v}\|}.$$

A cosine similarity near 1 or -1 shows a strong similarity whereas one near 0 shows little similarity, as the sign is irrelevant it is the absolute value of the cosine that we look at.

Figure 6.6 shows how the cosine similarity changes as the variability (σ^2) increases for Model 2 with α^2 set to 0.1. We see that generally as the variability increases the absolute cosine similarity tends to increase and eventually tends to 1 indicating for high variability the first PC is just dominated by the sample full Procrustes mean shape, and hence using $\boldsymbol{v}_R$ is not suitable in theses cases. As the variability decreases the cosine similarity overall is decreasing showing there is less link between the first PC vector and the mean shape and so $\boldsymbol{v}_R$ would be appropriate to use in this case. The bump in Figure 6.6 is unexpected as we expected the cosine to be increasing with σ^2, which clearly is

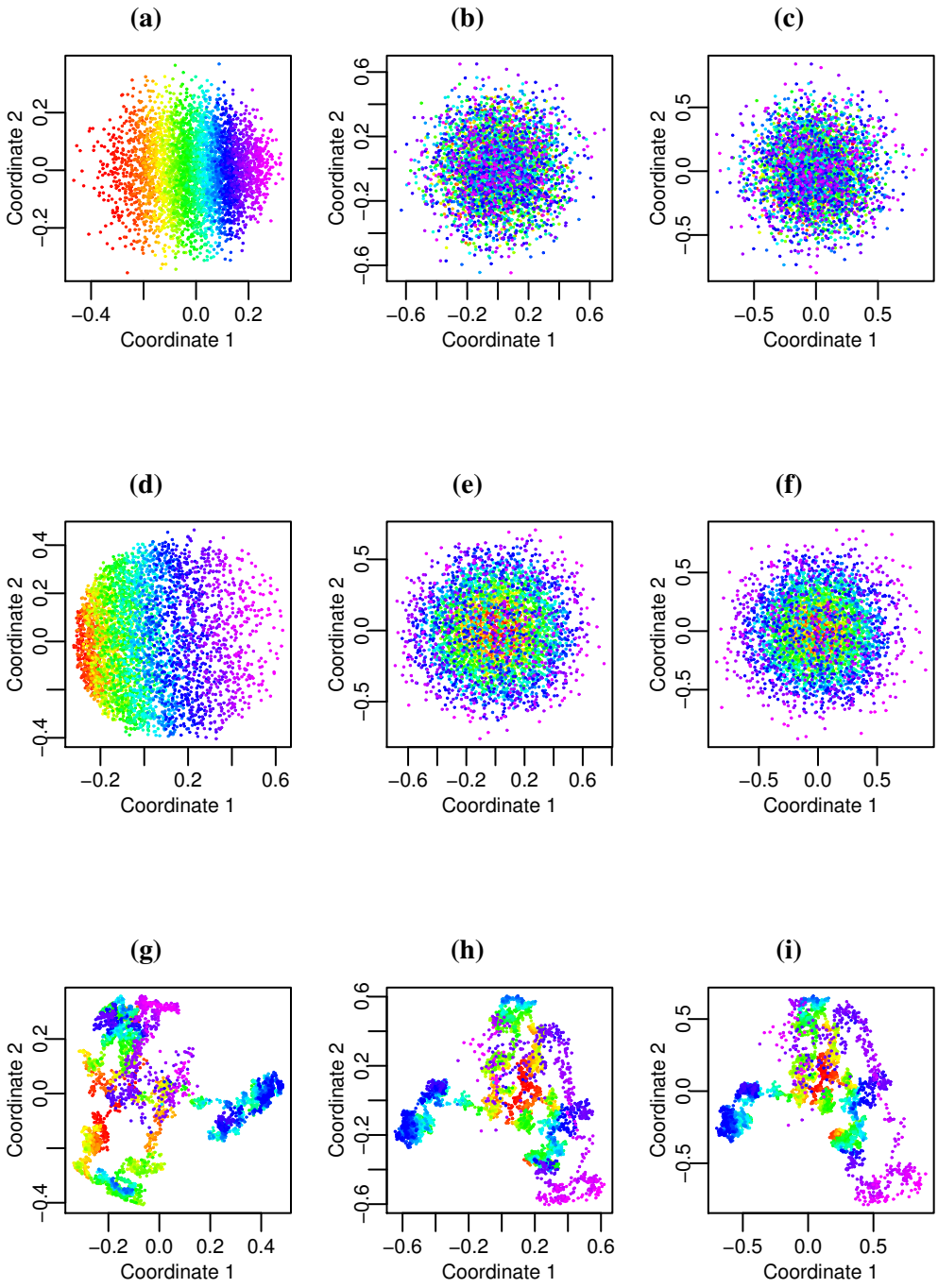


Figure 6.5: *Graphs of PC 1 and 2 scores for simulated data for Models (top to bottom) 1,2 and 3 using (left to right) $\boldsymbol{v}_R$, $\boldsymbol{v}_P$ and $\boldsymbol{v}_E$. Coloured by the Riemannian distance to the mean shape, Equation 6.1.4, with red showing closest to mean shape.*

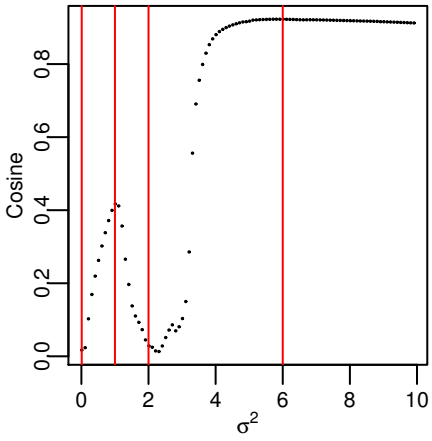


Figure 6.6: *Absolute cosine similarity between the mean and PC 1 vector as variability of configurations is altered for Model 2. Vertical lines at* $\sigma^2 = 0.01, 1, 2$ *and* 6.

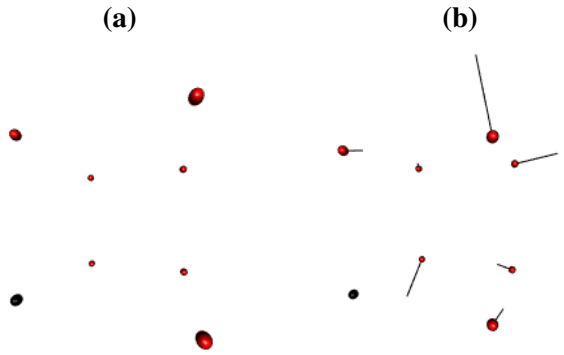
not true for σ^2 values between 1 and 2.

The plot in Figure 6.6 is marked with vertical red lines at σ^2 values that we have then used to produce 3D plots of sample Procrustes means and the first PC vector, found in Figure 6.7. In these 3D plots the black point represents the landmark with variance σ^2 whilst the red points are those with variance $\alpha^2 = 0.1$. These plots support a theory as follows that may explain the bump. For low σ^2 the sample Procrustes mean is similar to the population mean, a cube, seen in Figures 6.7a and 6.7c, when $\sigma^2 = 0.01$ and 1. As σ^2 increases variation from this mean increases hence the same effect seen multiple times before occurs; the first PC vector is dominated by the mean vector, which explains the first increase on the cosine graph. The dip may be as the sample mean shape around $\sigma^2 = 1$ starts to change until by $\sigma^2 = 2$ it is 7 landmarks getting closer together and the one landmark with σ^2 variance is much further seen in Figure 6.7e and more pronounced by 6.7g when $\sigma^2 = 6$. So as the mean is changing the first PC vector is not remaining similar to it, hence the drop in cosine score. We believe by $\sigma^2 = 2$ there is a new sample mean shape, from here the variance is increasing hence the cosine is tending to 1. Whilst the exact reasoning behind the cosine values especially the bump is still unclear it is clear the use of $\boldsymbol{v}_R$ leads to the first PC vector to be dominated by the mean at some instances and using $\boldsymbol{v}_R$ are not suitable in these instances.

6.4 Summary

We have seen that for high variability data, residual tangent coordinates, $\boldsymbol{v}_R$, give very different results in shape PCA to the other two tangent coordinates, the partial tangent coordinates, $\boldsymbol{v}_P$, and the inverse exponential tangent coordinates, $\boldsymbol{v}_E$. In these cases using $\boldsymbol{v}_R$ gives the first PC as just an indication of a configurations Riemannian distance to the mean and this makes their use unsuitable, as landmark configuration information is taken less into account and the effect of the mean is not being removed. Using $\boldsymbol{v}_P$ and $\boldsymbol{v}_E$ give very similar results throughout and therefore for high variability data only $\boldsymbol{v}_P$ and $\boldsymbol{v}_E$ should be used and never $\boldsymbol{v}_R$. For lower variability data all three tangent coordinates are approximately equivalent and all are suitable for use.

No definite measure exists on when a dataset is too variable for $\boldsymbol{v}_R$ to be suitable and so currently the only test is by comparison between them and the other two tangent coordinates, therefore we suggest if there is doubt on a dataset's variability it is best to use $\boldsymbol{v}_P$ or $\boldsymbol{v}_E$s from the start, so the comparison is not necessary. Further work would be looking at more empirical conditions on what we mean by a dataset being 'too varied' for $\boldsymbol{v}_R$ to be appropriate.



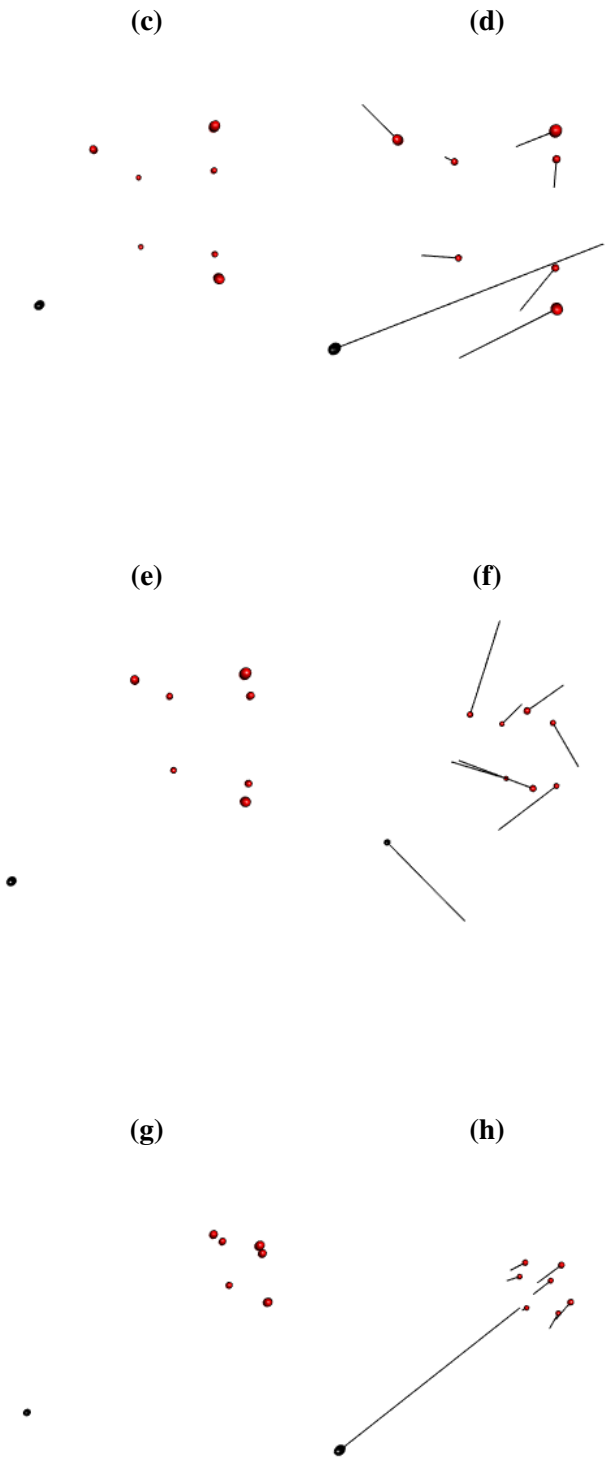


Figure 6.7: *3D plots for model 2 with (top to bottom)* $\alpha^2 = 0.1$ *and* $\sigma^2 = 0.01$, $\alpha^2 = 0.1$ *and* $\sigma^2 = 1$, $\alpha^2 = 0.1$ *and* $\sigma^2 = 2$, *and* $\alpha^2 = 0.1$ *and* $\sigma^2 = 6$ *of the (left to right) sample full Procrustes mean and sample full Procrustes mean with the first PC vector plotted on.*

CHAPTER 7

Conclusion

In this work we have developed a novel framework for the statistical analysis of networks by representing them as graph Laplacians. With this framework we defined two general metrics between graph Laplacians, the Euclidean power metric and the Procrustes power metric, and developed for network data analogues of many methods of classical multivariate analysis such as calculating means, interpolating and extrapolating and performing PCA.

The framework however, remains general for metrics and the use of other metrics between graph Laplacians could be considered. For example a metric to consider in future work is the log-Euclidean metric defined for $\boldsymbol{L}_1, \boldsymbol{L}_2 \in \mathcal{L}_m$ as

$$d_{log}(\boldsymbol{L}_1, \boldsymbol{L}_2) = \| \log(\boldsymbol{L}_1) - \log(\boldsymbol{L}_2) \|.$$

where the log of the graph Laplacian $\boldsymbol{L} = \boldsymbol{U} \begin{pmatrix} \boldsymbol{\Lambda}' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \boldsymbol{U}^T$ is defined as $\log(\boldsymbol{L}) = \boldsymbol{U} \begin{pmatrix} \log(\boldsymbol{\Lambda}') & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \boldsymbol{U}^T$. This metric is of interest to consider as logarithm-based metrics have been used previously to interpolate between graph Laplacians in Bakker et al. (2018).

Using the graph Laplacian framework we also explored different regression models. When we took the graph Laplacian as the response with Euclidean covariates we defined a linear regression model and Nadaraya-Watson model to estimate a graph Laplacian. We also used a Nadaraya-Watson model to predict Euclidean responses from Graph Laplacian covariates. We often considered the covariate being a scalar, for example time, and we therefore investigated the horseshoe effect present in PCA and MDS plots

for graph Laplacians with a time structure. We developed a new method for visualising graph Laplacians removing this horseshoe effect. There is an increasing number of temporal network data, for example in Friel et al. (2016), which considers temporal networks representing the connection of leading Irish companies and board directors, and in Dubey and Mueller (2019) where one example includes temporal networks representing the taxi trips in New York. As temporal network datasets become more common future work will involve applying our regression models to more temporal network data to confirm it works consistently well for a range of data. We also considered the case with multidimensional covariates, e.g. spatial coordinates, for graph Laplacians and in this case we defined an adaptation of Kriging to predict a graph Laplacian from known spatial coordinates.

We have developed a two-sample test to test equality of means for samples of graph Laplacians. We focused on using the metrics d_1, $d_{\frac{1}{2}}$ and $d_{\frac{1}{2},S}$, giving test statistics T_E, T_H and T_S, these all performed well when using a permutation test, however using T_E was significantly faster as no square rooting of the graph Laplacian or optimising rotation was needed and therefore d_1 seemed to be the better metric to use. The test could be easily altered to facilitate using different metrics. We also compared all these test statistics to one previously defined in Ginestet et al. (2017), T_G. The test statistic T_G required the estimation of a large covariance matrix and so we saw when m was large, approximately over 40, the estimation of the covariance matrix was poor and T_G is not advisable to use. We also provided a method of studying why the means between networks differed which we applied to the novel data. Understanding differences between bodies of text is an interesting challenge in corpus linguistics. Further work should compare our method with the methods of comparing differences in text used as standard in corpus linguistics. We can also apply our new method to interesting corpus linguistic questions, such as how does character speech/quotes differ to narration/non-quotes in novels (Mahlberg and Wiegand, 2018).

We provided two methods for classifying graph Laplacians belonging to binary classes. One method took place in the manifold whilst the other used PC scores. Both performed well although the classification in the manifold performed slightly better. Further work should consider when graph Laplacians can belong to more than 2 classes, for the classification using PC scores this should be a relatively simple adaptation. Adapting deep learning classification methods, such as convolutional neural networks, for manifold-valued data could be considered to see how this compares with the methods we have

already proposed. Some work has been done with deep learning for manifold-valued data, for example in Chakraborty et al. (2018), but there is a lot of scope to expand this. We also provided a method to detect anomalies.

Throughout we have compared the Euclidean metric, d_1, the square root Euclidean metric, $d_{\frac{1}{2}}$, and the Procrustes size-and-shape metric, $d_{\frac{1}{2},S}$. We have seen using $d_{\frac{1}{2}}$ and $d_{\frac{1}{2},S}$ often gives visibly identical results. The results between d_1 and $d_{\frac{1}{2}}$ for the examples we have looked at are generally only slightly different and so neither metric seems advantageous over the other, except d_1 is computationally more appealing. Investigating the differences in these metrics for more datasets would be a useful next step to see if $d_{\frac{1}{2}}$ and $d_{\frac{1}{2},S}$ are advantageous in some cases, as they are when used for symmetric positive semi-definite matrices in Dryden et al. (2009). One advantage we hypothesise is that $d_{\frac{1}{2}}$ and $d_{\frac{1}{2},S}$ can distinguish differences in network connectivity better than d_1. To illustrate this advantage we provide examples in Figure 7.1 and Table 7.1.

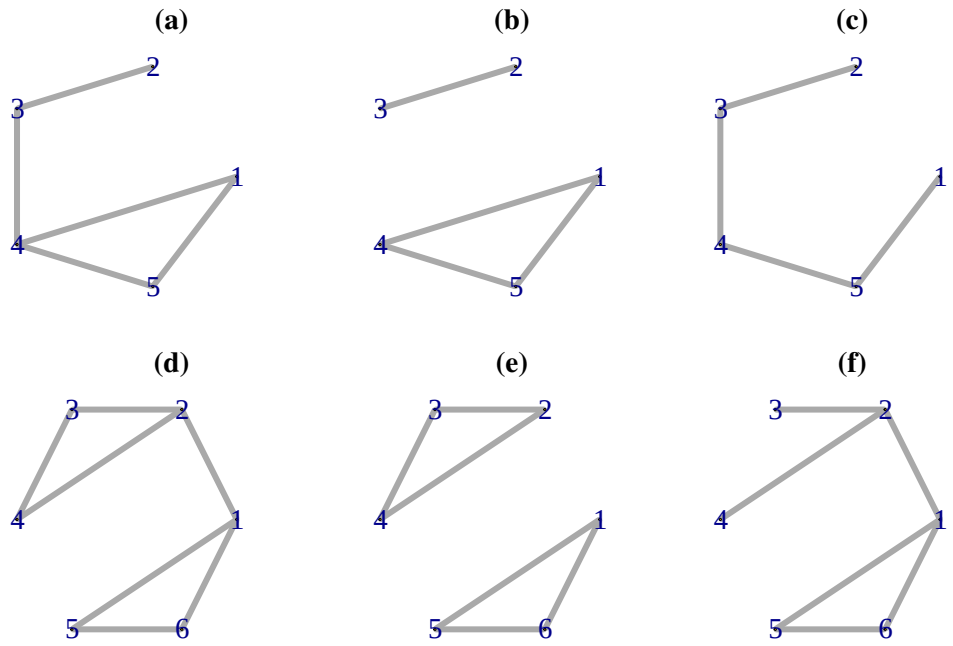


Figure 7.1: *(top) 5 node networks and (bottom) 6 node networks used to compare the Eucldiean, square root Euclidean and Procrustes size-and-shape metrics in Table 7.1.*

The connected network (a) in Figure 7.1 has had an edge deleted to form both networks (b) and (c), however (b) is now disconnected whilst (c) is still connected. From Table 7.1 d_1 gives identical distances between (a) and both (b) and (c) however $d_{\frac{1}{2}}$ and $d_{\frac{1}{2},S}$ give a larger distance between (a) and (b), for certain application this could be advantageous as the connectivity of the networks is taken into account in these metrics. Similarly for

Distance between		d_1	$d_{\frac{1}{2}}$	$d_{\frac{1}{2},S}$
(a)	(b)	2	1.027	0.971
(a)	(c)	2	0.687	0.675
(d)	(e)	2	0.940	0.888
(d)	(f)	2	0.732	0.732

Table 7.1: *Comparing the Eucldiean, square root Euclidean and Procrustes size-and-shape metrics for the networks in Figure 7.1*

(d) in Figure 7.1 an edge is deleted to create a disconnected network, (e), and connected one, (f). Table 7.1 shows the same effect that $d_{\frac{1}{2}}$ and $d_{\frac{1}{2},S}$ give a larger distance between the connected and disconnected network. For the analysis we have done the connectivity is not a property of interest, as it is not something we have interpreted for our applications, hence the proposed advantage of $d_{\frac{1}{2}}$ and $d_{\frac{1}{2},S}$ distinguishing differences in networks connectivity is not seen as advantageous in our applications. However this advantage for $d_{\frac{1}{2}}$ and $d_{\frac{1}{2},S}$ could be advantageous in other application which may be of interest to study (Bao et al., 2018, Section 4.1). This is just one possible difference between the metrics and we expect as the novel framework for the statistical analysis of networks is used with these metrics for a wider array of applications and dataset more differences and advantages between metrics will become apparent. We can then provide guidelines as to which metric one should use based on the application.

Finally we looked at a different application of manifold-valued data analysis, namely shape analysis. In this application we used the shape space and compared several tangent coordinates that are commonly used. We found that the residual tangent coordinates that are commonly used by practitioners are not suitable for datasets with large variation. It would be useful to define empirical conditions on when a dataset is too varied to use the residual tangent coordinate, so we can guide practitioners which tangent coordinate to use.

References

Agarwal, A., Omuya, A., Harnly, A., and Rambow, O. (2012). A comprehensive gold standard for the Enron organizational hierarchy. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 161–165, Stroudsburg, PA, USA. Association for Computational Linguistics.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of machine learning research*, 9(Sep):1981–2014.

Akoglu, L., Tong, H., and Koutra, D. (2015). Graph based anomaly detection and description: a survey. *Data mining and knowledge discovery*, 29(3):626–688.

Amaral, G. J. A., Dryden, I. L., and Wood, A. T. A. (2007). Pivotal bootstrap methods for k-sample problems in directional statistics and shape analysis. *Journal of the American Statistical Association*, 102(478):695–707.

Anderson, E. (2018). *rosqp: Quadratic Programming Solver using the 'OSQP' Library*. R package version 0.1.0.

Antiqueira, L., Pardo, T. A. S., Nunes, M. d. G. V., and Oliveira Jr, O. N. (2007). Some issues on complex networks for author characterization. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, 11(36):51–58.

Arsigny, V., Fillard, P., Pennec, X., and Ayache, N. (2007). Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM journal on matrix analysis and applications*, 29(1):328–347.

Bakker, C., Halappanavar, M., and Sathanur, A. V. (2018). Dynamic graphs, community detection, and Riemannian geometry. *Applied Network Science*, 3(1):3.

Banerjee, A. and Jost, J. (2008). On the spectrum of the normalized graph Laplacian. *Linear algebra and its applications*, 428(11-12):3015–3022.

Bao, D., You, K., and Lin, L. (2018). Network distance based on Laplacian flows on graphs. *arXiv preprint arXiv:1810.02906*.

Bhattacharya, R. and Patrangenaru, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds. *Ann. Statist.*, 31(1):1–29.

Bhattacharya, R. and Patrangenaru, V. (2005). Large sample theory of intrinsic and extrinsic sample means on manifolds II. *Ann. Statist.*, 33(3):1225–1259.

Bierens, H. (1988). The Nadaraya-Watson kernel regression function estimator. Working Paper 1988-58, Faculty of Economics and Business Administration, Vrije Universiteit Amsterdam.

Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., Beckmann, C. F., Adelstein, J. S., Buckner, R. L., Colcombe, S., et al. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10):4734–4739.

Bookstein, F. L. (1978). *The measurement of biological shape and shape change*. Springer.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, i. effect of inequality of variance in the one-way classification. *Ann. Math. Statist.*, 25(2):290–302.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Bunke, H., Dickinson, P. J., Kraetzl, M., and Wallis, W. D. (2007). *A graph-theoretic approach to enterprise network dynamics*, volume 24. Springer Science & Business Media.

Burrows, J. F. (1987). *Computation into criticism: A study of Jane Austen's novels and an experiment in method*. Clarendon Pr.

Chakraborty, R., Bouza, J., Manton, J. H., and Vemuri, B. C. (2018). Manifoldnet: A deep network framework for manifold-valued data. *CoRR*, abs/1809.06211.

Charles Dickens Info (2018). Charles Dickens timeline. `https://www.charlesdickensinfo.com/life/timeline/`, Last accessed on 2018-11-12.

Chen, H., Perozzi, B., Al-Rfou, R., and Skiena, S. (2018). A tutorial on network embeddings. *arXiv preprint arXiv:1808.02590*.

Chilès, J.-P. and Desassis, N. (2018). *Fifty Years of Kriging*, pages 589–612. Springer International Publishing, Cham.

Chung, F. R. (1997). *Spectral graph theory*. Number 92. American Mathematical Soc.

Conte, D., Foggia, P., Sansone, C., and Vento, M. (2004). Thirty years of graph matching in pattern recognition. *International journal of pattern recognition and artificial intelligence*, 18(03):265–298.

Cootes, T., Taylor, C., Cooper, D., and Graham, J. (1992). *Training models of shape from sets of examples*. Springer Verlag, Germany.

Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1994). Image search using flexible shape models generated from sets of examples. In Mardia, K. V., editor, *Statistics and Images: Vol. 2*, pages 111–139. Carfax, Oxford.

Coulthard, M. (2004). Author identification, idiolect, and linguistic uniqueness. *Applied linguistics*, 25(4):431–447.

Cullinane, M. J. (2011). Metric axioms and distance. *The Mathematical Gazette*, 95(534):414–419.

da Fontoura Costa, L., Jr., O. N. O., Travieso, G., Rodrigues, F. A., Boas, P. R. V., Antiqueira, L., Viana, M. P., and Rocha, L. E. C. (2011). Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics*, 60(3):329–412.

Davis, B. C., Fletcher, P. T., Bullitt, E., and Joshi, S. (2010). Population shape regression from random design data. *International Journal of Computer Vision*, 90(2):255–266.

De Klerk, E. (2006). *Aspects of semidefinite programming: interior point algorithms and selected applications*, volume 65. Springer Science & Business Media.

Diaconis, P., Goel, S., Holmes, S., et al. (2008). Horseshoes in multidimensional scaling and local kernel methods. *The Annals of Applied Statistics*, 2(3):777–807.

Diesner, J. and Carley, K. M. (2005). Exploration of communication networks from the Enron email corpus 1.

Diesner, J., Frantz, T. L., and Carley, K. M. (2005). Communication networks from the Enron email corpus "it's always about the people. Enron is no different". *Computational & Mathematical Organization Theory*, 11(3):201–228.

Dryden, I. L. (2018). *Shapes: Statistical Shape Analysis*. R package version 1.2.4.

Dryden, I. L., Koloydenko, A., and Zhou, D. (2009). Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, 3(3):1102–1123.

Dryden, I. L. and Mardia, K. V. (1993). Multivariate shape analysis. *Sankhya : The Indian Journal of Statistics, Series A (1961-2002)*, 55(3):460–480.

Dryden, I. L. and Mardia, K. V. (2016). *Statistical shape analysis with applications in R*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, second edition.

Dubey, P. and Mueller, H.-G. (2019). Functional models for time-varying random objects. *arXiv preprint arXiv:1907.10829*.

Engelmann, G., Smith, G., and Goulding, J. (2018). The unbanked and poverty: Predicting area-level socio-economic vulnerability from M-money transactions. *2018 IEEE International Conference on Big Data (Big Data)*, pages 1357–1366.

Evert, S. (2008). Corpora and collocations. *Corpus linguistics. An international handbook*, 2:1212–1248.

Fiedler, M. (1973). Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2):298–305.

Fillard, P., Pennec, X., Arsigny, V., and Ayache, N. (2007). Clinical DT-MRI estimation, smoothing, and fiber tracking with log-Euclidean metrics. *IEEE transactions on medical imaging*, 26(11):1472–1482.

Fiori, S. (2009). Learning the Fréchet mean over the manifold of symmetric positive-definite matrices. *Cognitive Computation*, 1(4):279.

Fisher, R. (1953). Dispersion on a sphere. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 217, pages 295–305. The Royal Society.

Fletcher, P. T. and Joshi, S. (2004). *Principal Geodesic analysis on symmetric spaces: Statistics of diffusion tensors*, pages 87–98. Springer Berlin Heidelberg, Berlin, Heidelberg.

Fletcher, P. T. and Joshi, S. (2007). Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Processing*, 87(2):250 – 262. Tensor Signal Processing.

Fletcher, P. T., Lu, C., Pizer, S. M., and Joshi, S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE transactions on medical imaging*, 23(8):995–1005.

Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'institut Henri PoincarÃl'*, 10(4):215–310.

Friel, N., Rastelli, R., Wyse, J., and Raftery, A. E. (2016). Interlocking directorates in Irish companies using a latent space model for bipartite networks. *Proceedings of the National Academy of Sciences*, 113(24):6629–6634.

Fu, A., Narasimhan, B., Diamond, S., and Miller, J. (2018). *CVXR: Disciplined Convex Optimization*. R package version 0.99.

Gao, X., Xiao, B., Tao, D., and Li, X. (2010). A survey of graph edit distance. *Pattern Analysis and Applications*, 13(1):113–129.

Gaston, L. (2016). Gossip economies: Jane Austen, Lady Susan, and the right to self-fashion. *European Romantic Review*, 27(3):405–411.

Ginestet, C. E., Fournel, A. P., and Simmons, A. (2014). Statistical network analysis for functional MRI: summary networks and group comparisons. *Frontiers in Computational Neuroscience*, 8:51.

Ginestet, C. E., Li, J., Balachandran, P., Rosenberg, S., Kolaczyk, E. D., et al. (2017). Hypothesis testing for network data in functional neuroimaging. *The Annals of Applied Statistics*, 11(2):725–750.

Goodall, C. (1991). Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(2):285–339.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338.

Gower, J. C. (1975). Generalized Procrustes analysis. *Psychometrika*, 40(1):33–51.

Gross, J. L. and Yellen, J. (2004). *Handbook of graph theory*. CRC press.

Hammond, D. K., Vandergheynst, P., and Gribonval, R. (2011). Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129 – 150.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85.

Healy, P. M. and Palepu, K. G. (2003). The fall of Enron. *Journal of Economic Perspectives*, 17(2):3–26.

Hennessey, A., Wiegand, V., Mahlberg, M., Tench, C. R., and Lentin, J. (2017). *CorporaCoCo: Corpora Co-Occurrence Comparison*. R package version 1.1-0.

Higham, N. J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear algebra and its applications*, 103:103–118.

Huckemann, S. and Hotz, T. (2009). Principal component geodesics for planar shape spaces. *Journal of Multivariate Analysis*, 100(4):699 – 714.

Huckemann, S., Hotz, T., and Munk, A. (2010). Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric lie group actions. *Statistica Sinica*, 20(1):1–58.

Imhof, J. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48(3/4):419–426.

Joyce, D. (2009). On manifolds with corners. *arXiv preprint arXiv:0910.3518*.

Juola, P. (2015). The Rowling case: A proposed standard analytic protocol for authorship questions. *Digital Scholarship in the Humanities*, 30(suppl_1):i100–i113.

Kendall, D. G. (1970). A mathematical approach to seriation. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 269(1193):125–134.

Kendall, D. G. (1971). Abundance matrices and seriation in archaeology. *Probability Theory and Related Fields*, 17(2):104–112.

Kendall, D. G. (1977). The diffusion of shape. *Advances in Applied Probability*, 9(3):428–430.

Kendall, D. G. (1984). Shape manifolds, Procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society*, 16(2):81–121.

Kendall, D. G., Barden, D., Carne, T. K., and Le, H. (1999). *Shape and Shape Theory*. Wiley, Chichester.

Kendall, W. S. (1990). Probability, convexity, and harmonic maps with small image I: uniqueness and fine existence. *Proceedings of the London Mathematical Society*, 3(2):371–406.

Kent, J. T. (1994). The complex Bingham distribution and shape analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(2):285–299.

Klimt, B. and Yang, Y. (2004). Introducing the Enron corpus. In *CEAS*.

Klopp, O. and Verzelen, N. (2017). Optimal graphon estimation in cut distance. *Probability Theory and Related Fields*, pages 1–58.

Kolaczyk, E., Lin, L., Rosenberg, S., and Walters, J. (2017). Averages of unlabeled networks: Geometric characterization and asymptotic behavior. *arXiv preprint arXiv:1709.02793*.

Kolaczyk, E. D. (2009). *Statistical analysis of network data: methods and models*. Springer Science & Business Media.

Koutra, D., Vogelstein, J. T., and Faloutsos, C. (2013). Deltacon: A principled massive-graph similarity function. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 162–170. SIAM.

LaFarge, L. (2009). The workings of forgiveness: Charles Dickens and David Copperfield. *Psychoanalytic Inquiry*, 29(5):362–373.

Le, H. (1995). Mean size-and-shapes and mean shapes: A geometric point of view. *Advances in Applied Probability*, 27(1):44–55.

Le, H. and Kendall, D. G. (1993). The Riemannian structure of Euclidean shape spaces: A novel environment for statistics. *The Annals of Statistics*, 21(3):1225–1271.

Lee, J. M. (2003). Smooth manifolds. In *Introduction to Smooth Manifolds*, pages 1–29. Springer.

Li, T., Zhu, S., and Ogihara, M. (2006). Using discriminant analysis for multi-class classification: an experimental investigation. *Knowledge and Information Systems*, 10(4):453–472.

Maggie Kopp (2011). Thackeray and Charles Dickens. `https://sites.lib.byu.edu/special-collections/2011/07/21/thackeray-and-charles-dickens/`, Last accessed on 2019-06-10.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. National Institute of Science of India.

Mahlberg, M. (2010). Corpus linguistics and the study of nineteenth-century fiction. *Journal of Victorian Culture*, 15(2):292–298.

Mahlberg, M. (2013). *Corpus stylistics and Dickens's fiction*. Routledge.

Mahlberg, M., Smith, C., and Preston, S. (2013). Phrases in literary contexts. *International Journal of Corpus Linguistics*, 18(1):35–56.

Mahlberg, M., Stockwell, P., de Joode, J., Smith, C., and O'Donnell, M. B. (2016). CLiC Dickens: novel uses of concordances for the integration of corpus stylistics and cognitive poetics. *Corpora*, 11(3):433–463.

Mahlberg, M. and Wiegand, V. (2018). *Corpus stylistics, norms and comparisons: Studying speech in Great Expectations*, pages 123–143.

Mardia, K. V. and Dryden, I. L. (1999). The complex Watson distribution and shape analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(4):913–926.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate analysis*. Academic Press, London.

McDaid, A. F., Murphy, T. B., Friel, N., and Hurley, N. J. (2013). Improved Bayesian inference for the stochastic block model with application to large networks. *Computational Statistics & Data Analysis*, 60:12–31.

Merris, R. (1994). Laplacian matrices of graphs: a survey. *Linear Algebra and its Applications*, 197-198(Supplement C):143 – 176.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81.

Moakher, M. and Zéraï, M. (2011). The Riemannian geometry of the space of positive-definite matrices and its application to the regularization of positive-definite matrix-valued data. *Journal of Mathematical Imaging and Vision*, 40(2):171–187.

Moisl, H. (2015). *Cluster analysis for corpus linguistics*, volume 66. Walter de Gruyter GmbH & Co KG.

Molenaar, W. (1970). *Approximations to the Poisson, Binomial and Hypergeometric distribution functions*. C.

Morton, J. T., Toran, L., Edlund, A., Metcalf, J. L., Lauber, C., and Knight, R. (2017). Uncovering the horseshoe effect in microbial analyses. *MSystems*, 2(1):e00166–16.

Mpogole, H., Tweve, Y., Mwakatobe, N., Mlasu, S., and Sabokwigina, D. (2016). Towards non-cash payments in Tanzania: The role of mobile phone money services. In *2016 IST-Africa Week Conference*, pages 1–11.

Nadaraya, É. (1965). On non-parametric estimates of density functions and regression curves. *Theory of Probability & Its Applications*, 10(1):186–190.

Newman, M. (2010). *Networks: an introduction*. Oxford university press.

Pennec, X. et al. (2018). Barycentric subspace analysis on manifolds. *The Annals of Statistics*, 46(6A):2711–2746.

Pennec, X., Fillard, P., and Ayache, N. (2006). A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66.

Penrose, R. (1955). A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):406âĂŞ413.

Phillips, M. (1983). *Lexical Macrostructure in Science Text*. University of Birmingham.

Pigoli, D., Menafoglio, A., and Secchi, P. (2016). Kriging prediction for manifold-valued random fields. *Journal of Multivariate Analysis*, 145:117–131.

Prabhu, N., Chang, H.-C., and deGuzman, M. (2005). Optimization on Lie manifolds and pattern recognition. *Pattern Recognition*, 38(12):2286 – 2300.

Preston, S. P. and Wood, A. T. A. (2010). Two-sample bootstrap hypothesis tests for three-dimensional labelled landmark data. *Scandinavian Journal of Statistics*, 37(4):568–587.

Preston, S. P. and Wood, A. T. A. (2011). Bootstrap inference for mean reflection shape and size-and-shape with three-dimensional landmark data. *Biometrika*, 98(1):49–63.

Pržulj, N. (2010). Biological network comparison using graphlet degree distribution. *Bioinformatics*, 26(6):853–854.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rastelli, R., Latouche, P., and Friel, N. (2018). Choosing the number of groups in a latent stochastic blockmodel for dynamic networks. *Network Science*, 6(4):469–493.

Rockafellar, R. T. (1993). Lagrange multipliers and optimality. *SIAM review*, 35(2):183–238.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4:Article32.

Scheffe, H. (1999). *The analysis of variance*, volume 72. John Wiley & Sons.

Schmidt, R., Grimm, C., and Wyvill, B. (2006). Interactive decal compositing with discrete exponential maps. *ACM Trans. Graph.*, 25(3):605–613.

Severn, K., Dryden, I. L., and Preston, S. P. (2019). Manifold valued data analysis of samples of networks, with applications in corpus linguistics. *arXiv preprint arXiv:1902.08290*.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13:2498–2504.

Shaw, N. (1990). Free indirect speech and Jane Austen's 1816 revision of Northanger Abbey. *Studies in English Literature, 1500-1900*, 30(4):591–601.

Shetty, J. and Adibi, J. (2004). The Enron email dataset database schema and brief statistical report. *Information sciences institute technical report, University of Southern California*, 4(1):120–128.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905.

Shimada, Y., Hirata, Y., Ikeguchi, T., and Aihara, K. (2016). Graph distance for complex networks. *Scientific reports*, 6:34944.

Snijders, T. A. and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100.

Spielman, D. A. (2007). Spectral graph theory and its applications. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, pages 29–38. IEEE.

Stoyan, D. (1997). Geometrical means, medians and variances for samples of particles. *Particle & particle systems characterization*, 14(1):30–34.

Tang, M., Athreya, A., Sussman, D. L., Lyzinski, V., Park, Y., and Priebe, C. E. (2017). A semiparametric two-sample hypothesis testing problem for random graphs. *Journal of Computational and Graphical Statistics*, 26(2):344–354.

Ten Berge, J. M. F. (1977). Orthogonal Procrustes rotation for two or more matrices. *Psychometrika*, 42(2):267–276.

The Guardian (2006). Timeline: Enron. `https://www.theguardian.com/business/2006/jan/30/corporatefraud.enron`, Last accessed on 2019-05-10.

The Jane Austen Society of North America (2018). Jane austen's works. `http://jasna.org/austen/works/`, Last accessed on 2018-11-12.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, 15(1):273–289.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

Wallach, H. M. (2006). Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 977–984, New York, NY, USA. ACM.

Wang, H., Tang, M., Park, Y., and Priebe, C. E. (2014). Locality statistics for anomaly detection in time series of graphs. *IEEE Transactions on Signal Processing*, 62(3):703–717.

Wang, S., Arroyo, J., Vogelstein, J. T., and Priebe, C. E. (2017). Joint Embedding of Graphs. *arXiv e-prints*, page arXiv:1703.03862.

Ward, Jr., J. H. (1963). Hierarchical grouping to optimize an objective function. *J. Amer. Statist. Assoc.*, 58:236–244.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)*, 26(4):359–372.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of Small-world networks. *nature*, 393(6684):440–442.

Wegner, A. E., Ospina-Forero, L., Gaunt, R. E., Deane, C. M., and Reinert, G. (2018). Identifying networks with common organizational principles. *Journal of Complex Networks*, 6(6):887–913.

Weinberger, S. (1994). *The topological classification of stratified spaces*. University of Chicago Press.

Wilks, S. S. (1962). *Mathematical Statistics*. Wiley, New York.

Williams, C. K. (2002). On a connection between kernel PCA and metric multidimensional scaling. *Machine Learning*, 46(1):11–19.

Wilson, G. and Banzhaf, W. (2009). Discovery of email communication networks from the Enron corpus with a genetic algorithm using social network analysis. In *2009 IEEE Congress on Evolutionary Computation*, pages 3256–3263.

Zhou, D., Dryden, I. L., Koloydenko, A. A., Audenaert, K. M., and Bai, L. (2016). Regularisation, interpolation and visualisation of diffusion tensor images using non-Euclidean statistics. *Journal of Applied Statistics*, 43(5):943–978.