Adjacency Spectral Embedding Beyond Unweighted, Undirected Networks

Ian Gallagher

School of Mathematics
University of Bristol

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Doctor of Philosophy in the Faculty of Science.

23rd June 2022

Word count: 22940 words

Abstract

There exists a rich variety of real-world networks beyond those with unweighted and undirected edges. This thesis considers the statistical analysis of adjacency spectral embedding beyond these standard networks. By embedding the nodes of these networks into low dimensional space in a consistent and meaningful manner, it is possible to make better inferences from the embedded data. This thesis extends the existing theory for unweighted, undirected networks with two major contributions.

First, we extend the generalised random dot product graph, a latent position graph model, to allow for weighted networks and provide new results regarding the asymptotic distribution of the adjacency spectral embedding of these networks. This opens up new possibilities as a network can be transformed using a different representation of the edge weights, for example, embedding an adjacency matrix of p-values or an adjacency matrix of log p-values. In the case of the weighted stochastic block model, we can compare the quality of the different embeddings using the size-adjusted Chernoff information and consider the optimal transformation for a network.

Second, we consider dynamic networks where individual nodes, communities or the entire graph may change over time. We show that unfolded adjacency spectral embedding produces an embedding for every node at every time point in a statistically consistent way. Up to noise, nodes behaving similarly at a given time are given the same embedding (cross-sectional stability), as is a single node behaving similarly across different times (longitudinal stability). We show that many common dynamic network embedding algorithms often lack one, or both, of these desirable properties.

Acknowledgments

I would like to thank my supervisor Patrick Rubin-Delanchy for their support and guidance throughout my PhD. As you said in our first meeting, the aim of a PhD is to produce a piece of research you are proud of, and I am incredibly pleased with this work. I would also like to thank my collaborators during the course of my PhD; Anna Bertiger, Joshua Cape, Andrew Jones, Alex Modell, Melissa Turcotte and Carey Priebe. I feel privileged to have worked with you all in such an exciting area of research.

My time in Bristol has introduced me to a wide range of excellent people and I feel grateful to have met you all. In particular, I am thankful to Jenny for taking me under her wing when I first arrived into what should have been the Fry Building. Even today, I turn to you whenever I have a problem and you are always so happy to help.

I am also grateful for my friends who have supported me in my decision to return to university. Thank you Katie for giving me the encouragement to continue my studies, and for being there to celebrate when things go well. Your kind words and actions keep me going and, hopefully, I will be halfway to your two doctorates as a result.

Last but not least, thank you to my family for your love and support in all my endeavours. Your belief in me to succeed in whatever makes me happy is the reason this thesis is even here at all.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE: .....

Contents

List of Tables

List of Figures

Chapter 1

Introduction

A wide variety of data sets can be interpreted as networks. Nodes representing entities are connected by a possibly weighted or temporal edge representing some interaction between those entities. It includes social networks, either face-to-face human interaction or collaboration, or online communication via email or an online social network. It includes computer networks where interaction between computers is a mixture of automated activity and those initiated by people. It includes biological networks such as neuron communication within the brain and protein-protein interaction. Many physical infrastructures can be interpreted as a network, such as the road network, the power grid or flights between airports. Our thoughts and opinions on products and media can be converted into bipartite networks. The list goes on and on.

One approach to analysing these networks is to convert the graph into a low dimensional space representing the nodes that captures their structure in the network. By mapping the nodes into a low dimensional space, the embedding can be used as input for other techniques, such as node clustering [23] and edge weight prediction [7].

The rest of the thesis is structured as follows. In Chapter 2 we introduce the concept of graph embedding as a whole, highlighting the advantages of these approaches as a method of reducing the dimensionality of a network while maintaining some of its structural properties. We then gradually narrow our attention by looking at graph embeddings based on matrix factorisation, moving to those based on spectral decomposition before finally

focusing on adjacency spectral embedding.

In Chapter 3 we describe some simple unweighted graph models such as the stochastic block model and show how adjacency spectral embedding can recover the underlying structure of these models. Based on this, we cover the existing theory of the generalised random dot product model and the asymptotic behaviours under adjacency spectral embedding [61]. This theory is the basis of the rest of the research in this thesis.

In Chapter 4 we extend the theory of the generalised random dot product model to allow for weighted networks. This requires a new way of thinking as the latent position now encapsulates information about a family of distributions rather than edge probabilities. This adds an extra layer of abstraction between the latent positions of a weighted generalised random dot product graph and the resulting adjacency spectral embedding. We provide new results regarding the asymptotic behaviour of this extended model under adjacency spectral embedding [21].

In Chapter 5 we discuss how to compare the quality of different spectral embeddings, in particular measuring the community separation in a stochastic block model using size-adjusted Chernoff information. By extending the theory to weighted networks in the previous chapter, we are able to transform a weighted network by an entry-wise transformation of every edge. This can result in some representations having better community separation as shown by our synthetic example of pairwise p-value data first shown in Gallagher et al. [21].

In Chapter 6 we study dynamic networks in which nodes can change behaviour over time, forming new connections accordingly. The goal is to create an extension of adjacency spectral embedding for every node at every time point in a statistically consistent way. We propose two desirable properties for such an embedding; cross-sectional stability where nodes behaving similarly at a particular time have comparable embeddings, and longitudinal

stability where a node behaving similarly at different time points have comparable embeddings. We show that an existing technique called unfolded adjacency spectral embedding has both of these stability properties while many existing algorithms have neither. This research was first published in Gallagher et al. [22].

Chapter 2

Graph embedding

2.1 Graph embedding overview

There are numerous approaches to convert a graph into a low dimensional embedding. For example, the surveys [9, 24] bring together different techniques focusing on unweighted, undirected networks. Matrix factorisation based embeddings compute decompositions of some matrix of node proximity in a network, for example, the adjacency matrix directly or the symmetric Laplacian matrix defined in Section 2.3. Spectral embedding [72] refers to a subset of these algorithms that uses a spectral decomposition to perform the matrix factorisation.

Algorithms inspired by deep learning techniques have been adjusted to create graph embedding by the computer science community, such as DeepWalk [55] and node2vec [25]. These algorithms provide useful embeddings for large real-world networks, but often lack the statistical rigour explaining why these are sensible approaches. For example, many deep learning approaches apply ideas from natural language processing to random walks in the network and it was only until later that a link to matrix factorisation was discovered [43] and proof that they produce consistent embeddings for simple graph models [78].

A major advantage of spectral embedding algorithms is that these approaches can be given the necessary statistical rigour to show that they perform in a sensible and consistent way. In Chapter 3 we introduce the generalised random dot product graph [61], whose approach to analysing the

adjacency spectral embedding algorithm (given in Definition 1 below) is a key inspiration for the new extensions given in the rest of this thesis.

2.2 Adjacency spectral embedding

Definition 1 (Adjacency spectral embedding)**.** Given a d-truncated singular value decomposition of an adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$,

$$\mathbf{A} = \mathbf{U}_\mathbf{A} \mathbf{\Sigma}_\mathbf{A} \mathbf{V}_\mathbf{A}^\top + \mathbf{U}_{\mathbf{A},\perp} \mathbf{\Sigma}_{\mathbf{A},\perp} \mathbf{V}_{\mathbf{A},\perp}^\top,$$

where $\mathbf{U}_\mathbf{A}, \mathbf{V}_\mathbf{A} \in \mathbb{O}(n \times d) = \{\mathbf{Q} \in \mathbb{R}^{n \times d} : \mathbf{Q}^\top \mathbf{Q} = \mathbf{I}\}$ and $\mathbf{\Sigma}_\mathbf{A} \in \mathbb{R}^{d \times d}$ is the diagonal matrix comprising the d largest singular values of $\mathbf{A}$ arranged in decreasing order denote by $\hat{\mathbf{X}} \in \mathbb{R}^{n \times d}$ the *adjacency spectral embedding*

$$\hat{\mathbf{X}} = \mathbf{U}_\mathbf{A} \mathbf{\Sigma}_\mathbf{A}^{1/2} \in \mathbb{R}^{n \times d}.$$

We divide the adjacency spectral embedding into rows to represent the spectral embedding for each node in the graph. By writing $\hat{\mathbf{X}} = (\hat{X}_1 \mid \cdots \mid \hat{X}_n)^\top$, $\hat{X}_i$ is the spectral embedding representation for node i. Occasionally it will be useful to construct the adjacency spectral embedding for something other than an adjacency matrix. Given a d-truncated singular value decomposition of an arbitrary matrix $\mathbf{M}$, we denote the adjacency spectral embedding as $\mathbf{X}_\mathbf{M} = \mathbf{U}_\mathbf{M} \mathbf{\Sigma}_\mathbf{M}^{1/2}$.

In this thesis, we do not concern ourselves with how the embedding dimension d is chosen, instead assuming that it has been given to us by an oracle. The dimension d can be estimated based on the spectral decomposition of the adjacency matrix $\mathbf{A}$ [80, 12]. In Section 3.2 we show for simple graph models how to compute the correct embedding dimension and how it relates to the observed adjacency matrix. In these cases, we know which singular values of $\mathbf{A}$ can be discarded as they are close to zero and essentially noise. We will not treat estimation of d in any further detail, and assume it is known.

Example 1 (*Les Misérables* co-occurrence network)**.** To give an example of adjacency spectral embedding, consider the weighted graph of co-occurrences between characters in the novel *Les Misérables* [33]. A node represents one of the 77 characters in the book and the edge weight counts the number of times the characters appeared in the same chapter [40].

Figure 2.1 shows the adjacency spectral embedding of the network into $d = 2$ dimensions chosen using the singular values of $\mathbf{A}$ [80]. The embedding approximately creates two rays of points emanating from the origin. One of these rays contains some of the major characters in the novel with a few examples shown as red triangles with the protagonist Jean Valjean located at the end. We shall see in Section 3.1.3 that this is common feature of a degree-corrected stochastic block model.

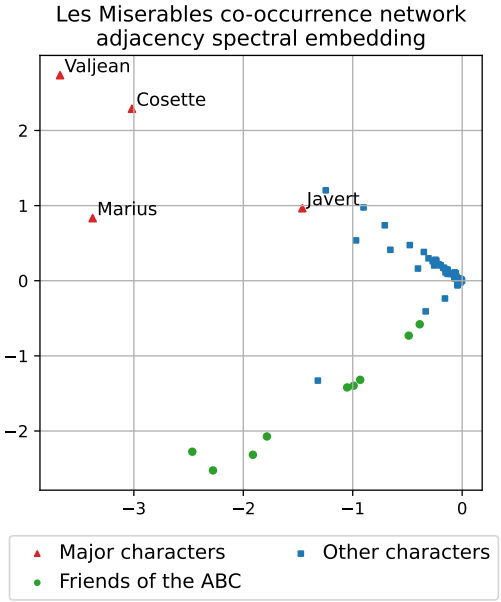

Figure 2.1: The embedding $\hat{\mathbf{X}}$ of the *Les Misérables* co-occurrence network into $d = 2$ dimensions. The points are coloured according to their role in the book.

The other ray shows members of the Friends of the ABC, a revolutionary club of students, shown as green circles. One point along this ray is different but corresponds to the street urchin Gavroche who is linked to the revolutionary group. The adjacency spectral embedding is able to capture some of the character interactions from the network and reduces the information to a more manageable dimension where further analysis can be performed. ◁

In previous work [60, 21] the adjacency spectral embedding was defined in terms of the eigendecomposition of the adjacency matrix $\mathbf{A}$. However, for generalisation, it is convenient to define adjacency spectral embedding using singular value decomposition. For example, the rectangular unfolded adjacency matrix [36] cannot be analysed at all using an eigendecomposition. Given the eigendecomposition of an adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$,

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{U}^\top + \mathbf{U}_\perp \mathbf{S}_\perp \mathbf{U}_\perp^\top,$$

where $\mathbf{U} \in \mathbb{O}(n \times d)$ and $\mathbf{S} \in \mathbb{R}^{d \times d}$ is the diagonal matrix comprising the d largest eigenvalues of $\mathbf{A}$ arranged in decreasing order of magnitude, the adjacency spectral embedding $\hat{\mathbf{X}} \in \mathbb{R}^{n \times d}$ can also be defined using the absolute value of the eigenvalues,

$$\hat{\mathbf{X}} = \mathbf{U}|\mathbf{S}|^{1/2}.$$

It is important the singular value decomposition is using the d largest eigenvalues of $\mathbf{A}$ in magnitude. A large collection of exemplary work on adjacency spectral embedding initially used the d largest positive eigenvalues of $\mathbf{A}$, reviewed by Athreya et al. [6]. By only considering the largest eigenvalues it is only possible to correctly analyse networks with positive definite adjacency matrices. However, in a bipartite network, for every positive eigenvalue there exists a negative one with the same magnitude and it is incorrect to say that one of these is more important to understanding the network structure than the other. This is just one example and we will see many more in this thesis where both the positive and negative eigenvalues are key.

To show that the two versions of the adjacency spectral embedding algorithm are the same, we introduce the indefinite orthogonal group.

Definition 2 (Indefinite orthogonal group)**.** Given $p > 0$, $q \geq 0$ and $d = p+q$, the *indefinite orthogonal group* is defined as the set of matrices in $\mathbb{R}^{d\times d}$ such that

$$\mathbb{O}(p,q) = \{\mathbf{Q} \in \mathbb{R}^{d\times d} : \mathbf{Q}^\top \mathbf{I}_{p,q} \mathbf{Q} = \mathbf{I}_{p,q}\},$$

where

$$\mathbf{I}_{p,q} = \operatorname{diag}(\underbrace{1,\ldots,1}_{\text{length } p}, \underbrace{-1,\ldots,-1}_{\text{length } q}).$$

When $p = d$ and $q = 0$ this reduces to the standard group of orthogonal matrices $\mathbb{O}(d) = \{\mathbf{Q} \in \mathbb{R}^{d\times d} : \mathbf{Q}^\top\mathbf{Q} = \mathbf{I}\}$. A matrix $\mathbf{M}$ has signature (p,q) if it has p positive and q negative eigenvalues. In both case, the adjacency spectral embedding satisfies

$$\mathbf{M} = \mathbf{X}_\mathbf{M}\mathbf{I}_{p,q}\mathbf{X}_\mathbf{M}^\top.$$

However, given an adjacency spectral embedding, we can provide another valid embedding $\mathbf{X}_\mathbf{M}\mathbf{Q}^\top$ for any $\mathbf{Q} \in \mathbb{O}(p,q)$. This non-uniqueness is a key feature of the adjacency spectral embedding algorithm and results regarding its behaviour, such as Theorem 1 and 2, have to account for these transformations.

Given an eigendecomposition adjacency spectral embedding, one can always construct one using the singular value decomposition by setting $\mathbf{U}_\mathbf{M} = \mathbf{U}$, $\mathbf{\Sigma}_\mathbf{M} = |\mathbf{S}|$ and $\mathbf{V}_\mathbf{M} = \mathbf{U}\mathbf{I}_{p,q}$. However, some care must be taken when considering the reverse statement. Using a singular value decomposition can lead to non-uniqueness issues. For example, bipartite graphs have repeated singular values and hence can have different basis for the singular vectors compared to the eigenvectors. However, it is always possible to find a transformation $\mathbf{Q} \in \mathbb{O}(p,q)$ that converts the singular value decomposition embedding into one obtained by an eigendecomposition.

For the remainder of the thesis, we will use adjacency spectral embedding to refer to the singular value decomposition version.

2.3 Other spectral embeddings

Adjacency spectral embedding is not the only way to use spectral techniques to embed a network into lower dimensions. Rather than analysing the graph adjacency matrix, different representations of the network can be used. For example, Laplacian spectral embedding is the same as adjacency spectral embedding in Definition 1 except the unweighted adjacency matrix is replaced with the symmetric normalised Laplacian matrix,

$$\mathbf{L}^{\text{sym}} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2},$$

where $\mathbf{D} \in \mathbb{R}^{n \times n}$ is the diagonal matrix of node degrees with diagonal entries $D_i = \sum_j \mathbf{A}_{ij}$. Another embedding technique instead uses the random walk Laplacian matrix [50],

$$\mathbf{L}^{\text{rw}} = \mathbf{D}^{-1}\mathbf{A}.$$

Using different matrix regularisation techniques for the adjacency matrix can result in different types of structure being detected by the spectral embedding [63, 68, 10]. For example, in brain connectivity networks, clustering using the Laplacian spectral embedding separates the left and right hemispheres of the brain, while adjacency spectral embedding distinguishes between white and grey matter [57]. Using the random walk embedding to analyse a social network captures a person's friend preferences separately from how popular or active they may be in that social network [50].

Moving from unweighted to weighted networks, for Laplacian-based spectral embedding techniques one needs to be aware what the degree of a node actually means for a weighted network. In many situations defining the degree as the sum of weighted edges, $D_i = \sum_j \mathbf{A}_{ij}$, makes sense. For example,

in computer networks where edge weights represent the number of bytes being transferred between computer nodes and the degree represents the total amount of data leaving or arriving at a computer. Although, things may not always be so clear.

Example 2 (Network of pairwise p-values)**.** Suppose we have a weighted network where edge weights represent the observed p-value between a pair of nodes quantifying our surprise in their level of activity. The degree of a node is the sum of p-values and, while there are approaches based on adding p-values [18], probably the most common method for combining p-values $p_1, \ldots, p_n$ is Fisher's method [20] using the test statistic,

$$-2 \sum_{i=1}^{n} \log p_i.$$

This suggests using the log p-values as an alternative edge weight in this network as the node degree is now proportional to this test statistic.

Alternatively, one can convert the matrix of p-values into a binary graph where edges represent p-values below a particular threshold that is deemed to be of interest. After this transformation, the degree is the number of interesting p-values involving a particular node. This network is discussed in much more detail in Section 5.3. ◁

It is often assumed that all of the weights in a network are positive, however, there is nothing preventing negative edge weights in the network, perhaps to denote disassociation between nodes. This can lead to problems if some nodes in the network have negative degrees as spectral embedding using the symmetric normalised Laplacian $\mathbf{L}^{\text{sym}}$ cannot be computed without resorting to complex numbers, although the random walk Laplacian $\mathbf{L}^{\text{rw}}$ can still be used.

To avoid the issues around node degree in weighted networks, this thesis will focus solely on adjacency spectral embedding. There is plenty of scope

for future research to extend Laplacian-based spectral embedding techniques to weighted networks and dynamic networks in a meaningful and consistent way.

Chapter 3

Unweighted networks

In this chapter we give a review of research related to the adjacency spectral embedding algorithm for unweighted networks. We describe a few basic unweighted graph models before introducing the generalised random dot product graph [60]. This will provide an important starting point for the rest of the thesis which introduces two new extensions to this model.

3.1 Simple graph models

This section introduces three graph models as motivation for the use of adjacency spectral embedding to recover the structure of these models. Section 3.3 gives the asymptotic distribution for adjacency spectral embedding in general, justifying the algorithms for recovering the graph model parameters discussed here.

3.1.1 Stochastic block model

Suppose we have a network of university staff where links represent communication between two faculty members. In this network, the probability of an edge between two members of staff will depend on their departments within the university. People within the same department are much more likely to communicate, but other pairs of departments may also be likely to communicate, for example, mathematics and computer science compared to biology and history. The departments divide the university in communities

and one basic network model assumes that the probability two members of staff communicate depends only on the departments they belong.

The stochastic block model is a graph generative model that creates networks with this community structure [31]. Nodes in a graph are assigned to one of K possible communities using a latent variable Z_i represents the community assignment of node i. The probability of an edge between two nodes i and j depends only on their community assignments Z_i and Z_j.

Definition 3 (Stochastic block model)**.** Let $\mathcal{Z} = \{1, \ldots, K\}$ be a sample space with probability distribution F. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is distributed as a *stochastic block model*, if $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} F$ and, for $i < j$,

$$\mathbf{A}_{ij} \mid Z_i, Z_j \overset{\text{ind}}{\sim} \text{Bernoulli}(\mathbf{B}_{Z_i Z_j}),$$

where $\mathbf{B} \in [0,1]^{K \times K}$ is the *mean block matrix*, a matrix of inter-community edge probabilities.

Given $Z \sim F$, we define $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ where $\pi_k = \mathbb{P}(Z = k)$ is the probability a node is assigned to community k.

To see why this is known as a block model, imagine that the rows and columns of the adjacency matrix $\mathbf{A}$ are arranged according to their community assignment Z_i. The adjacency matrix can be rewritten as a block matrix,

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}^{(1,1)} & \cdots & \mathbf{A}^{(1,K)} \\ \vdots & \ddots & \vdots \\ \mathbf{A}^{(K,1)} & \cdots & \mathbf{A}^{(K,K)} \end{pmatrix},$$

where $\mathbf{A}^{(k,\ell)}$ represents the adjacency matrix where one node is in community k, and the other is in community ℓ. In each of these blocks, the edge density is fixed since

$$\mathbf{A}_{ij}^{(k,\ell)} \overset{\text{ind}}{\sim} \text{Bernoulli}(\mathbf{B}_{k\ell}),$$

for all nodes i, j such that $Z_i = k$ and $Z_j = \ell$. The entry $\mathbf{B}_{k\ell}$ is the Bernoulli parameter for the adjacency block $\mathbf{A}^{(k,\ell)}$, the mean of the Bernoulli distribution. The importance of the block mean matrix will become more apparent when extending the stochastic block model to weighted graphs.

Example 3 (Erdős-Rényi model)**.** In the $G(n, p)$ version of the Erdős-Rényi random graph model [19], a graph is constructed on n nodes where every possible edge is included with probability p, independently from all other edges. This can be thought of as a stochastic block model with a single community ($Z_i = 1$ for all nodes) and block mean matrix $\mathbf{B} = (p)$. The stochastic block model is an extension of the Erdős-Rényi random graph model to include community structure. ◁

Example 4 (Two-community stochastic block model)**.** Consider a stochastic block model with two communities with block mean matrix

$$\mathbf{B} = \begin{pmatrix} 0.10 & 0.20 \\ 0.20 & 0.05 \end{pmatrix},$$

with signature $(1, 1)$, and a distribution F that assigns equal probability to the two communities; $\pi_1 = 1/2$ and $\pi_2 = 1/2$.

Using this model, we generate a stochastic block model with $n = 1000$ nodes and perform adjacency spectral embedding into $d = 2$ dimensions. The number of embedding dimensions is chosen to be the rank of the block mean matrix $\mathbf{B}$, which in this case is the number of communities in the stochastic block model. The reasoning behind this choice is explained in Section 3.2.1. The leading two singular values of the resulting adjacency matrix $\mathbf{A}$ correspond to one positive and one negative eigenvalue matching the signature of the block mean matrix $\mathbf{B}$. Figure 3.1 shows the embedding $\hat{\mathbf{X}}$ of the stochastic block model. Points are coloured according to their community assignment Z_i and form two clusters.

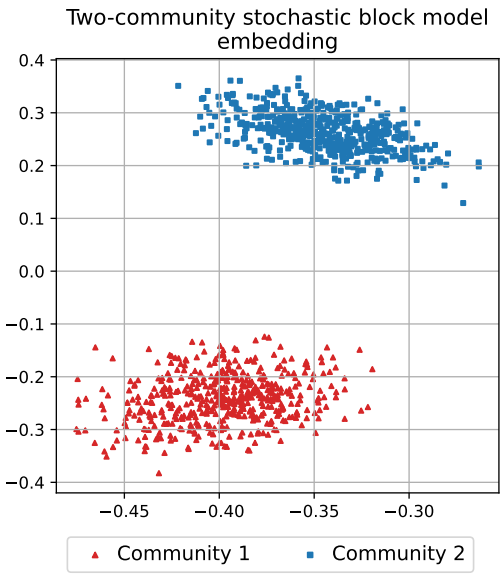

Figure 3.1: The embedding $\hat{\mathbf{X}}$ of a two-community stochastic block model into $d = 2$ dimensions. The points are coloured according to their true community assignment Z_i.

Suppose we did not know the true community assignment of the nodes within the graph. By performing the adjacency spectral embedding and fitting a Gaussian mixture model with two components, we can recover the community labels of the nodes, up to community relabeling. As shown in Figure 3.1, the cluster covariances can be unequal and non-spherical, which is justified by the asymptotic distribution of the embedding given in Section 3.3.1. Therefore, the chosen clustering algorithm should account for that possibility meaning an algorithm like k-means may not be appropriate, as seen in Example 20. ◁

3.1.2 Mixed membership stochastic block model

While the stochastic block model is a straightforward way to generate random graphs with community structure, it is perhaps unrealistic to assume that a node in the network belongs to a single community. For example, returning to the university communications network in the previous section, a data scientist may be employed by the mathematics department, but also be considered part of the biology and social science departments.

The mixed membership stochastic block model extends the stochastic block model and allows nodes within the graph to belong to multiple communities [3]. Every node is assigned a probability distribution over the K communities, which dictates how they behave for a possible connection in the graph.

To model this, node i generates a random variable $Z_{i\to j} \sim F_i$ representing the community role node i adopts when communicating with node j. Returning to our data scientist, suppose they have the majority of their probability mass assigned to the mathematics department, with smaller amounts assigned to the biology and social science departments and the remaining departments within the university have probability mass zero. This defines their distribution F_i. Most of the time they will behave like a mathematician when interacting with other staff members, but occasionally behave like a biologist or social scientist.

To see if there is a link between the two nodes, node j generates a random variable $Z_{j\to i} \sim F_j$ represents the community role node j adopts when communicating with node i. The probability of an edge between two nodes i and j depends only on their community assignments $Z_{i\to j}$ and $Z_{j\to i}$. Once the community assignments $Z_{i\to j}$ and $Z_{j\to i}$ have been selected, the model works in the same way as the stochastic block model.

Definition 4 (Mixed membership stochastic block model)**.** Let $\mathcal{Z} = \{1, \ldots, K\}$

be a sample space where each node has a probability distribution F_i. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is distributed as a *mixed membership stochastic block model*, if $Z_{i \to j} \overset{\text{ind}}{\sim} F_i$, $Z_{j \to i} \overset{\text{ind}}{\sim} F_j$, and, for $i < j$,

$$\mathbf{A}_{ij} \mid Z_{i \to j}, Z_{j \to i} \overset{\text{ind}}{\sim} \text{Bernoulli}(\mathbf{B}_{Z_{i \to j} Z_{j \to i}}),$$

where $\mathbf{B} \in [0,1]^{K \times K}$ is the matrix of inter-community edge probabilities.

Given $Z_i \sim F_i$, we define $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK})$, where $\pi_{ik} = \mathbb{P}(Z_i = k)$ is the probability a node is assigned to community k.

The stochastic block model is a special case of the mixed membership stochastic block model where the distributions F_i places all of its probability mass on the community Z_i. In this case $Z_{i \to j} = Z_i$, $Z_{j \to i} = Z_j$ and the model reduces to Definition 3.

Example 5 (Three-community mixed membership stochastic block model)**.** Consider a three-community mixed membership stochastic block model with block mean matrix

$$\mathbf{B} = \begin{pmatrix} 0.10 & 0.20 & 0.22 \\ 0.20 & 0.05 & 0.16 \\ 0.22 & 0.16 & 0.26 \end{pmatrix}.$$

The block mean matrix is not full rank and has signature $(1, 1)$. Each node generates a probability distribution F_i over the three communities independently using a Dirichlet distribution with parameter $\boldsymbol{\alpha} = (0.2, 0.2, 0.2)$ so each community is given equal preference. Since $\alpha_k < 1$ in each component, F_i is likely to be unbalanced meaning a large amount of probability is assigned to a single community.

We generate a mixed membership stochastic block model with $n = 1000$ nodes and perform adjacency spectral embedding into $d = 2$ dimensions, the rank of the block mean matrix $\mathbf{B}$, in this case, smaller than the number of communities.

Figure 3.2 shows the embedding $\hat{\mathbf{X}}$ of the mixed membership stochastic block model. Points are coloured a mixture of red, blue and green according to their probability distribution F_i over the communities. Nodes that are more likely to behave like a single community, $\pi_{ik} = \mathbb{P}(Z_{i\to j} = k) \geq 0.9$ for some community k, are shown as larger points; $\pi_{i1} \geq 0.9$ as red triangles, $\pi_{i2} \geq 0.9$ as blue squares, and $\pi_{i3} \geq 0.9$ as green circles.

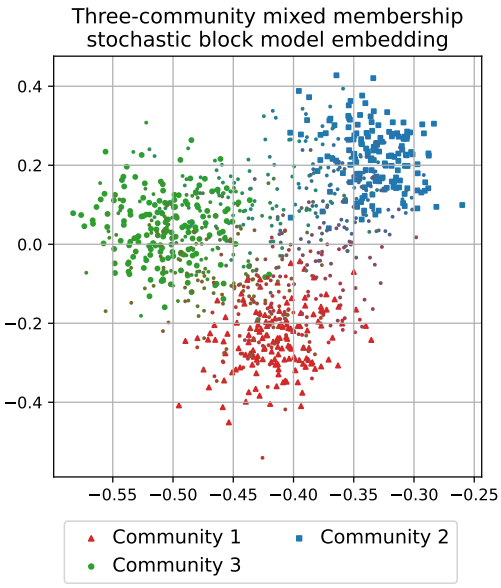

Figure 3.2: The embedding $\hat{\mathbf{X}}$ of a three-community mixed membership stochastic block model. The points are coloured a mixture of red (triangle), blue (square) and green (circle) according to their probability distribution F_i over the communities.

The node embedding $\hat{\mathbf{X}}$ forms a noisy triangle, a two-dimensional simplex, where the corners represent the three communities. Points near the corners have the majority of their probability mass in F_i assigned to a single community, a red triangle, blue square or green circle depending on the dominant community. The remaining points approximately lie within this

simplex where their position represents the community probabilities π_{ik} in F_i.

Suppose we did not know the the community distributions F_i for the nodes that we wish to recover. Rubin-Delanchy et al. [61] describes an algorithm that gives consistent estimation of the parameters for a K-community mixed membership stochastic block model, although for this example, the method for recovering the distributions F_i is sufficient. In this example, their approach finds the smallest area triangle enclosing the embedding. The point $\hat{X}_i$ in the embedding can be expressed as a convex combination of the corners giving an estimator of the community distribution F_i. The reason why this approach gives a consistent estimator for F_i is discussed in Section 3.3.2. ◁

3.1.3 Degree-corrected stochastic block model

Another potential downside of the stochastic model block is that nodes within a community approximately have the same amount of activity in the network. To demonstrate this, we compute the expected degree of a node in a stochastic block model,

$$\mathbb{E}(D_i) = \sum_{j \neq i} \mathbb{E}(\mathbf{A}_{ij}) = \sum_{j \neq i} \mathbf{B}_{Z_i Z_j}.$$

The expected degree of a node depends only on its community, therefore, two nodes in the same community have the same expected degree.

While a network may have groups of similarly behaving nodes, it may be desirable to allow nodes within a community to display different levels of activity. Returning again to the university communications network, different members of a department may be more active. For example, a professor may communicate more and have a higher degree compared to a new PhD student, even though they both communicate to similar types of people.

The degree-corrected stochastic block model extends the stochastic block model by introducing a node-specific weight parameter $w_i \in (0, 1]$ which

represents their activity [37]. The probability of an edge between two nodes i and j still depends on their community assignments Z_i and Z_j, but also the weights w_i and w_j. In the university communication network example, the professor above may have weight close to one, while the new PhD student may have weight closer to zero.

Definition 5 (Degree-corrected stochastic block model)**.** Let $\mathcal{Z} = \{1, \ldots, K\}$ be a sample space with probability distribution F, and let G_k, $k \in \mathcal{Z}$ be a collection of probability distributions on the interval $(0, 1]$. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is distributed as a *degree-corrected stochastic block model*, if $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} F$ and, for $i < j$,

$$\mathbf{A}_{ij} \mid w_i, w_j, Z_i, Z_j \overset{\text{ind}}{\sim} \text{Bernoulli}(w_i w_j \mathbf{B}_{Z_i Z_j}),$$

where $\mathbf{B} \in [0, 1]^{K \times K}$ is the matrix of inter-community edge probabilities and $w_i \mid Z_i \overset{\text{ind}}{\sim} G_{Z_i}$ is a node-specific weight parameter.

Similarly to the stochastic block model, we define $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, where $\pi_k = \mathbb{P}(Z = k)$ is the probability a node is assigned to community k.

In this model, the weight parameter w_i for a node can be distributed differently depending on the community assignment Z_i. Perhaps one community contains more active nodes with larger weight parameters on average than another community.

To demonstrate why this model is known as the degree-corrected stochastic block model, given the weight parameters w_i, the expected degree of any given node is

$$\mathbb{E}(D_i) = \sum_{j \neq i} \mathbb{E}(\mathbf{A}_{ij}) = w_i \sum_{j \neq i} w_j \mathbf{B}_{Z_i Z_j}.$$

Unlike the stochastic block model, the expected degree of a node increases with the weight parameter w_i. For two nodes i and j in the same community ($Z_i = Z_j$), the node with the larger weight parameter has the larger expected

degree, for $w_i > w_j$,

$$\begin{aligned}\mathbb{E}(D_i) - \mathbb{E}(D_j) &= w_i \sum_{k \neq i} w_k \mathbf{B}_{Z_i Z_k} - w_j \sum_{k \neq j} w_k \mathbf{B}_{Z_j Z_k} \\ &= (w_i - w_j) \sum_{k \neq i,j} w_k \mathbf{B}_{Z_i Z_k} > 0.\end{aligned}$$

Note that for nodes in different communities ($Z_i \neq Z_j$) this may not be the case but, in the university communication network example, the professor would have a larger expected degree than a PhD student in the same department.

The stochastic block model is a special case of the degree-corrected stochastic model where $w_i = 1$ for all nodes. More interestingly, the degree-corrected stochastic block model can also be interpreted as a mixed membership stochastic block model.

Example 6 (Degree-corrected to mixed membership stochastic block model)**.** Consider a K-community degree-corrected stochastic block model with block mean matrix $\mathbf{B}$ and weight parameters w_i. This can be written as a $(K+1)$-community mixed membership stochastic block model by introducing a new community that generates no edges in the network. Define a new mean block matrix $\mathbf{B}' \in [0,1]^{(K+1)\times(K+1)}$ such that nodes assigned to community $K+1$ have probability zero of forming an edge,

$$\mathbf{B}'_{ij} = \begin{cases} \mathbf{B}_{ij} & \text{if } 1 \leq i, j \leq K, \\ 0 & \text{otherwise.} \end{cases}$$

The community assignment distribution F_i is defined such that, for $Z_{i\to j} \sim F_i$, the probability a node is assigned to a community k in the mixed membership stochastic block model is given by

$$\pi_{ik} = \mathbb{P}(Z_{i\to j} = k) = \begin{cases} w_i & \text{if } k = Z_i, \\ 1 - w_i & \text{if } k = K+1 \\ 0 & \text{otherwise.} \end{cases}$$

Each node can be considered as a mixture of its true community and the new null community where the probability a node is assigned to its true community Z_i is equal to w_i, and the probability it is assigned to the null community $K+1$ is $1-w_i$. By summing over the possible community assignments in this new mixed membership stochastic block model, the probability of an edge between nodes i and j is given by

$$\mathbb{P}(\mathbf{A}_{ij} = 1 \mid Z_i, Z_j) = \sum_{k,\ell=1}^{K+1} \pi_{ik}\pi_{j\ell}\mathbf{B}'_{k\ell} = w_i w_j \mathbf{B}_{Z_i Z_j}.$$

Therefore, the probability of an edge between two nodes in the $(K+1)$-community mixed membership stochastic block model is identical to the K-community degree-corrected stochastic block model. ◁

Example 7 (Two-community degree-corrected stochastic block model)**.** Consider a two-community degree-corrected stochastic block model with the same block mean matrix as the stochastic block model in Example 4, and a distribution F that assigns equal probability to the two communities; $\pi_1 = 1/2$ and $\pi_2 = 1/2$. Each node is independently assigned a weight parameter $w_i \overset{\text{ind}}{\sim} \text{Beta}(2,2)$. Note that in this case, the distribution is the same for both communities, but in general that does not need to be the case. We generate a degree-corrected stochastic block model with $n = 1000$ nodes and perform adjacency spectral embedding into $d = 2$ dimensions, chosen to be equal to the number of communities.

Figure 3.3 shows the embedding $\hat{\mathbf{X}}$ of the degree-corrected stochastic block model. Points are coloured according to their community assignment $\mathbf{Z}_i$ and form two rays starting from the origin. Points near the origin correspond to a weight parameter $w_i \approx 0$, while points near the ends of the rays correspond to a weight parameter $w_i \approx 1$.

Suppose that we did not know the community assignments or the weight parameters for each node. There are other spectral embedding techniques

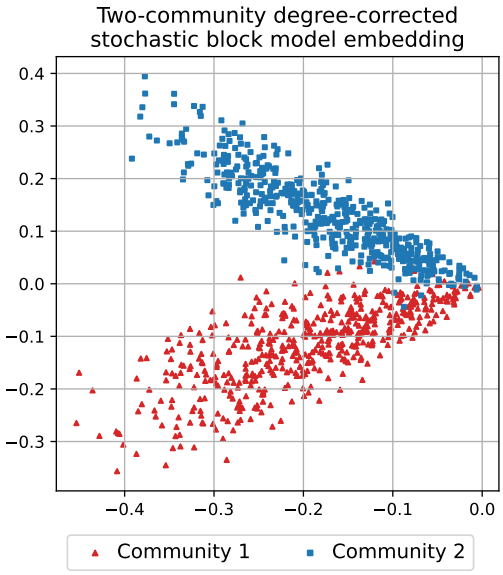

Figure 3.3: The embedding $\hat{\mathbf{X}}$ of a two-community degree-corrected stochastic block model into $d = 2$ dimensions. The points are coloured according to their true community assignment Z_i.

designed for degree-corrected stochastic block models [34, 50], however we can recover the communities by projecting the adjacency spectral embedding on to the unit sphere [53] or a hyperplane [50] then clustering the projected points. A line can be fitted to each ray of points using linear regression and the weight w_i can be estimated by the distance from the origin, the justification for this technique given in Section 3.3.3. ◁

3.2 Generalised random dot product graph

This section introduces the generalised random dot product graph [60]. It is an extension of the random dot product graph where the probability of an edge is the inner product of the corresponding latent positions [77, 51].

Under this model the resulting adjacency matrix $\mathbf{A}$ will very likely be positive definite. In general, the random dot product graph cannot model graphs that are not positive definite, like the examples given so far in this chapter. Constructing a spectral embedding of an adjacency matrix by only using the largest positive eigenvalues necessarily fails to capture some of the underlying structure of the network [6]. By allowing the probability of an edge to be the indefinite inner product of the corresponding latent positions, the generalised random dot product graph models a wider variety of network.

Definition 6 (Generalised random dot product graph)**.** Let $\mathcal{X}$ be a subset of $\mathbb{R}^d$ such that $x_1^\top \mathbf{I}_{p,q} x_2 \in [0,1]$ for all $x_1, x_2 \in \mathcal{X}$, with $F_\mathcal{X}$ a probability distribution over $\mathcal{X}$. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is distributed as a *generalised random dot product graph* with signature (p,q), if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F_\mathcal{X}$ and, for $i < j$,

$$\mathbf{A}_{ij} \mid X_i, X_j \overset{\text{ind}}{\sim} \text{Bernoulli}(X_i^\top \mathbf{I}_{p,q} X_j).$$

As with the adjacency spectral embedding, it is convenient to call $\mathbf{X} = (X_1 \mid \cdots \mid X_n)^\top \in \mathbb{R}^{n\times d}$, the latent positions of the generalised random dot product graph. In this section we show how the latent positions of the generalised random dot product graph $\mathbf{X}$ relate to the adjacency spectral embedding $\hat{\mathbf{X}}$. Note that the conditional distribution of $\mathbf{A}$ is unchanged if the latent positions X_i are replaced by $\mathbf{Q}X_i$ for any indefinite orthogonal matrix $\mathbf{Q} \in \mathbb{O}(p,q)$, so the connection between the two will need to account for this unidentifiability.

The standard definition of the generalised random dot product allow for sparse networks by introducing a parameter ρ_n that prevents the average degree of a node growing linearly as the number of nodes n increases. This is done by scaling the latent positions $\rho_n^{1/2} X_i$ so that the probability of an edge between two nodes is distributed as

$$\mathbf{A}_{ij} \mid X_i, X_j \overset{\text{ind}}{\sim} \text{Bernoulli}(\rho_n X_i^\top \mathbf{I}_{p,q} X_j).$$

In this setting, the expected degree of a node grows as $n\rho_n$ so by having $\rho_n \to 0$, the model can produce sparse networks. In this thesis, we will focus on the dense setting where $\rho_n = 1$ as this is our setting for making extensions for weighted networks in Chapter 4 and dynamic networks in Chapter 6.

Example 8 (Multipartite networks)**.** To show the flexibility of the generalised random dot product graph, consider a subset $\mathcal{X}$ that is union of K isotropic subspaces,

$$\mathcal{X} = \bigcup_{k=1}^{K} \mathcal{X}_k,$$

meaning, for all $x_1, x_2 \in \mathcal{X}_k$, $x_1^\top \mathbf{I}_{p,q} x_2 = 0$. One example of isotropic subspaces is given by the latent positions in two dimensions $\mathcal{X}_1 = \{(x, x) : x \in \mathbb{R}\}$ and $\mathcal{X}_2 = \{(x, -x) : x \in \mathbb{R}\}$ with signature $(1, 1)$. For two nodes represented by latent positions in the subspace $\mathcal{X}_k$, there is zero probability of an edge between them, generating a K-partite graph [49]. ◁

The three versions of the stochastic block model defined in Section 3.1 are examples of the generalised random dot product graph. Each example uses the adjacency spectral embedding of the block mean matrix $\mathbf{B} \in [0, 1]^{K \times K}$. If $\mathbf{B}$ has signature (p, q), then the embedding dimension is given by $d = p + q = \text{rank}(\mathbf{B}) \leq K$ resulting in the adjacency spectral embedding $\mathbf{X_B} = \mathbf{U_B}\mathbf{\Sigma_B}^{1/2} \in \mathbb{R}^{K \times d}$. As shown in Section 2.2, the adjacency spectral embedding satisfies

$$\mathbf{B} = \mathbf{X_B}\mathbf{I}_{p,q}\mathbf{X_B}^\top.$$

This resembles the Bernoulli probability $X_i^\top \mathbf{I}_{p,q} X_j$ given in Definition 6.

Let $(\mathbf{M})_k$ denote row k of a matrix $\mathbf{M}$. The rows of the adjacency spectral embedding $(\mathbf{X_B})_1^\top, \ldots, (\mathbf{X_B})_K^\top$ provide a good starting point for constructing latent positions for generalised random dot product graphs for these simple graph models. The statistical theory in this section requires that the embedding dimension for the adjacency spectral embedding is equal to the rank

of the block mean matrix $\mathbf{B}$. The next three sections show how the different stochastic block models in Section 3.1 can be expressed as a generalised random dot product graph [60].

3.2.1 Stochastic block model

Lemma 1. The stochastic block model is an instance of the generalised random dot product graph.

Proof. Recall Definition 3 for a stochastic block model. Given a sample space $\mathcal{Z} = \{1, \ldots, K\}$ with probability distribution F, then, for $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} F$, a symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ satisfies, for $i < j$,

$$\mathbf{A}_{ij} \mid Z_i, Z_j \overset{\text{ind}}{\sim} \text{Bernoulli}(\mathbf{B}_{Z_i Z_j}),$$

where $\mathbf{B} \in [0,1]^{K \times K}$ is the matrix of inter-community edge probabilities.

To construct the equivalent generalised random dot product graph, map the latent position Z_i to the corresponding row of the adjacency spectral embedding $\mathbf{X_B}$,

$$X_i = (\mathbf{X_B})_{Z_i}^\top .$$

The latent space $\mathcal{X}$ consists of the K points $(\mathbf{X_B})_1^\top, \ldots, (\mathbf{X_B})_K^\top$ with distribution $F_{\mathcal{X}}$ inherited from the stochastic block model, for $X \sim F_{\mathcal{X}}$,

$$\pi_k = \mathbb{P}(X = (\mathbf{X_B})_k^\top).$$

By the construction of the adjacency spectral embedding $\mathbf{X_B}$,

$$X_i^\top \mathbf{I}_{p,q} X_j = \mathbf{B}_{Z_i Z_j},$$

therefore, the probability of an edge $\mathbf{A}_{ij}$ is identical under the stochastic block model and generalised random dot product graph. □

From this construction, there are K possible latent positions X_i in the generalised random dot product graph assigned depending on the latent positions Z_i in the stochastic block model. There is evidence of this behaviour in Figure 3.1 from Example 4. For this particular two-community stochastic block model, there are two clusters of points in the adjacency spectral embedding. We show in Section 3.3.1 how the centres of these clusters are related to the latent positions $(\mathbf{X_B})_1^\top, \ldots, (\mathbf{X_B})_K^\top$.

For sensibly defined stochastic block models, there is a separate latent position for each community in the network, regardless of the dimension of the adjacency spectral embedding. Suppose that two distinct communities were assigned the same latent positions; $(\mathbf{X_B})_k^\top = (\mathbf{X_B})_\ell^\top$ for some $k \neq \ell$. Since $\mathbf{B} = \mathbf{X_B}\mathbf{I}_{p,q}\mathbf{X_B}^\top$, this implies that the corresponding rows of $\mathbf{B}$ must be equal. Therefore, we could have defined the stochastic block model more efficiently by combining communities k and ℓ into a single community with assignment probability $\pi_k + \pi_\ell$. In general, it is assumed that the stochastic block models are defined using as few communities as necessary.

There is no issue if the mean block matrix $\mathbf{B}$ is rank-deficient and the embedding dimension d is strictly smaller than the number of communities K in the stochastic block model. It is possible for the K communities to coexist in $\mathbb{R}^d$ and still be distinguishable.

3.2.2 Mixed-membership stochastic block model

Lemma 2. The mixed membership stochastic block model is an instance of the generalised random dot product graph.

Proof. Recall Definition 4 for a mixed membership stochastic block model. Given a sample space $\mathcal{Z} = \{1, \ldots, K\}$ where each node has a probability distribution F_i over $\mathcal{Z}$, then, for $Z_{i\to j} \overset{\text{ind}}{\sim} F_i$, $Z_{j\to i} \overset{\text{ind}}{\sim} F_j$, a symmetric matrix

$\mathbf{A} \in \mathbb{R}^{n \times n}$ satisfies, for $i < j$,

$$\mathbf{A}_{ij} \mid Z_{i \to j}, Z_{j \to i} \overset{\text{ind}}{\sim} \text{Bernoulli}(\mathbf{B}_{Z_{i \to j} Z_{j \to i}}),$$

where $\mathbf{B} \in [0,1]^{K \times K}$ is the matrix of inter-community edge probabilities.

To construct the generalised random dot product graph, we define $\mathcal{X}$ as the simplex with corners given by the rows of the adjacency spectral embedding $(\mathbf{X_B})_1^\top, \ldots, (\mathbf{X_B})_K^\top$. The latent position X_i of a node is defined as a convex combination of these corners weighted by the probability a node is assigned to the k communities, $\pi_{ik} = \mathbb{P}(Z_{i \to j} = k)$ given by F_i,

$$X_i = \sum_{k=1}^{K} \pi_{ik} \left(\mathbf{X_B}\right)_k^\top.$$

The embedding X_i is a point within the simplex $(K-1)$-dimensional $\mathcal{X}$ with barycentric coordinates determined by the distribution F_i. The exact form of the distribution $F_{\mathcal{X}}$ will depend on the how the distributions F_i are assigned to each node in the mixed membership stochastic block model. The indefinite inner product of these latent positions is given by

$$\begin{aligned}
X_i^\top \mathbf{I}_{p,q} X_j &= \left\{ \sum_{k=1}^{K} \pi_{ik} \left(\mathbf{X_B}\right)_k^\top \right\}^\top \mathbf{I}_{p,q} \left\{ \sum_{k=1}^{K} \pi_{jk} \left(\mathbf{X_B}\right)_k^\top \right\} \\
&= \sum_{k,\ell=1}^{K} \pi_{ik} \pi_{j\ell} \left(\mathbf{X_B}\right)_k \mathbf{I}_{p,q} \left(\mathbf{X_B}\right)_\ell^\top \\
&= \sum_{k,\ell=1}^{K} \pi_{ik} \pi_{j\ell} \mathbf{B}_{k\ell} \\
&= \mathbb{P}(\mathbf{A}_{ij} = 1 \mid Z_i, Z_j).
\end{aligned}$$

Therefore, the probability of an edge $\mathbf{A}_{ij}$ is identical under the mixed membership stochastic block model and generalised random dot product graph. □

In the above construction of the generalised random dot product graph, the corner $(\mathbf{X_B})_k^\top$ corresponds to a node that has all the probability in the distribution F_i assigned to community k; $\pi_{ik} = 1$, $\pi_{i\ell} = 0$ for all $k \neq \ell$.

In Example 5, the adjacency spectral embedding of the three-community mixed membership stochastic block model form a noisy triangle, a two-dimensional simplex. In Section 3.1.2, we show how the noisy simplex formed by the embedding $\hat{\mathbf{X}}$ relates to the simplex with corners given by the latent positions $(\mathbf{X_B})_1^\top, \ldots, (\mathbf{X_B})_K^\top$.

The mean block matrix $\mathbf{B}$ being rank-deficient has an interesting effect on the embedding of the mixed membership stochastic block model. When the embedding dimension d is equal to $K - 1$, it is still possible to place a $(K-1)$-dimensional simplex in $\mathbb{R}^d$ as seen in Example 5. However, when $d < K - 1$, the embedding X_i can no longer be uniquely written as a convex combination of the simplex corners.

3.2.3 Degree-corrected stochastic block model

Lemma 3. The degree-corrected stochastic block model is an instance of the generalised random dot product graph.

Proof. Recall Definition 5 for a degree-corrected stochastic block model. Given a sample space $\mathcal{Z} = \{1, \ldots, K\}$ with probability distribution F, and a collection of probability distributions G_k on the interval $(0, 1]$ then, for $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} F$, a symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ satisfies, for $i < j$,

$$\mathbf{A}_{ij} \mid w_i, w_j, Z_i, Z_j \overset{\text{ind}}{\sim} \text{Bernoulli}(w_i w_j \mathbf{B}_{Z_i Z_j}),$$

where $\mathbf{B} \in [0, 1]^{K \times K}$ is the matrix of inter-community edge probabilities, and $w_i \mid Z_i \overset{\text{ind}}{\sim} G_{Z_i}$ is a node-specific weight parameter.

To construct the equivalent generalised random dot product graph, map the latent position Z_i to the corresponding row of the adjacency spectral

embedding $\mathbf{X_B}$,

$$X_i = w_i \left(\mathbf{X_B}\right)_{Z_i}^\top .$$

The latent space $\mathcal{X}$ consists of the K rays starting at the origin going to the endpoints $\left(\mathbf{X_B}\right)_1^\top, \ldots, \left(\mathbf{X_B}\right)_K^\top$. For a particular node, the latent variable Z_i says which ray the embedded point lines on, and the weight parameter w_i says how far the point lies along the ray. The distribution $F_{\mathcal{X}}$ is defined such that, for $X \sim F_{\mathcal{X}}$,

$$X = W_Z \left(\mathbf{X_B}\right)_Z^\top ,$$

for $Z \sim F$ and $W_k \sim G_k$, using uppercase to distinguish a weight drawn for a particular community, rather than a node-specific weight parameter.

By the construction of the adjacency spectral embedding $\mathbf{X_B}$,

$$X_i^\top \mathbf{I}_{p,q} X_j = w_i w_j \mathbf{B}_{Z_i Z_j},$$

therefore, the probability of an edge $\mathbf{A}_{ij}$ is identical under the degree-corrected stochastic block model and generalised random dot product graph. □

For a sensibly defined K-community degree-corrected stochastic block model, there are K rays of latent positions in the generalised random dot product graph. If two rays of points overlapped, that would imply that, for some $k \neq \ell$, the endpoint $\left(\mathbf{X_B}\right)_k^\top$ is positioned somewhere along the ray with endpoint $\left(\mathbf{X_B}\right)_\ell^\top$, meaning that $\left(\mathbf{X_B}\right)_k^\top = c\left(\mathbf{X_B}\right)_\ell^\top$ for some $c \leq 1$. Therefore, any node in community k could instead be described as a node in community ℓ with new weight parameter cw_i. It is assumed that the degree-corrected stochastic block model uses as few communities as necessary.

Figure 3.3 shows an example adjacency spectral embedding of a two-community degree-corrected stochastic block model forming two noisy rays of points starting from the origin and ending at the points giving by the points $\left(\mathbf{X_B}\right)_1^\top, \ldots, \left(\mathbf{X_B}\right)_K^\top$.

Example 6 showed that the degree-corrected stochastic block model can be considered as a special case of the mixed membership stochastic block model. However, as with the stochastic block model, if the mean block matrix $\mathbf{B}$ is rank-deficient, it is possible for the K rays of points to coexist in $\mathbb{R}^d$ and still be distinguishable.

3.3 Asymptotic results

This section describes how the adjacency spectral embedding for each node $\hat{X}_i$ converges to the latent position of the generalised random dot product graph X_i. Recall that the latent positions of the generalised random dot product graph X_i can all be replaced by $\mathbf{Q}X_i$ for any indefinite orthogonal matrix $\mathbf{Q} \in \mathbb{O}(p, q)$ to create an equivalent parameterisation of the model. Therefore, the adjacency spectral embedding $\hat{X}_i$ can only hope to estimate the latent position X_i up to some transformation by $\mathbf{Q}$. This needs to be taken into account when describing the asymptotic theory.

There exists an indefinite orthogonal matrix $\mathbf{Q} \in \mathbb{O}(p, q)$ such that the adjacency spectral embedding is uniform consistent meaning that even the largest difference between X_i and its estimate $\mathbf{Q}\hat{X}_i$ across all nodes tends to zero with high probability. This is known as the two-to-infinity norm, the maximal Euclidean distance between X_i and $\mathbf{Q}\hat{X}_i$ [11]. Furthermore, a central limit theorem shows that this convergence is with an asymptotically Gaussian error that we can quantify.

In order to obtain these convergence results, we must first make an assumption that the generalised random dot product graph is somehow sensibly defined. This is done via the second moment matrix $\mathbf{\Delta} = \mathbb{E}(XX^\top) \in \mathbb{R}^{d \times d}$ where $X \sim F_{\mathcal{X}}$. This matrix plays a key role in the asymptotic distribution of the adjacency spectral embedding.

Assumption 1 (Minimal dimensionality)**.** For $X \sim F_X$, the second moment matrix $\mathbf{\Delta} = \mathbb{E}(XX^\top)$ has full rank d.

Assumption 1 is not a real constraint on the generalised random dot product graph, existing to ensure that the latent positions $X_i \in \mathbb{R}^d$ have been defined economically. If $\mathbf{\Delta}$ is not full rank d, then it is possible to define new latent positions $X_i' \in \mathbb{R}^{d'}$ with smaller dimension $d' < d$ such that the resulting second moment matrix $\mathbf{\Delta}' \in \mathbb{R}^{d' \times d'}$ is full rank.

For Theorem 1 below, we say that a sequence of random variables X_n converges in probability to the random variable X, denoted by $X_n \xrightarrow{\mathbb{P}} X$, if for all $\epsilon > 0$,

$$\mathbb{P}(|X_n - X| > \epsilon) \to 0,$$

as $n \to \infty$. For Theorem 2, let $\Phi(\mathbf{z}, \mathbf{\Sigma})$ denote the cumulative distribution function of a multivariate Gaussian random variable with mean $\mathbf{0}$ and covariance $\mathbf{\Sigma}$. For $Z \sim \text{Normal}(\mathbf{0}, \mathbf{\Sigma})$, $\Phi(\mathbf{z}, \mathbf{\Sigma})$ is the probability that Z is less than or equal to $\mathbf{z}$ for all components, $\Phi(\mathbf{z}, \mathbf{\Sigma}) = \mathbb{P}(Z \leq \mathbf{z})$.

The following two theorems were first stated and proved in Rubin-Delanchy et al. [60]. We provide a sketch proof highlighting the main ideas of the results.

Theorem 1 (Adjacency spectral embedding uniform consistency)**.** Let $\mathbf{A}$ be an instance of a generalised random dot product graph satisfying Assumption 1. There exists a sequence of matrices $\mathbf{Q}_n \in \mathbb{O}(p, q)$ such that

$$\max_{i \in \{1,\ldots,n\}} \left\| \mathbf{Q}_n \hat{X}_i - X_i \right\| \xrightarrow{\mathbb{P}} 0.$$

Theorem 2 (Adjacency spectral embedding central limit theorem)**.** Let $\mathbf{A}$ be an instance of a generalised random dot product graph satisfying Assumption 1. Given $x \in \mathcal{X}$, define the covariance-valued function

$$\mathbf{\Sigma}(x) = \mathbf{I}_{p,q} \mathbf{\Delta}^{-1} \mathbb{E}\left[(x^\top \mathbf{I}_{p,q} X)(1 - x^\top \mathbf{I}_{p,q} X) X X^\top \right] \mathbf{\Delta}^{-1} \mathbf{I}_{p,q},$$

where $X \sim F_{\mathcal{X}}$. Then, there exists a sequence of matrices $\mathbf{Q}_n \in \mathbb{O}(p, q)$ such that, for all $\mathbf{z} \in \mathbb{R}^d$,

$$\mathbb{P}\left(n^{1/2}\left(\mathbf{Q}_n\hat{X}_i - X_i\right) \leq \mathbf{z} \mid X_i = x\right) \to \Phi\left(\mathbf{z}, \mathbf{\Sigma}(x)\right).$$

The original paper gives the convergence rate for uniform consistency [60], but for the results in this thesis we are only concerned that the limit tends to zero. The result was later strengthen by explaining how the sequence of matrices $\mathbf{Q}_n$ converge up to some other transformation [2]. We shall see this in the equivalent central limit theorem for dynamic networks in Chapter 6, which was developed once this discovery was made.

Sketch proof. This proof outline brings together ideas from similar arguments from several papers involving some form of the generalised random dot product graph [60, 21, 36, 22]. Theorem 1 and 2 are proved simultaneously by breaking down the connection between the spectral embedding $\hat{\mathbf{X}}$ and the latent positions $\mathbf{X}$ as a series of matrix perturbations. Here we describe how the sequence of matrices $\mathbf{Q}_n$ are constructed.

For a fixed n, let $\mathbf{P} \in [0, 1]^{n \times n}$ be the matrix of probabilities that two nodes in the network form an edge,

$$\mathbf{P}_{ij} = \mathbb{E}(\mathbf{A}_{ij}) = X_i^\top \mathbf{I}_{p,q} X_j.$$

The adjacency matrix $\mathbf{A}$ can be thought of as a noisy observation of the probability matrix $\mathbf{P}$.

The aim is to construct $\mathbf{Q}_n$ in two steps by aligning the adjacency spectral embedding $\hat{\mathbf{X}}$ to the adjacency spectral embedding of the matrix $\mathbf{P}$ given by $\mathbf{X_P}$, before aligning $\mathbf{X_P}$ to the latent positions $\mathbf{X}$. The first step is to find $\mathbf{W}$ such that $\hat{\mathbf{X}}\mathbf{W}^\top \approx \mathbf{X_P}$. This is done by solving a one-mode orthogonal Procrustes problem [69] to find an orthogonal matrix $\mathbf{Q}$ that best aligns the left and right singular vectors of the matrices $\mathbf{A}$ and $\mathbf{P}$,

$$\mathbf{W} = \underset{\mathbf{Q} \in \mathbb{O}(d)}{\arg\min} \left\{ \|\mathbf{U} - \mathbf{U_P}\mathbf{Q}\|_{\mathrm{F}}^2 + \|\mathbf{V} - \mathbf{V_P}\mathbf{Q}\|_{\mathrm{F}}^2 \right\},$$

where $\|\cdot\|_{\mathrm{F}}^2$ represents the Frobenius norm. The solution is given by $\mathbf{W} = \mathbf{W}_1\mathbf{W}_2^\top$ constructed using the singular value decomposition

$$\mathbf{U}_{\mathbf{P}}^\top\mathbf{U} + \mathbf{V}_{\mathbf{P}}^\top\mathbf{V} = \mathbf{W}_1\mathbf{\Sigma}\mathbf{W}_2^\top,$$

when it can also be shown that $\mathbf{W} \in \mathbb{O}(p, q)$.

For the second step, note that $\mathbf{X}_{\mathbf{P}}\mathbf{I}_{p,q}\mathbf{X}_{\mathbf{P}}^\top = \mathbf{P} = \mathbf{X}\mathbf{I}_{p,q}\mathbf{X}^\top$ meaning there exists $\mathbf{Q}_{\mathbf{X}} \in \mathbb{O}(p, q)$ such that $\mathbf{X}_{\mathbf{P}} = \mathbf{X}\mathbf{Q}_{\mathbf{X}}$. Together with the first step, this gives $\hat{\mathbf{X}}\mathbf{W}^\top\mathbf{Q}_{\mathbf{X}}^{-1} \approx \mathbf{X}$ which gives the matrix $\mathbf{Q}_n = \mathbf{W}^\top\mathbf{Q}_{\mathbf{X}}^{-1}$ in Theorem 1 and 2.

The bulk of the two proofs revolves around quantifying the expression $\hat{\mathbf{X}}\mathbf{Q}_n \approx \mathbf{X}$ more precisely. After some matrix perturbation decompositions, we arrive at the expression

$$n^{1/2}(\hat{\mathbf{X}}\mathbf{Q}_n - \mathbf{X}) = n^{1/2}(\mathbf{A} - \mathbf{P})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{I}_{p,q} + n^{1/2}\mathbf{R}\mathbf{Q}_n,$$

where $\mathbf{R} \in \mathbb{R}^{n\times d}$ is some residual matrix whose behaviour can be well-controlled. Theorem 1 is obtained by bounding the two-to-infinity norm of the right hand side [11], while Theorem 2 is obtained by applying the standard central limit theorem on each row of the matrix $n^{1/2}(\mathbf{A}-\mathbf{P})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{I}_{p,q}$. More details can be found in the Appendix of Rubin-Delanchy et al. [60]. □

3.3.1 Stochastic block model asymptotic distribution

When showing that the stochastic block model is an instance of the generalised random dot product graph in Lemma 1, we considered the latent positions

$$X_i = (\mathbf{X}_{\mathbf{B}})_{Z_i}^\top.$$

Under this construction, there are only K possible latent positions a node can take. Theorem 1 and 2 show that the adjacency spectral embedding $\hat{X}_i$ converges to these K latent positions consistently with asymptotically Gaussian

error up to an indefinite orthogonal transformation $\mathbf{Q}$. The covariance-valued function $\mathbf{\Sigma}(x)$ only needs to be computed for the K latent positions in $\mathcal{X}$, where we denote

$$\mathbf{\Sigma}_k = \mathbf{\Sigma}((\mathbf{X_B})_k^\top)$$

to be the covariance corresponding to community k.

From Theorem 1 and 2, the points $\mathbf{Q}_n\hat{X}_i$ have an approximate Gaussian mixture model distribution with components Normal$((\mathbf{X_B})_k, \mathbf{\Sigma}_k/n)$ with components probability π_k. Since $\mathbf{Q}_n$ is a linear transformation, this implies that the adjacency spectral embedding $\hat{\mathbf{X}}$ also has an approximate Gaussian mixture model distribution, which can be seen in Figure 3.1. This motivates the choice of using a Gaussian mixture model to cluster points in an adjacency spectral embedding to find community structure compared to an algorithm that does not account for unequal and non-spherical cluster covariances.

In the remainder of this section, we show how the parameters of the approximate Gaussian mixture model are computed for a stochastic block model. For $X \sim F_{\mathcal{X}}$, the second moment matrix is given by

$$\begin{aligned}\mathbf{\Delta} &= \mathbb{E}(XX^\top) \\ &= \sum_{k=1}^{K} \pi_k \left(\mathbf{X_B}\right)_k^\top \left(\mathbf{X_B}\right)_k \\ &= \mathbf{X_B}^\top \mathbf{\Pi} \mathbf{X_B},\end{aligned}$$

where $\mathbf{\Pi} = \text{diag}(\boldsymbol{\pi}) \in \mathbb{R}^{K\times K}$ is the diagonal matrix consisting of the community assignment probabilities. To compute the covariance $\mathbf{\Sigma}_k$ corresponding

to latent position $x = (\mathbf{X_B})_k^\top$, we need to compute the following expectation,

$$\begin{aligned}
&\mathbb{E}\left[\left(x^\top \mathbf{I}_{p,q} X\right)\left(1 - x^\top \mathbf{I}_{p,q} X\right) X X^\top\right] \\
&= \sum_{\ell=1}^{K} \pi_\ell \left((\mathbf{X_B})_k \mathbf{I}_{p,q} (\mathbf{X_B})_\ell^\top\right)\left(1 - (\mathbf{X_B})_k \mathbf{I}_{p,q} (\mathbf{X_B})_\ell^\top\right)(\mathbf{X_B})_\ell^\top (\mathbf{X_B})_\ell \\
&= \sum_{\ell=1}^{K} \pi_\ell \mathbf{B}_{k\ell}\left(1 - \mathbf{B}_{k\ell}\right)(\mathbf{X_B})_\ell^\top (\mathbf{X_B})_\ell \\
&= \mathbf{X_B}^\top \mathbf{\Pi} \mathbf{\Lambda}_k \mathbf{X_B},
\end{aligned}$$

where $\mathbf{\Lambda}_k = \operatorname{diag}(\mathbf{B}_{k1}(1-\mathbf{B}_{k1}), \ldots, \mathbf{B}_{kK}(1-\mathbf{B}_{kK})) \in \mathbb{R}^{K\times K}$. Recall that $\mathbf{B}_{k\ell}$ is the mean of an edge between communities k and ℓ, meaning $\mathbf{B}_{k\ell}(1-\mathbf{B}_{k\ell})$ is the corresponding variance term. Therefore, $\mathbf{\Lambda}_k$ is the diagonal matrix with entries equal to the variances for edges involving community k. Piecing these terms together gives the covariance matrix

$$\mathbf{\Sigma}_k = \mathbf{I}_{p,q} \mathbf{\Delta}^{-1}\left(\mathbf{X_B}^\top \mathbf{\Pi} \mathbf{\Lambda}_k \mathbf{X_B}\right)\mathbf{\Delta}^{-1} \mathbf{I}_{p,q}.$$

Example 9 (Two-community stochastic block model continued)**.** We compute the asymptotic Gaussian mixture model for the adjacency spectral embedding of the two-community stochastic block model described in Example 4. Firstly, the asymptotic distribution for the points $\mathbf{Q}_n \hat{X}_i$ is constructed by computing the Gaussian components $\text{Normal}((\mathbf{X_B})_k, \mathbf{\Sigma}_k/n)$ for each community. In this case $\mathbf{Q}_n$ can be constructed since we know the matrix of probabilities $\mathbf{P}$ used to generate the adjacency matrix $\mathbf{A}$. Transformation by $\mathbf{Q}_n$ provides a Gaussian mixture model for the adjacency spectral embedding $\hat{\mathbf{X}}$.

Figure 3.4 shows the asymptotic distribution for the embedding $\hat{\mathbf{X}}$ of the stochastic block model. Points are coloured according to their community assignment Z_i. The ellipses show the 95% contours of the asymptotic Gaussian components, which fit the adjacency spectral embedding points well. ◁

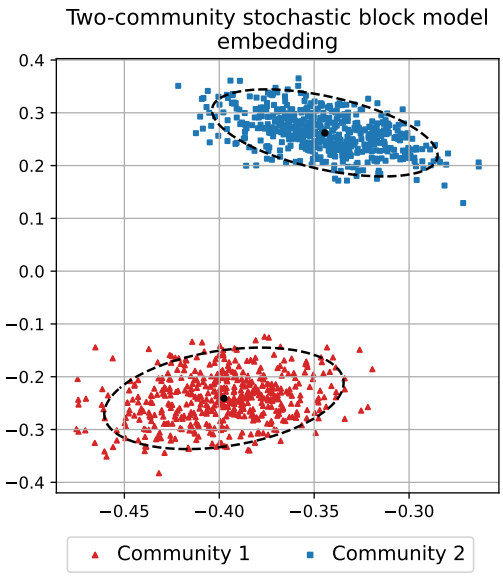

Figure 3.4: The embedding $\hat{\mathbf{X}}$ of a two-community stochastic block model into $d = 2$ dimensions. The points are coloured according to their true community assignment Z_i. The ellipses show the 95% contours of the asymptotic Gaussian components.

3.3.2 Mixed membership stochastic block model asymptotic distribution

When showing that the mixed membership stochastic block model is an instance of the generalised random dot product graph in Lemma 2, we considered the latent positions

$$X_i = \sum_{k=1}^{K} \pi_{ik} \left(\mathbf{X_B}\right)_k^\top .$$

There are now infinitely many latent positions to consider within $\mathcal{X}$, the simplex with corners given by the latent positions $\left(\mathbf{X_B}\right)_1^\top, \ldots, \left(\mathbf{X_B}\right)_K^\top$. To describe the asymptotic distribution of the adjacency spectral embedding,

we need to be able to compute the covariance structure for any point in the simplex.

To do this, we need to know how nodes are assigned their probability distribution F_i over the K possible communities. In Definition 4 for the mixed membership stochastic block model, no assumption was made about these distributions. Often the probability masses for the distributions F_i are sampled from a Dirichlet distribution with parameter $\boldsymbol{\alpha}$, part of the original definition of a mixed membership stochastic block model [3],

$$\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK}) \overset{\text{ind}}{\sim} \text{Dirichlet}(\boldsymbol{\alpha}).$$

The distribution $F_{\mathcal{X}}$ over the simplex $\mathcal{X}$ is defined inheriting this distribution; if $\boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{\alpha})$, then $X = \sum_{k=1}^{K} \pi_k \left(\mathbf{X_B}\right)_k^\top \sim F_{\mathcal{X}}$.

For $\boldsymbol{\rho} = (\rho_1 \ldots, \rho_K) \in \mathbb{R}^K$ in the standard simplex where $\rho_k \geq 0$ and $\sum_{k=1}^{K} \rho_k = 1$, we compute the covariance $\boldsymbol{\Sigma}(x)$ for $x = \sum_{k=1}^{K} \rho_k \left(\mathbf{X_B}\right)_k^\top \in \mathcal{X}$. This is more involved than the corresponding calculation for the stochastic block model as it requires the moments of the Dirichlet distribution. For $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$, these can be neatly expressed using the gamma function,

$$\mathbb{E}\left(\prod_{k=1}^{K} \pi_k^{\beta_k}\right) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k + \beta_k)} \times \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + \beta_k)}{\Gamma(\alpha_k)}.$$

Using these moments, the second moment matrix is given by

$$\begin{aligned}
\boldsymbol{\Delta} &= \mathbb{E}(XX^\top) \\
&= \sum_{k,\ell=1}^{K} \mathbb{E}(\pi_k \pi_\ell) \left(\mathbf{X_B}\right)_k^\top \left(\mathbf{X_B}\right)_\ell \\
&= \frac{1}{\alpha_0(\alpha_0+1)} \left(\sum_{k=1}^{K} \alpha_k(\alpha_k+1) \left(\mathbf{X_B}\right)_k^\top \left(\mathbf{X_B}\right)_k + \sum_{k \neq \ell}^{K} \alpha_k \alpha_\ell \left(\mathbf{X_B}\right)_k^\top \left(\mathbf{X_B}\right)_\ell \right) \\
&= \frac{1}{\alpha_0(\alpha_0+1)} \mathbf{X_B}^\top \left(\text{diag}(\boldsymbol{\alpha}) + \boldsymbol{\alpha}^\top \boldsymbol{\alpha} \right) \mathbf{X_B}
\end{aligned}$$

For $X = \sum_{k=1}^{K} \pi_k \left(\mathbf{X_B}\right)_k^\top \in \mathcal{X}$, the indefinite inner product is given by

$$\begin{aligned} x^\top \mathbf{I}_{p,q} X &= \sum_{k,\ell=1}^{K} \rho_k \pi_\ell \left(\mathbf{X_B}\right)_k \mathbf{I}_{p,q} \left(\mathbf{X_B}\right)_\ell^\top \\ &= \sum_{k,\ell=1}^{K} \rho_k \pi_\ell \mathbf{B}_{k\ell}. \end{aligned}$$

Finally, using this, the expectation in the covariance term is given by

$$\begin{aligned} &\mathbb{E}\left[\left(x^\top \mathbf{I}_{p,q} X\right)\left(1 - x^\top \mathbf{I}_{p,q} X\right) X X^\top\right] \\ &= \mathbb{E}\left[\left(\sum_{k,\ell=1}^{K} \rho_k \pi_\ell \mathbf{B}_{k\ell}\right)\left(1 - \sum_{k,\ell=1}^{K} \rho_k \pi_\ell \mathbf{B}_{k\ell}\right)\left(\sum_{k,\ell=1}^{K} \pi_k \pi_\ell \left(\mathbf{X_B}\right)_k^\top \left(\mathbf{X_B}\right)_\ell\right)\right] \end{aligned}$$

There is not much insight to be gained by continuing this calculation; it is a complicated expression involving the third and fourth moments of the Dirichlet distribution. However, it gives some indication about how the calculation of $\mathbf{\Sigma}(x)$ can be performed on a computer, combining this equation with the second moment matrix $\mathbf{\Delta}$, as shown in the following example.

Example 10 (Three-community mixed membership stochastic block model continued)**.** We compute the asymptotic distribution for the adjacency spectral embedding of the three-community mixed membership stochastic block model described in Example 5. To visualise this, we aim to find the asymptotic distribution for the points at the corners of the simplex, which correspond to a node where all the probability mass in F_i is assigned to a single community. These are constructed using the equations above then transformed using the $\mathbf{Q}_n$ from Theorem 1 and 2 to get the asymptotic distribution for the adjacency spectral embedding $\hat{\mathbf{X}}$ at those positions.

Figure 3.5 shows the embedding $\hat{\mathbf{X}}$ of the mixed membership stochastic block model. Points are coloured a mixture of red, blue and green according to their probability distribution F_i over the communities. The solid black

triangle represents the simplex $\mathcal{X}$ after transformation by $\mathbf{Q}_n^{-1}$ corresponding to the points $\boldsymbol{\rho} = (1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$ in the standard simplex. The ellipses show the 95% contours of the asymptotic Gaussian distribution at the corners of this transformed simplex.

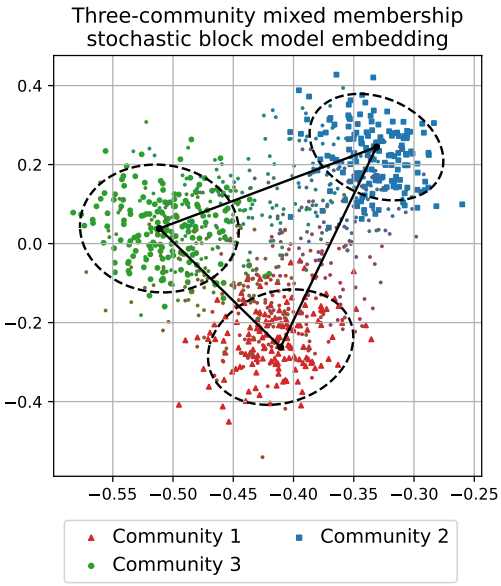

Figure 3.5: The embedding $\hat{\mathbf{X}}$ of a three-community mixed membership stochastic block model. The points are coloured a mixture of red (triangle), blue (square) and green (circle) according to their probability distribution F_i over the communities. The lines show the transformed simplex $\mathcal{X}$ and the ellipses show the 95% contours of the asymptotic Gaussian components at the corners of the simplex.

Given the choice of Dirichlet distribution parameter $\boldsymbol{\alpha} = (0.2, 0.2, 0.2)$, many points in the embedding, shown by red triangles, blue squares and green circles, are very close to the simplex corners. These points fit the asymptotic distributions particularly well. ◁

3.3.3 Degree-corrected stochastic block model asymptotic distribution

When showing that the degree-corrected stochastic block model is an instance of the generalised random dot product graph in Lemma 2, we considered the latent positions

$$X_i = w_i \left(\mathbf{X_B}\right)_{Z_i}^\top .$$

Recall that the distribution $F_{\mathcal{X}}$ is defined such that, for $X \sim F_{\mathcal{X}}$,

$$X = W_Z \left(\mathbf{X_B}\right)_Z^\top ,$$

where $Z \sim F$ and $W_k \sim G_k$ for $k = 1, \ldots, K$. Using $\mathbb{E}_k(\cdot)$ to denote expectation with respect to the distribution G_k, we can compute the second moment matrix,

$$\begin{aligned}
\mathbf{\Delta} &= \mathbb{E}(XX^\top) \\
&= \sum_{k=1}^{K} \pi_k \mathbb{E}_k(W_k^2) \left(\mathbf{X_B}\right)_k^\top \left(\mathbf{X_B}\right)_k \\
&= \mathbf{X_B^\top \Pi \Gamma X_B},
\end{aligned}$$

where $\mathbf{\Gamma} = \operatorname{diag}(\mathbb{E}_1(W_1^2), \ldots, \mathbb{E}_K(W_K^2)) \in \mathbb{R}^{K \times K}$ is the diagonal matrix of second moments for the weight parameter for each community. To compute the covariance $\mathbf{\Sigma}(x)$ for $x = w \left(\mathbf{X_B}\right)_k^\top \in \mathcal{X}$, we need to compute the following expectation,

$$\begin{aligned}
&\mathbb{E}\left[\left(x^\top \mathbf{I}_{p,q} X\right)\left(1 - x^\top \mathbf{I}_{p,q} X\right) XX^\top\right] \\
&= \sum_{\ell=1}^{K} \pi_\ell \mathbb{E}_\ell \left[w W_\ell \mathbf{B}_{k\ell} (1 - w W_\ell \mathbf{B}_{k\ell}) W_\ell^2 \left(\mathbf{X_B}\right)_\ell^\top \left(\mathbf{X_B}\right)_\ell \right] \\
&= \sum_{\ell=1}^{K} \pi_\ell \left[w \mathbb{E}_\ell(W_\ell^3) \mathbf{B}_{k\ell} - w^2 \mathbb{E}_\ell(W_\ell^4) \mathbf{B}_{k\ell}^2 \right] \left(\mathbf{X_B}\right)_\ell^\top \left(\mathbf{X_B}\right)_\ell \\
&= \mathbf{X_B^\top \Pi \Lambda}_k(w) \mathbf{X_B},
\end{aligned}$$

where $\mathbf{\Lambda}_k(w) \in \mathbb{R}^{K\times K}$ is the diagonal matrix where $(\mathbf{\Lambda}_k(w))_{\ell\ell} = w\mathbb{E}_\ell(W_\ell^3)\mathbf{B}_{k\ell} - w^2\mathbb{E}_\ell(W_\ell^4)\mathbf{B}_{k\ell}^2$. The asymptotic covariance $\mathbf{\Sigma}(x)$ can be calculated by combining this equation with the second moment matrix $\mathbf{\Delta}$.

Example 11 (Two-community degree-corrected stochastic block model continued)**.** We compute the asymptotic distribution for the adjacency spectral embedding of the two-community degree-corrected stochastic block model described in Example 7. These are constructed using the equations above then transformed using the $\mathbf{Q}_n$ from Theorem 1 and 2 to get the asymptotic distribution for the adjacency spectral embedding $\hat{\mathbf{X}}$ at those positions.

Figure 3.6 shows the embedding $\hat{\mathbf{X}}$ of the stochastic block model. Points are coloured according to their community assignment Z_i. The solid black lines represent the rays of latent positions $\mathcal{X}$ after transformation by $\mathbf{Q}_n^{-1}$.

The ellipses show the 95% contours of the asymptotic Gaussian distribution for two points along each ray for the two communities, $x = w(\mathbf{X_B})_k^\top$ for $w = 0.5$ and $w = 1$. Given the form of $\mathbf{\Lambda}_k(w)$, the covariance $\mathbf{\Sigma}(x)$ tends to get smaller as w gets closer to zero. This can be seen in Figure 3.6 where points have smaller error closer to the origin. ◁

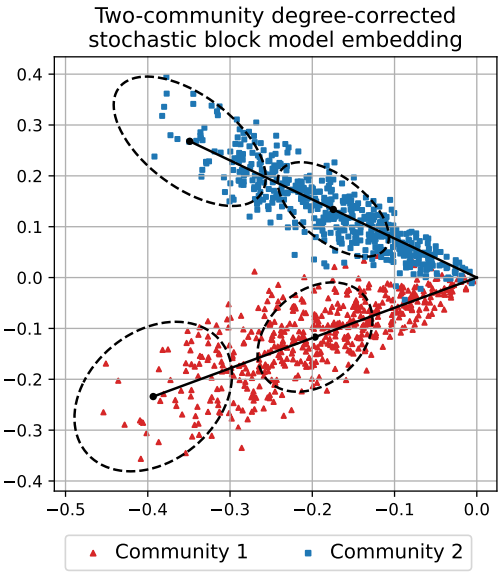

Figure 3.6: The embedding $\hat{\mathbf{X}}$ of a two-community degree-corrected stochastic block model into $d = 2$ dimensions. The points are coloured according to their true community assignment Z_i. The lines show the transformed rays of $\mathcal{X}$ and the ellipses show the 95% contours of the asymptotic Gaussian components for latent positions corresponding to $w = 0.5$ and $w = 1$.

Chapter 4

Weighted networks

In this chapter, we extend the generalised random dot product model to allow for weighted networks. This requires a new way of thinking as the latent position now must encapsulate information about a family of distributions rather than edge probabilities. This adds an extra layer of abstraction between the latent positions and an instance of a weighted generalised random dot product graph, which must be included when linking this model to adjacency spectral embedding. The work was first published in Gallagher et al. [21].

4.1 Weighted graph models

In order to model weighted networks, we need a collection of distributions that can be used to generate the edge weights.

Definition 7 (Symmetric family of distributions)**.** For a set $\mathcal{Z}$, $\{H(z_1, z_2) : z_1, z_2 \in \mathcal{Z}\}$ is a *symmetric family of real-valued distributions*, if $H(z_1, z_2)$ is a distribution on $\mathbb{R}$ for all $z_1, z_2 \in \mathcal{Z}$ such that $H(z_1, z_2) = H(z_2, z_1)$.

The distributions in $\{H(z_1, z_2) : z_1, z_2 \in \mathcal{Z}\}$ do not need to be from the same parameter family of distributions, for example, it could contain both discrete and continuous distributions. This allows for a lot of flexibility for the weighted graph models introduced in the remainder of the chapter.

The adjacency spectral embedding of a weighted graph is encapsulated by the mean and variance of the distributions in the family $\{H(z_1, z_2) : z_1, z_2 \in$

$\mathcal{Z}\}$, similar to how Theorem 1 and 2 rely on the mean and variance of the Bernoulli distributions in the generalised random dot product graph. The following example shows how the unweighted setting can be written using this new notation.

Example 12 (Symmetric family of Bernoulli distributions)**.** The family of distributions used for the stochastic block model in Definition 3 is a symmetric family of distributions over $\mathcal{Z} = \{1, \ldots, K\}$. For all $z_1, z_2 \in \mathcal{Z}$,

$$H(z_1, z_2) = \text{Bernoulli}(\mathbf{B}_{z_1 z_2}). \qquad \triangleleft$$

In this section, the three versions of the stochastic block model described in Section 3.1 are updated to allow for weighted edges. In each case, the symmetric family of distributions are initially defined over a finite number of latent positions $\mathcal{Z} = \{1, \ldots, K\}$ and the mean and variance of the distributions in the family $\{H(z_1, z_2) : z_1, z_2 \in \mathcal{Z}\}$ can be written in a convenient manner.

Definition 8 (Block mean and variance matrices)**.** For $\mathcal{Z} = \{1, \ldots, K\}$ and a symmetric family of distributions $\{H(z_1, z_2) : z_1, z_2 \in \mathcal{Z}\}$, denote $\mathbf{B} \in \mathbb{R}^{K \times K}$ and $\mathbf{C} \in \mathbb{R}_+^{K \times K}$ respectively as the *block mean matrix* and *block variance matrix*, where, for $X \sim H(k, \ell)$, $\mathbf{B}_{k\ell} = \mathbb{E}(X)$ and $\mathbf{C}_{k\ell} = \text{var}(X)$, if those moments exist.

4.1.1 Weighted stochastic block model

We define the most general version of the weighted stochastic block model. Previous work has focused on weighted homogeneous stochastic block models where there is a fixed behaviour for within community edges and between community edges [35, 76] and labelled stochastic block models [29, 41].

Definition 9 (Weighted stochastic block model)**.** Let $\mathcal{Z} = \{1, \ldots, K\}$ be a sample space with probability distribution F and $\{H(z_1, z_2) : z_1, z_2 \in \mathcal{Z}\}$ be a

symmetric family of real-valued distributions. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is distributed as a *weighted stochastic block model*, if $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} F$ and, for $i < j$,

$$\mathbf{A}_{ij} \mid Z_i, Z_j \overset{\text{ind}}{\sim} H(Z_i, Z_j).$$

Unsurprisingly, given Example 12, the standard stochastic block model from Definition 3 is also a weighted stochastic block model. In the same way as before, given $Z \sim F$, we define $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ where $\pi_k = \mathbb{P}(Z = k)$ is the probability a node is assigned to community k.

The sample space for this model is $\mathcal{Z} = \{1, \ldots, K\}$ so we can compute the block mean and variance matrices $\mathbf{B}$ and $\mathbf{C}$ using Definition 8. In the standard stochastic block model, this definition returns the same mean block matrix $\mathbf{B}$ previously defined as part of Definition 3. The success probability p of a Bernoulli distribution is also the mean. We have also seen the variance block matrix $\mathbf{C}$ previously appear when discussing the standard stochastic block model. In Section 3.3.1, when calculating the asymptotic distribution of the adjacency spectral embedding, the covariance term $\boldsymbol{\Sigma}_k$ involved the variance of the Bernoulli distribution, $\mathbf{C}_{k\ell} = \mathbf{B}_{k\ell}(1 - \mathbf{B}_{k\ell})$.

Example 13 (Two-community Poisson distribution stochastic block model)**.** Consider a weighted stochastic block model with two communities where matrix entries are Poisson random variables. The symmetric family of distributions $\{H(z_1, z_2) : z_1, z_2 \in \mathcal{Z}\}$ is given by

$$H(z_1, z_2) = \begin{cases} \text{Poisson}(\lambda_1) & \text{if } z_1 = z_2, \\ \text{Poisson}(\lambda_2) & \text{if } z_1 \neq z_2, \end{cases}$$

where $\lambda_1 \neq \lambda_2$.

We generate a weighted stochastic block model with $n = 1000$ nodes, $\lambda_1 = 0.5$, $\lambda_2 = 0.6$ and a community assignment distribution F such that

$\pi_1 = 1/2$ and $\pi_2 = 1/2$. The block mean and variance matrices are given by

$$\mathbf{B} = \begin{pmatrix} \lambda_1 & \lambda_2 \\ \lambda_2 & \lambda_1 \end{pmatrix}, \ \mathbf{C} = \begin{pmatrix} \lambda_1 & \lambda_2 \\ \lambda_2 & \lambda_1 \end{pmatrix}.$$

Figure 4.1 shows the embedding $\hat{\mathbf{X}}$ of the weighted stochastic block model in $d = 2$ dimensions. The points form two clusters based on their community assignment Z_i. ◁

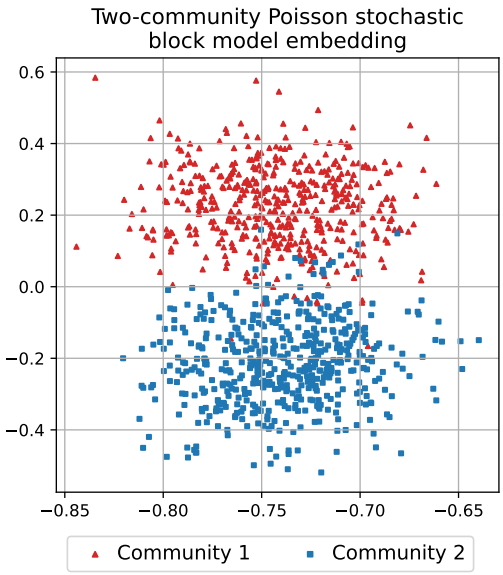

Figure 4.1: The embedding $\hat{\mathbf{X}}$ of a two-community weighted stochastic block model into $d = 2$ dimensions. The points are coloured according to their true community assignment Z_i.

Example 14 (Two-community Gaussian distribution stochastic block model with equal means)**.** Consider a weighted stochastic block model with two communities where matrix entries are Gaussian random variables with symmetric family of distributions $\{H(z_1, z_2) : z_1, z_2 \in \mathcal{Z}\}$, where

$$H(z_1, z_2) = \text{Normal}(1, \sigma^2_{z_1 z_2}).$$

The mean of the distributions are the same regardless of the community assignments, only the variance changes. In this example, $\mathbf{B}$ is equal to the all-one matrix and $\text{rank}(\mathbf{B}) = 1$.

We generate a weighted stochastic block model with $n = 1000$ nodes, $\sigma^2_{k\ell} = 2$ if $k = 1, \ell = 1$, and $\sigma^2_{k\ell} = 1$ otherwise, with community assignment distribution F such that $\pi_1 = 1/2$ and $\pi_2 = 1/2$. Figure 4.2 shows the embedding $\hat{\mathbf{X}}$ of the weighted stochastic block model in $d = 1$ dimensions, the rank of the block mean matrix. The empirical distributions for the two communities are approximately centred around -1.0 but the variances are different; the embedding for nodes with $Z_i = 1$ having a larger variance compared to the embedding for nodes with $Z_i = 2$.

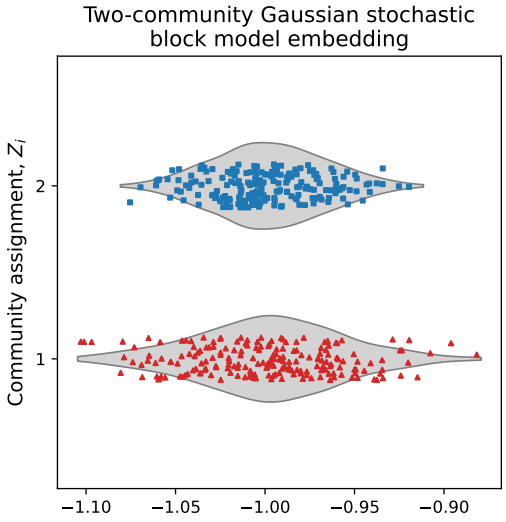

Figure 4.2: The embedding $\hat{\mathbf{X}}$ of a two-community weighted stochastic block model into $d = 1$ dimensions. The violin plot show the empirical distribution of the points separated by their true community assignment Z_i. A sample of 200 jittered points are shown coloured according to their true community assignment Z_i.

Suppose we did not know the true community assignment of the nodes

within the graph. While it is not possible to separate the nodes into two clusters, fitting a Gaussian mixture model would identify the embedding as a mixture of two components. Points in the tail of the distribution are more likely to correspond to one community than the other, one can at least achieve better-than-random community recovery. ◁

4.1.2 Weighted mixed membership stochastic block model

Creating a weighted version of the mixed membership stochastic block model follows the same procedure as creating a weighted version of the stochastic block model.

Definition 10 (Weighted mixed membership stochastic block model)**.** Let $\mathcal{Z} = \{1, \ldots, K\}$ be a sample space with probability distribution F and $\{H(z_1, z_2) : z_1, z_2 \in \mathcal{Z}\}$ be a symmetric family of real-valued distributions. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is distributed as a *weighted mixed membership stochastic block model*, if $Z_{i \to j} \overset{\text{ind}}{\sim} F_i$ and, for $i < j$,

$$\mathbf{A}_{ij} \mid Z_{i \to j}, Z_{j \to i} \overset{\text{ind}}{\sim} H(Z_{i \to j}, Z_{j \to i}).$$

As with the standard mixed membership stochastic block model, given $Z_i \sim F_i$, we define $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK})$, where $\pi_{ik} = \mathbb{P}(Z_i = k)$ is the probability a node is assigned to community k.

4.1.3 Zero-inflated stochastic block model

The degree-corrected stochastic block model extends the stochastic block model by introducing a node-specific weight parameter w_i which represents their activity. With probability $w_i w_j$, both nodes are willing to form an edge, in which case, an edge is sampled using the stochastic block model.

The zero-inflated stochastic block model applies this interpretation to the weighted stochastic block model. Rather than the weight parameter w_i

adjusting the parameters within the distribution, it models the probability that an edge is made. If both endpoints are willing to form an edge, one is sampled using the distribution in the weighted stochastic block model, otherwise the edge weight is zero representing no connection.

Given a distribution F and probability π, the zero-inflated distribution is defined as the following mixture distribution,

$$(1-\pi)\delta_0 + \pi F,$$

where δ_0 is the delta distribution that places all its probability mass at zero. If X is a random variable with this zero-inflated distribution, then with probability π, X is sampled from F, and with probability $1-\pi$, X is sampled from δ_0 and $X = 0$. Note that this is not saying that $\mathbb{P}(X = 0) = 1 - \pi$ as there is a chance that the distribution F could also set X to zero.

Definition 11 (Zero-inflated stochastic block model)**.** Let $\mathcal{Z} = \{1, \ldots, K\}$ be a sample space with probability distribution F, and let G_k, $k \in \mathcal{Z}$ be a collection of probability distributions on the interval $(0, 1]$. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is distributed as a *weighted zero-inflated stochastic block model*, if $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} F$ and, for $i < j$,

$$\mathbf{A}_{ij} \mid w_i, w_j, Z_i, Z_j \overset{\text{ind}}{\sim} (1 - w_i w_j)\delta_0 + w_i w_j H(Z_i, Z_j),$$

where $w_i \mid Z_i \overset{\text{ind}}{\sim} G_{Z_i}$ is a node-specific weight parameter.

Given $Z \sim F$, we define $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ where $\pi_k = \mathbb{P}(Z = k)$ is the probability a node is assigned to community k.

Example 15 (Zero-inflated Bernoulli distribution stochastic block model)**.** In this example, we show that the zero-inflated stochastic block model is an extension of the standard degree-corrected stochastic block model. Consider a zero-inflated stochastic block model where matrix entries are Bernoulli

random variables with the symmetric family of distributions $\{H(z_1, z_2) : z_1, z_2 \in \mathcal{Z}\}$, where

$$H(z_1, z_2) = \text{Bernoulli}(\mathbf{B}_{z_1 z_2}).$$

In this case, the probability of an edge between nodes i and j in the zero-inflated stochastic block model is given by

$$\mathbb{P}(\mathbf{A}_{ij} = 1 \mid Z_i, Z_j) = w_i w_j \mathbf{B}_{Z_i Z_j}.$$

This is the probability of an edge in the degree-corrected stochastic block model with the same weights meaning the two models are identical. ◁

Example 16 (Two-community zero-inflated Poisson distribution stochastic block model)**.** Consider a zero-inflated stochastic block model with two communities where matrix entries are Poisson random variables. The symmetric family of distributions $\{H(z_1, z_2) : z_1, z_2 \in \mathcal{Z}\}$ is given by

$$H(z_1, z_2) = \begin{cases} \text{Poisson}(\lambda_1) & \text{if } z_1 = z_2, \\ \text{Poisson}(\lambda_2) & \text{if } z_1 \neq z_2, \end{cases}$$

where $\lambda_1 \neq \lambda_2$.

We generate a zero-inflated stochastic block model with $n = 1000$ nodes, $\lambda_1 = 0.5$, $\lambda_2 = 1.0$ and a community assignment distribution F such that $\pi_1 = 1/2$ and $\pi_2 = 1/2$. Each node is independently assigned a weight parameter $w_i \overset{\text{ind}}{\sim} \text{Beta}(2, 2)$. Figure 4.3 shows the embedding $\hat{\mathbf{X}}$ of the zero-inflated stochastic block model in $d = 2$ dimensions. The points form two rays starting from the origin based on their community assignment Z_i. ◁

4.2 Weighted generalised random dot product graph

First, we make an assumption about the means of the symmetric family of distributions $\{H(z_1, z_2) : z_1, z_2 \in \mathcal{Z}\}$, akin to the form of the Bernoulli

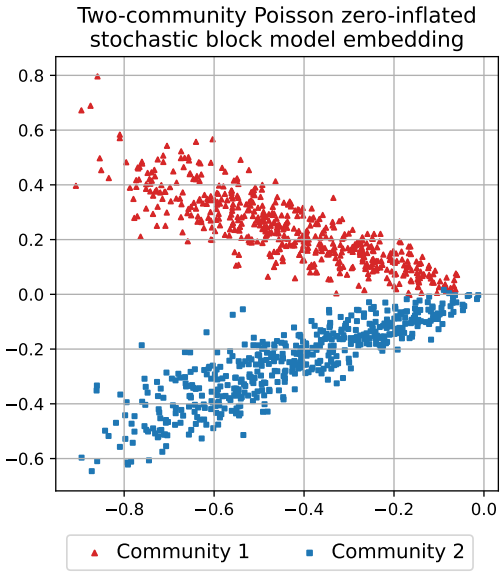

Figure 4.3: The embedding $\hat{\mathbf{X}}$ of a two-community weighted stochastic block model into $d = 2$ dimensions. The points are coloured according to their true community assignment Z_i.

probability given in Definition 6 for the generalised random dot product graph.

Assumption 2 (Low-rank expectation)**.** Given a symmetric family of real-valued distributions $\{H(z_1, z_2) : z_1, z_2 \in \mathcal{Z}\}$, there exists a map $\phi : \mathcal{Z} \to \mathbb{R}^d$ such that, for all $z_1, z_2 \in \mathcal{Z}$, if $X \sim H(z_1, z_2)$, then

$$\mathbb{E}(X) = \phi(z_1)^\top \mathbf{I}_{p,q} \phi(z_2).$$

This low-rank assumption is not a major constraint on our graph model as, from a practical view point, many real-world networks are observed to have low rank [71]. A generalised version of Mercer's theorem [48] shows that it is always possible to at least construct a possibly infinite-dimensional map ϕ. In the case where ϕ is infinite-dimensional, it can be shown for unweighted

networks that the latent positions $\phi(Z)$ often live in a low-dimensional manifold in infinite-dimensional space [59]. For weighted networks different representations of the edge weights (introduced in Section 5.2) can produce embeddings into a different number of dimensions, potentially converting a network into one where Assumption 2 is satisfied.

Note that the map ϕ is not unique. Given a valid map, it is possible to create a new map $\phi'(z) = \mathbf{Q}\phi(z)$ for $\mathbf{Q} \in \mathbb{O}(p, q)$ such that Assumption 2 remains satisfied.

Definition 12 (Weighted generalised random dot product graph)**.** Let $\mathcal{Z}$ be a sample space with probability distribution F and $\{H(z_1, z_2) : z_1, z_2 \in \mathcal{Z}\}$ be a symmetric family of real-valued distributions satisfying Assumption 2. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is distributed as a *weighted generalised random dot product graph* with signature (p, q), if $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} F$ and, for $i < j$,

$$\mathbf{A}_{ij} \mid Z_i, Z_j \overset{\text{ind}}{\sim} H(Z_i, Z_j).$$

The map ϕ in Assumption 2 is key to understanding the adjacency spectral embedding of a weighted generalised random dot product graph. In Section 4.3, we show that the adjacency spectral embedding $\hat{\mathbf{X}}_i$ only allows inference about the value $\phi(Z_i)$ rather than the latent position Z_i directly.

The three versions of the weighted stochastic block model defined in Section 4.1 are examples of the weighted generalised random dot product graph. As with the generalised random dot product graph, each example uses the adjacency spectral embedding of the block mean matrix $\mathbf{B} \in \mathbb{R}^{K \times K}$. If $\mathbf{B}$ has signature (p, q), then the embedding dimension is given by the rank of the block mean matrix $\text{rank}(\mathbf{B}) = p + q = d$.

4.2.1 Weighted stochastic block model

Lemma 4. The weighted stochastic block model is an instance of the weighted generalised random dot product graph.

Proof. Since the conditional distribution of $\mathbf{A}_{ij}$ under the weighted stochastic block model is already of the right form, all that is required is to show that Assumption 2 is satisfied. Define the map $\phi : \mathcal{Z} \to \mathbb{R}^d$ using the rows of the adjacency spectral embedding $\mathbf{X_B}$,

$$\phi(Z_i) = (\mathbf{X_B})_{Z_i}^\top.$$

Using the signature (p, q) of the block mean matrix $\mathbf{B}$,

$$\begin{aligned}\phi(Z_i)^\top \mathbf{I}_{p,q} \phi(Z_j) &= (\mathbf{X_B})_{Z_i} \mathbf{I}_{p,q} (\mathbf{X_B})_{Z_j}^\top \\ &= \mathbf{B}_{Z_i Z_j} \\ &= \mathbb{E}(\mathbf{A}_{ij} \mid Z_i, Z_j).\end{aligned}$$

Therefore, Assumption 2 is satisfied. □

Note that, unlike for the standard stochastic block model, it is possible for two distinct communities to have the same latent positions, $\phi(Z_i) = \phi(Z_j)$ for $Z_i \neq Z_j$. Two communities can have the same rows in the block mean matrix $\mathbf{B}$ but have different distribution, something we observe in Example 14.

4.2.2 Weighted mixed membership stochastic block model

Lemma 5. The weighted mixed membership stochastic block model is an instance of the weighted generalised random dot product graph.

Proof. This proof finds a new sample space $\mathcal{Z}'$ and a family of real-valued distributions $\{H'(z_1, z_2) : z_1, z_2 \in \mathcal{Z}\}$ satisfying the conditions in Definition 12. Let $\mathcal{Z}'$ be the standard $(K-1)$-dimensional simplex,

$$\mathcal{Z}' = \left\{ z = (z_1, \ldots, z_K) \in \mathbb{R}^K : \sum_{k=1}^{K} z_k = 1; z_k \geq 0 \right\},$$

and let $\{H'(z_1, z_2) : z_1, z_2 \in \mathcal{Z}'\}$ be the symmetric family of mixture distributions using the distributions from the weighted mixed membership stochastic block model,

$$H'(z_1, z_2) = \sum_{k,\ell=1}^{K} z_{1k} z_{2\ell} H(k, \ell).$$

This converts the weighted mixed membership stochastic block model into the form of the weighted generalised random dot product graph. Finally, to show Assumption 2 is satisfied, define the map $\phi : \mathcal{Z}' \to \mathbb{R}^d$,

$$\phi(Z_i') = \sum_{k=1}^{K} z_{ik} (\mathbf{X_B})_k^\top.$$

Using the adjacency spectral embedding and signature (p, q) of the block mean matrix $\mathbf{B}$,

$$\begin{aligned}
\phi(Z_i')^\top \mathbf{I}_{p,q} \phi(Z_j') &= \left\{ \sum_{k=1}^{K} z_{ik} \left(\mathbf{X_B}\right)_k^\top \right\}^\top \mathbf{I}_{p,q} \left\{ \sum_{k=1}^{K} z_{jk} \left(\mathbf{X_B}\right)_k^\top \right\} \\
&= \sum_{k,\ell=1}^{K} z_{ik} z_{j\ell} \mathbf{B}_{k\ell} \\
&= \mathbb{E}(\mathbf{A}_{ij} \mid Z_i', Z_j'). \qquad \square
\end{aligned}$$

This proof differs from the corresponding proof for the standard mixed membership stochastic block model by constructing a new sample space $\mathcal{Z}'$ and symmetric family of distribution $\{H'(z_1, z_2) : z_1, z_2 \in \mathcal{Z}'\}$ for the weighted generalised random dot product graph. Naturally, this approach can also be used to prove Lemma 2, but for Bernoulli distributions it is easy to show the edge probabilities are identical under both models. By constructing the necessary sample space $\mathcal{Z}'$ for the weighted generalised random dot product graph, we explicitly describe the simplex of latent positions we see in the adjacency spectral embedding.

4.2.3 Zero-inflated stochastic block model

Lemma 6. The zero-inflated stochastic block model is an instance of the weighted generalised random dot product graph.

Proof. Define a new latent space $\mathcal{Z}' = (0, 1] \times \mathcal{Z}$ where $z_i' = (w_i, z_i)$, with probability distribution F' constructed using the distribution F for the original latent space $\mathcal{Z}$ and the distributions G_k for the weight parameters in each community. Let $\{H'(z_1', z_2') : z_1', z_2' \in \mathcal{Z}'\}$ be the symmetric family of real-valued distributions such that

$$H'(z_1', z_2') = (1 - w_1 w_2)\delta_0 + w_1 w_2 H(z_1, z_2).$$

To show Assumption 2 is satisfied, define the map $\phi : \mathcal{Z}' \to \mathbb{R}^d$,

$$\phi(Z_i') = w_i(\mathbf{X_B})_{Z_i}^\top.$$

Using the signature (p, q) of the block mean matrix $\mathbf{B}$,

$$\begin{aligned}\phi(Z_i')^\top \mathbf{I}_{p,q} \phi(Z_j') &= \left\{ w_i \left(\mathbf{X_B}\right)_{Z_i}^\top \right\}^\top \mathbf{I}_{p,q} \left\{ w_j \left(\mathbf{X_B}\right)_{Z_j}^\top \right\} \\ &= w_i w_j \mathbf{B}_{k\ell} \\ &= \mathbb{E}(\mathbf{A}_{ij} \mid Z_i', Z_j').\end{aligned}$$

□

As with the weighted mixed membership stochastic block model, by constructing the sample space $\mathcal{Z}'$ we explicitly describe the rays of latent positions we see in the adjacency spectral embedding.

4.3 Asymptotic results

This section describes how the adjacency spectral embedding for each node $\hat{X}_i$ is connected to the latent position of the generalised random dot product

graph Z_i. As hinted to previously, the embedding $\hat{X}_i$ can only hope to estimate $\phi(Z_i)$ rather than the latent position Z_i directly, and this statement is formalised here. Once again, the adjacency spectral embedding $\hat{X}_i$ this time can only estimate the latent position $\phi(Z_i)$ up to some unknown transformation by $\mathbf{Q} \in \mathbb{O}(p, q)$, which needs to be incorporated into the asymptotic theory. As with the generalised random dot product graph, there are two types of statistical results here. The first is that the adjacency spectral embedding is uniform consistent meaning that the largest difference between $\phi(Z_i)$ and its estimate $\mathbf{Q}\hat{X}_i$ for all nodes tends to zero with high probability. The second is a central limit theorem that shows this convergence is with an asymptotically Gaussian error.

In order to obtain these convergence results, we need to make an assumption that the weighted generalised random dot product graph is sensibly defined.

Assumption 3 (Minimal dimensionality)**.** For $Z \sim F$, the second moment matrix $\mathbf{\Delta} = \mathbb{E}\left[\phi(Z)\phi(Z)^\top\right]$ has full rank d.

This is playing the same role as Assumption 1 for the generalised random dot product graph. Rather than assuming that the latent positions Z_i are sensibly chosen, instead the responsibility is being placed on the choice of map $\phi : \mathcal{Z} \to \mathbb{R}^d$. Assumption 3 ensures that the dimension d for the map has been defined economically. If $\mathbf{\Delta}$ is not full rank d, then it is possible to find a new map $\phi' : \mathcal{Z} \to \mathbb{R}^{d'}$ also satisfying Assumption 2 with smaller dimension $d' < d$ such that the resulting second moment matrix $\mathbf{\Delta'} \in \mathbb{R}^{d' \times d'}$ is full rank.

The following two assumptions are related to the distributions in the symmetric family of distributions $\{H(z_1, z_2) : z_1, z_2 \in \mathcal{Z}\}$. This ensures the edge weights in the network are suitably well-behaved to allow for convergence.

Assumption 4 (Bounded expectation)**.** There exists universal constants $a, b \in \mathbb{R}$ such that, for all $z_1, z_2 \in \mathcal{Z}$, if $X \sim H(z_1, z_2)$ then

$$\mathbb{E}(X) = \phi(z_1)^\top \mathbf{I}_{p,q} \phi(z_2) \in [a, b].$$

Assumption 5 (Exponential tails)**.** There exists universal constants $\alpha > 0$ and $\beta_\rho > 0$ for each $\rho \in \mathbb{R}$, such that, for all $z_1, z_2 \in \mathcal{Z}$, if $X \sim H(z_1, z_2)$ then

$$\mathbb{P}(|X| \geq \beta_\rho \log^\alpha t) \leq t^{-\rho}.$$

Of the two, Assumption 5 is the more restrictive insisting that the tails of the distributions are not too large. This trivially holds for bounded distributions like the Bernoulli and beta distribution. Many other commonly used distributions satisfy this condition such as the Gaussian and Poisson distributions. For example, for the exponential distribution $X \sim \text{Exp}(\lambda)$, $\mathbb{P}(X \geq x) = e^{-\lambda x}$ for $x > 0$. Therefore,

$$\mathbb{P}(X \geq \frac{\rho}{\lambda} \log t) = t^{-\rho},$$

meaning the exponential distribution with parameter λ satisfies Assumption 5 with $\alpha = 1$ and $\beta_\rho = \rho/\lambda$.

The following two theorems were first stated and proved in my paper Gallagher et al. [21]. We provide a sketch proof highlighting the main ideas of the results.

Theorem 3 (Weighted adjacency spectral embedding uniform consistency)**.** Let $\mathbf{A}$ be an instance of a weighted generalised random dot product graph. There exists a sequence of matrices $\mathbf{Q}_n \in \mathbb{O}(p, q)$ such that

$$\max_{i \in \{1,\ldots,n\}} \left\| \mathbf{Q}_n \hat{X}_i - \phi(Z_i) \right\| \xrightarrow{\mathbb{P}} 0.$$

Theorem 4 (Weighted adjacency spectral embedding central limit theorem)**.** Let $\mathbf{A}$ be an instance of a weighted generalised random dot product graph. Given $z \in \mathcal{Z}$, define the covariance-valued function

$$\mathbf{\Sigma}(z) = \mathbf{I}_{p,q}\mathbf{\Delta}^{-1}\mathbb{E}\left[v(z,Z)\phi(Z)\phi(Z)^{\top}\right]\mathbf{\Delta}^{-1}\mathbf{I}_{p,q},$$

where $v(z_1, z_2)$ is the variance of the distribution $H(z_1, z_2)$ and $Z \sim F$. Then, there exists a sequence of matrices $\mathbf{Q}_n \in \mathbb{O}(p, q)$ such that, for all $\mathbf{x} \in \mathbb{R}^d$,

$$\mathbb{P}\left(n^{1/2}\left(\mathbf{Q}_n\hat{X}_i - \phi(Z_i)\right) \leq \mathbf{x} \mid Z_i = z\right) \to \Phi\left(\mathbf{x}, \mathbf{\Sigma}(z)\right).$$

The proofs also provide the convergence rate for the uniform consistency, which are also a function of the constant α in Assumption 5 [21]. These two results are very similar in nature to Theorem 1 and 2 for the generalised random dot product graph. In the weighted versions of these theorems, the quantity $\phi(Z_i)$ is now playing the role of X_i. The adjacency spectral embedding $\hat{X}_i$ estimates $\phi(Z_i)$ rather than the latent position Z_i directly, and only up to an indefinite orthogonal transformation $\mathbf{Q}_n$. In the unweighted case, the latent positions X_i in the generalised random dot product graph are purposely chosen so there is no need to include a map $\phi : \mathcal{X} \to \mathbb{R}^d$.

To show that the asymptotic theorems for the generalised random dot product graph are a special case of these given in Theorem 3 and 4, we need to show that all the necessary assumptions hold for the symmetric family of distributions $\{H(x_1, x_2) : x_1, x_2 \in \mathcal{X}\}$,

$$H(x_1, x_2) = \text{Bernoulli}(\mathbf{B}_{x_1 x_2}).$$

Assumption 2 about the low-rank expectation is built into the generalised random dot product graph as the mean of a Bernoulli distribution for $\mathbf{A}_{ij}$ was explicitly written as $X_i^{\top}\mathbf{I}_{p,q}X_j$. This also explains why there is no map ϕ present in those results; it is equal to the identity map and we can equate X_i

and $\phi(Z_i)$ in these equations. Knowing that ϕ is the identity map, Assumption 3 about minimal dimensionality is precisely the same as Assumption 1 required for the unweighted results. Assumptions 4 and 5 trivially hold as a Bernoulli random variable is bounded by $[0, 1]$.

Finally, given the symmetric family of Bernoulli distributions and the low-rank expectation assumption, the variance function for $x_1, x_2 \in \mathcal{X}$ is given by

$$v(x_1, x_2) = \left(x_1^\top \mathbf{I}_{p,q} x_2\right)\left(1 - x_1^\top \mathbf{I}_{p,q} x_2\right).$$

Therefore, the variance term in Theorem 2 is equal to the weighted version in Theorem 4.

The proofs of Theorem 3 and 4 are largely identical to those for the generalised random dot product model but we include a sketch proof to highlight the key areas where extra work is required.

Sketch proof. For a fixed n, let $\mathbf{P} \in [a, b]^{n \times n}$ be the matrix of expected edge weights in the network using Assumption 4 to bound the entries of this matrix. By Assumption 3 the entries can be written as

$$\mathbf{P}_{ij} = \mathbb{E}(\mathbf{A}_{ij}) = \phi(Z_i)^\top \mathbf{I}_{p,q} \phi(Z_j),$$

and the weighted adjacency matrix $\mathbf{A}$ can be thought of as a noisy observation of the matrix $\mathbf{P}$.

The matrix $\mathbf{Q}_n$ is then constructed in exactly the same way as before. An orthogonal Procrustes problem is solved to align the adjacency spectral embedding $\hat{\mathbf{X}}$ and $\mathbf{X_P}$, so $\hat{\mathbf{X}}\mathbf{W}^\top \approx \mathbf{X_P}$ for some $\mathbf{W} \in \mathbb{O}(p, q)$. The embedding $\mathbf{X_P}$ is then aligned to the matrix consisting of the latent positions $\phi(Z_i)$,

$$\mathbf{X} = \left(\phi(Z_1) \mid \cdots \mid \phi(Z_n)\right)^\top \in \mathbb{R}^{n \times d},$$

so that $\mathbf{X_P} = \mathbf{X}\mathbf{Q_X}$ for some $\mathbf{Q_X} \in \mathbb{O}(p, q)$. This results in the expression $\hat{\mathbf{X}}\mathbf{W}^\top \mathbf{Q}_\mathbf{X}^{-1} \approx \mathbf{X}$ when $\mathbf{Q}_n = \mathbf{W}^\top \mathbf{Q}_\mathbf{X}^{-1}$ in Theorem 3 and 4.

Again the main part of the proof resolves around the matrix perturbation

$$n^{1/2}(\hat{\mathbf{X}}\mathbf{Q}_n - \mathbf{X}) = n^{1/2}(\mathbf{A} - \mathbf{P})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{I}_{p,q} + n^{1/2}\mathbf{R}\mathbf{Q}_n,$$

for some residual matrix $\mathbf{R} \in \mathbb{R}^{n\times d}$. This time controlling the behaviour of the right hand side is more involved as the difference $\mathbf{A} - \mathbf{P}$ now depends on the edge weight distributions, which is where Assumption 4 and 5 are required. More details can be found in the Appendix of Gallagher et al. [21]. □

4.3.1 Weighted stochastic block model asymptotic distribution

To show that the weighted stochastic block model is an instance of the weighted generalised random dot product graph in Lemma 4, we considered the latent position map

$$\phi(Z_i) = (\mathbf{X_B})^\top_{Z_i}.$$

Due to the similarity of these positions to the unweighted stochastic block model much of the asymptotic distribution calculations remain unchanged. The function $\mathbf{\Sigma}(z)$ needs to be evaluated for the K latent positions in $\mathcal{Z}$, $\mathbf{\Sigma}_k = \mathbf{\Sigma}(\phi(k))$. For $Z \sim F$, the second moment matrix is given by

$$\begin{aligned}\mathbf{\Delta} &= \mathbb{E}(\phi(Z)\phi(Z)^\top) \\ &= \sum_{k=1}^{K} \pi_k \left(\mathbf{X_B}\right)^\top_k \left(\mathbf{X_B}\right)_k \\ &= \mathbf{X}_\mathbf{B}^\top \mathbf{\Pi} \mathbf{X}_\mathbf{B},\end{aligned}$$

where $\mathbf{\Pi} = \text{diag}(\boldsymbol{\pi}) \in \mathbb{R}^{K\times K}$ is the diagonal matrix consisting of the community assignment probabilities. The covariance for community k is given

by

$$\begin{aligned}\boldsymbol{\Sigma}_k &= \mathbf{I}_{p,q}\boldsymbol{\Delta}^{-1}\mathbb{E}\left[v(k,Z)\phi(Z)\phi(Z)^\top\right]\boldsymbol{\Delta}^{-1}\mathbf{I}_{p,q} \\ &= \mathbf{I}_{p,q}\boldsymbol{\Delta}^{-1}\left\{\sum_{\ell=1}^{K}\pi_\ell \mathbf{C}_{k\ell}\left(\mathbf{X}_\mathbf{B}\right)_\ell^\top\left(\mathbf{X}_\mathbf{B}\right)_\ell\right\}\boldsymbol{\Delta}^{-1}\mathbf{I}_{p,q} \\ &= \mathbf{I}_{p,q}\boldsymbol{\Delta}^{-1}\mathbf{X}_\mathbf{B}^\top\boldsymbol{\Pi}\boldsymbol{\Lambda}_k\mathbf{X}_\mathbf{B}\boldsymbol{\Delta}^{-1}\mathbf{I}_{p,q},\end{aligned}$$

where $\boldsymbol{\Lambda}_k = \operatorname{diag}((\mathbf{C})_k) \in \mathbb{R}^{K\times K}$ is the diagonal matrix consisting of the row of the block variance matrix corresponding to community k.

Example 17 (Two-community Poisson distribution stochastic block model continued)**.** Figure 4.4 shows the asymptotic distribution for the embedding $\hat{\mathbf{X}}$ of the Poisson distribution stochastic block model from Example 13. $\mathbf{Q}_n$ can be constructed since we know the matrix of expected values $\mathbf{P}$ for the distributions used to generate the adjacency matrix $\mathbf{A}$. The asymptotic distribution is constructed using the equations above then transformed using $\mathbf{Q}_n$ from Theorem 3 and 4 to get the asymptotic distribution for the adjacency spectral embedding $\hat{\mathbf{X}}$ at those positions. Points are coloured according to their community assignment Z_i. The ellipses show the 95% contours of the asymptotic Gaussian components. ◁

Example 18 (Two-community Gaussian distribution stochastic block model with equal means continued)**.** We compute the asymptotic distribution for the embedding $\hat{\mathbf{X}}$ of the Gaussian distribution stochastic block model from Example 14. Computing the adjacency spectral embedding of the block mean matrix $\mathbf{B}$ gives $\phi(Z_i) = 1$ for both $Z_i = 1$ and $Z_i = 2$. The asymptotic distributions differ in the variance terms; $\boldsymbol{\Sigma}_1 = 1.5$ and $\boldsymbol{\Sigma}_2 = 1.0$, which match the observed variances when normalised by $n = 1000$. ◁

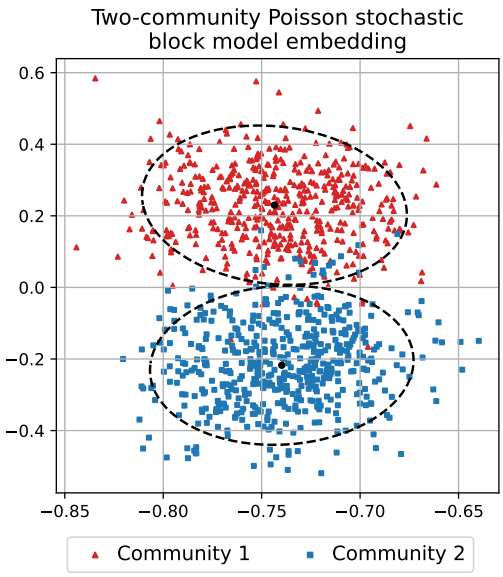

Figure 4.4: The embedding $\hat{\mathbf{X}}$ of a two-community weighted stochastic block model into $d = 2$ dimensions. The points are coloured according to their true community assignment Z_i. The ellipses show the 95% contours of the asymptotic Gaussian components.

4.3.2 Weighted mixed membership and zero-inflated stochastic block model asymptotic distribution

Section 4.3.1 showed how the asymptotic distribution calculation changed when going from the generalised stochastic block model to the weighted version. Similarly, the calculations for the mixed membership and degree-corrected stochastic block model in Sections 3.3.2 and 3.3.3 do not change meaningfully when converted to their weighted versions so we conclude with an example.

Example 19 (Two-community zero-inflated Poisson distribution stochastic block model continued)**.** Figure 4.5 shows the embedding $\hat{\mathbf{X}}$ of the zero-

inflated stochastic block model from Example 16. Points are coloured according to their community assignment Z_i. The solid black lines represent the rays of latent positions $\mathcal{X}$ after transformation using $\mathbf{Q}_n$ and the ellipses show the 95% contours of the asymptotic Gaussian components for latent positions corresponding to $w = 0.5$ and $w = 1$. ◁

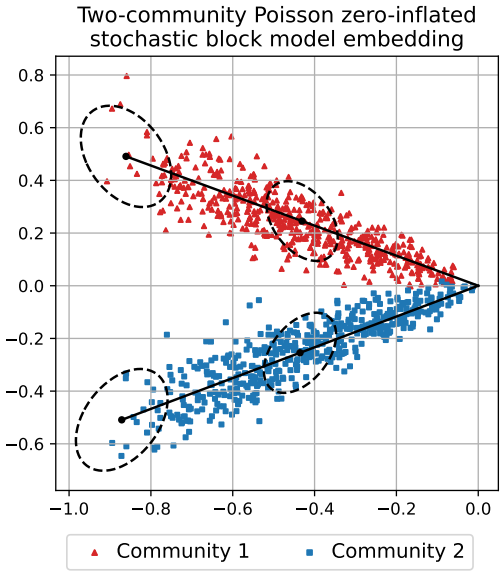

Figure 4.5: The embedding $\hat{\mathbf{X}}$ of a two-community weighted zero-inflated stochastic block model into $d = 2$ dimensions. The points are coloured according to their true community assignment Z_i. The lines show the transformed rays of $\mathcal{X}$ and the ellipses show the 95% contours of the asymptotic Gaussian components for latent positions corresponding to $w = 0.5$ and $w = 1$.

Chapter 5

Embedding comparison

In this chapter we discuss how to compare different spectral embeddings. For example, in a stochastic block model, we may wish for the embedding to create the best separation of the different communities in the network. We begin by motivating the use of Chernoff information commonly used as a method for comparing embeddings [68, 10, 57]. We show how this metric can be used to compare different embeddings of a weighted network where the entries in the adjacency matrix have been transformed by some entry-wise function and explain what benefits this may have. The latter part of this work was first published in Gallagher et al. [21].

The following example shows the danger of using standard clustering algorithms to evaluate the quality of a spectral embedding.

Example 20 (Stochastic block model community recovery)**.** Consider a stochastic block model with two communities with block mean matrix

$$\mathbf{B} = \begin{pmatrix} 0.10 & 0.12 \\ 0.12 & 0.05 \end{pmatrix},$$

and a distribution F that assigns equal probability to the two communities; $\pi_1 = 1/2$ and $\pi_2 = 1/2$. We generate a stochastic block model with $n = 1000$ nodes and perform adjacency spectral embedding into $d = 2$ dimensions.

Suppose we did not know the community assignment of the nodes within the graph. Knowing the asymptotic distribution of the embedding from Theorem 2, we fit a Gaussian mixture model using the *sklearn* machine learning package for Python with two communities and unequal, non-spherical co-

variance matrices. This correctly classifies all but two of the data points, an accuracy of 99.8%.

Alternatively, we can use another clustering algorithm such as k-means also available in the *sklearn* package. Figure 5.1a shows the embedding $\hat{\mathbf{X}}$ of the stochastic block model where points are now coloured according to their k-means cluster assignment, while the ellipses show the 95% contours of the asymptotic Gaussian components. Using k-means performs only slightly worse correctly classifying all but six of the data points, an accuracy of 99.4%, which is comparable to the Gaussian mixture model although there are a few noticeable misclassifications.

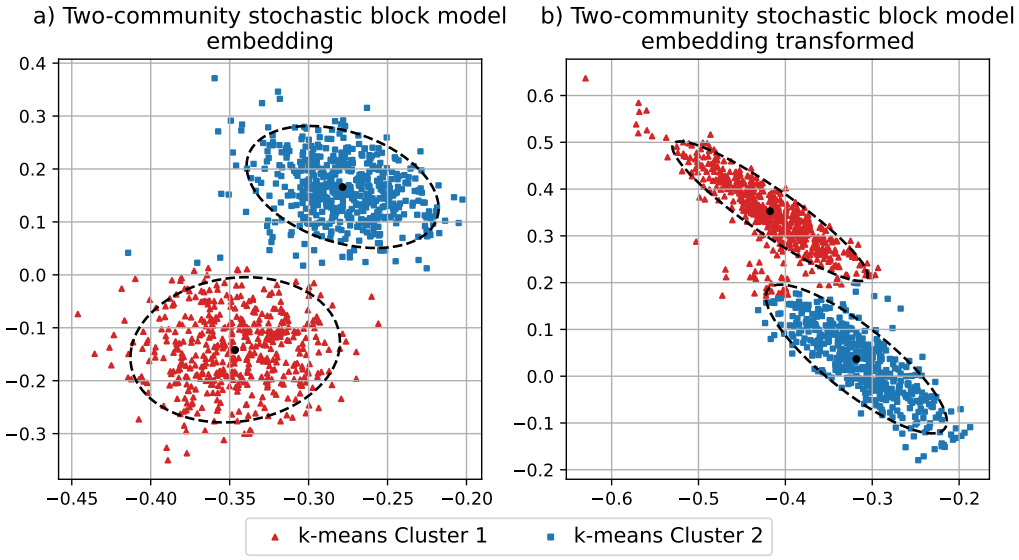

Figure 5.1: The embedding $\hat{\mathbf{X}}$ of a two-community stochastic block model into $d = 2$ dimensions, a) unaltered and b) transformed by $\mathbf{Q} \in \mathbb{O}(p, q)$. The points are coloured according to their k-means cluster assignment. The ellipses show the 95% contours of the asymptotic Gaussian components.

However, since the signature of the block mean matrix $\mathbf{B}$ is $(1, 1)$, any transformation of the embedding by $\mathbf{Q} \in \mathbb{O}(1, 1)$ is also a valid embedding

of the network. Figure 5.1b shows the embedding $\hat{\mathbf{X}}$ after transformation by one such matrix $\mathbf{Q}$ chosen to elongate the two clusters. This has no effect on fitting a Gaussian mixture model with unequal, non-spherical covariance matrices; the accuracy remains at 99.8%. However, the transformation does have an effect on the k-means output, which misclassifies 38 of the data points, an accuracy of only 96.2%. This highlights the issue that any metric chosen to evaluate the quality of an adjacency spectral embedding will need to be invariant under transformation by $\mathbf{Q} \in \mathbb{O}(p, q)$. ◁

Even in this example there is another issue with using classification accuracy as a measure of the quality of a spectral embedding. As the number of nodes tends to infinity, the community clusters will separate more due to the smaller covariance in the central limit theorems, Theorem 2 and Theorem 4. A worse representation of the data will eventually be just as good at separating points as the number of nodes increase.

5.1 Chernoff information

In this section, we introduce Chernoff information, a commonly used method for evaluating the spectral embedding of a stochastic block model [68, 10, 57].

Definition 13 (Chernoff information)**.** Let F_1 and F_2 be continuous multivariate distributions on $\mathbb{R}^d$ with density functions f_1 and f_2 respectively. The Chernoff information [14] is defined by

$$\mathcal{C}(F_1, F_2) = \sup_{t \in (0,1)} \left\{ -\log \int f_1^t(x) f_2^{1-t}(x) \, \mathrm{d}x \right\},$$

When given three or more distributions, one reports the Chernoff information of the critical pair, $\min_{k \neq \ell} \mathcal{C}(F_k, F_\ell)$.

To understand why we use Chernoff information, we follow Nielsen [52]. Consider a classification task between two classes C_1 and C_2 with density

functions f_1 and f_2 respectively where an observation x is drawn from the mixture distribution $f(x) = \mathbb{P}(C_1)f_1(x) + \mathbb{P}(C_2)f_2(x)$. The probability of that observation being from either class is given by Bayes' theorem,

$$\mathbb{P}(C_i \mid x) = \frac{\mathbb{P}(C_i)f_i(x)}{f(x)}.$$

The Bayes decision rule under a 0-1 loss function assigns the observation x to its most likely class, that is, x is assigned to class C_1 if $\mathbb{P}(C_1)f_1(x) > \mathbb{P}(C_2)f_2(x)$ and to class C_2 otherwise. The Bayes decision rule minimises the average probability of a misclassification,

$$E = \int \min\{\mathbb{P}(C_1 \mid x), \mathbb{P}(C_2 \mid x)\}f(x)\,\mathrm{d}x,$$

the probability that an observation from C_1 is assigned as C_2 and *vice versa*. Given a set of observations, this is the limit of the observed average misclassification error. To create an upper bound for the error, observe that $\min(a, b) \leq a^t b^{1-t}$ for all $t \in (0, 1)$, so

$$\begin{aligned} E &\leq \inf_{t\in(0,1)} \mathbb{P}(C_1)^t \mathbb{P}(C_2)^{1-t} \int f_1^t(x) f_2^{1-t}(x)\,\mathrm{d}x\,\mathrm{d}x \\ &= \mathbb{P}(C_1)^{t^*} \mathbb{P}(C_2)^{1-t^*} e^{-\mathcal{C}(F_1,F_2)}, \end{aligned}$$

where $t^* \in (0, 1)$ is the value that gives the smallest upper bound. In summary, the Chernoff information is the exponential rate of an upper bound for the average misclassification error.

The Chernoff information is invariant up to invertible transformations [16]. In particular, it is invariant to indefinite orthogonal transformations $\mathbf{Q} \in \mathbb{O}(p, q)$ meaning that it is suitable to use for adjacency spectral embeddings.

Example 21 (Chernoff information of Gaussian distributions)**.** Consider the multivariate Gaussian distributions $F_1 = \text{Normal}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $F_2 =$

Normal$(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ with probability density functions,

$$f_i(\mathbf{x}) = \frac{1}{\sqrt{2\pi \det(\boldsymbol{\Sigma}_i)}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i (\mathbf{x}-\boldsymbol{\mu}_i)\right\}.$$

The integral in the Chernoff information can be evaluated by rewriting it as the probability density function of another multivariate Gaussian distribution. The Chernoff information $\mathcal{C}(F_1, F_2)$ is given by [52]

$$\sup_{t\in(0,1)} \left[\frac{t(1-t)}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}(t)^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\frac{1}{2}\log\frac{\det(\boldsymbol{\Sigma}(t))}{\det(\boldsymbol{\Sigma}_1)^{1-t}\det(\boldsymbol{\Sigma}_2)^t}\right],$$

where $\boldsymbol{\Sigma}(t) = (1-t)\boldsymbol{\Sigma}_1 + t\boldsymbol{\Sigma}_2$. ◁

5.1.1 Stochastic block model Chernoff information

The larger the Chernoff information, the easier it is to distinguish the two distributions. For both the standard and weighted stochastic block model, Theorem 3 and 4 state that the asymptotic distribution for the adjacency spectral embedding for community k is Normal$((\mathbf{X_B})_k, \boldsymbol{\Sigma}_k/n)$. In Section 4.3.1 we derived the following expression for the asymptotic covariance,

$$\boldsymbol{\Sigma}_k = \mathbf{I}_{p,q}\boldsymbol{\Delta}^{-1}\mathbf{X}_\mathbf{B}^\top\boldsymbol{\Pi}\boldsymbol{\Lambda}_k\mathbf{X_B}\boldsymbol{\Delta}^{-1}\mathbf{I}_{p,q},$$

where $\boldsymbol{\Delta} = \mathbf{X}_\mathbf{B}^\top\boldsymbol{\Pi}\mathbf{X_B} \in \mathbb{R}^{d\times d}$ is the second moment matrix, $\boldsymbol{\Pi} = \operatorname{diag}(\boldsymbol{\pi}) \in \mathbb{R}^{K\times K}$ is the diagonal matrix consisting of the community assignment probabilities, and $\boldsymbol{\Lambda}_k = \operatorname{diag}(\mathbf{C}_k) \in \mathbb{R}^{K\times K}$ is the diagonal matrix consisting of the row of the block variance matrix corresponding to community k.

The covariance term is getting smaller as the number of nodes n tends to infinity making separating any pair of communities easier, so we want to know how the Chernoff information improves with n. For communities k and ℓ, the first term in the supremum of the Chernoff information from Example 21 is given by

$$\frac{nt(1-t)}{2}\left\{(\mathbf{X_B})_k - (\mathbf{X_B})_\ell\right\}^\top \boldsymbol{\Sigma}_{k\ell}(t)^{-1}\left\{(\mathbf{X_B})_k - (\mathbf{X_B})_\ell\right\},$$

where $\mathbf{\Sigma}_{k\ell}(t) = (1-t)\mathbf{\Sigma}_k + t\mathbf{\Sigma}_\ell$. This term is linear in n, while the second term

$$\frac{1}{2}\log\frac{\det(\mathbf{\Sigma}(t))}{\det(\mathbf{\Sigma}_1)^{1-t}\det(\mathbf{\Sigma}_2)^t}$$

is independent of n as contributions from the top and bottom of the fraction cancel. This motivates the following definition for the size-adjusted Chernoff information that explains how the Chernoff information in a stochastic block model grows as n tends to infinity.

Definition 14 (Size-adjusted Chernoff information)**.** The *size-adjusted Chernoff information* for the adjacency spectral embedding of a stochastic block model is

$$C = \min_{k\neq\ell}\sup_{t\in(0,1)}\left[\frac{t(1-t)}{2}\left\{(\mathbf{X_B})_k - (\mathbf{X_B})_\ell\right\}^\top \mathbf{\Sigma}_{k\ell}(t)^{-1}\left\{(\mathbf{X_B})_k - (\mathbf{X_B})_\ell\right\}\right],$$

where $\mathbf{\Sigma}_{k\ell}(t) = (1-t)\mathbf{\Sigma}_k + t\mathbf{\Sigma}_\ell$.

This definition of size-adjusted Chernoff information is related to the idea of Chernoff information ratios [56]. In this work, the authors compare the ratio of two Chernoff information terms as n tends to infinite. By focusing on the contribution of the term linear in n, the size-adjusted Chernoff information captures the most relevant part for a single embedding without having to immediately compare it against another embedding.

The following lemma shows how the expression can be simplified for the stochastic block model. This is a new contribution to be included in a revision for Gallagher et al. [21].

Lemma 7. For a K-community stochastic block model with full rank mean block matrix, $\operatorname{rank}(\mathbf{B}) = K$, the size-adjusted Chernoff information is

$$C = \min_{k\neq\ell}\sup_{t\in(0,1)}\left[\frac{t(1-t)}{2}\left\{(\mathbf{e}_k - \mathbf{e}_\ell)^\top \mathbf{B\Pi\Lambda}_{k\ell}(t)^{-1}\mathbf{B}(\mathbf{e}_k - \mathbf{e}_\ell)\right\}\right],$$

where $\mathbf{e}_k \in \mathbb{R}^K$ is the standard basis vector with one in position k, and zero elsewhere, and $\mathbf{\Lambda}_{k\ell}(t) = (1-t)\mathbf{\Lambda}_k + t\mathbf{\Lambda}_\ell$.

This is a vast simplification when $\mathbf{B}$ is full rank. Rather than having to compute the spectral embedding $\mathbf{X_B}$ and the second moment matrix $\mathbf{\Delta}$, the size-adjusted Chernoff information can be expressed directly in terms of the block mean and variance matrices $\mathbf{B}$ and $\mathbf{C}$. This makes the optimisation steps in the calculation much more manageable.

Proof. If $\mathbf{D} \in \mathbb{R}^{K\times K}$ is full rank, then for $\mathbf{Y} \in \mathbb{R}^{K\times d}$, the following matrix inequality holds [46],

$$\mathbf{Y}\left(\mathbf{Y}^\top\mathbf{D}\mathbf{Y}\right)^{-1}\mathbf{Y}^\top \preceq \mathbf{D}^{-1},$$

where $\mathbf{M} \succeq \mathbf{0}$ means that $\mathbf{M}$ is a positive semi-definite matrix. However, in the case where $\mathbf{D}$ and $\mathbf{X} \in \mathbb{R}^{K\times K}$ are full rank, then the two sides of this inequality are equal,

$$\mathbf{X}\left(\mathbf{X}^\top\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}^\top = \mathbf{D}^{-1}.$$

If the block mean matrix $\mathbf{B}$ is full rank, this implies that the adjacency spectral embedding $\mathbf{X_B}$ is also full rank. Since $\mathbf{B} = \mathbf{X_B}\mathbf{I}_{p,q}\mathbf{X_B}^\top$, this means $\operatorname{rank}(\mathbf{B}) \leq \operatorname{rank}(\mathbf{X_B})$ implying $\operatorname{rank}(\mathbf{X_B}) = K$ when $\operatorname{rank}(\mathbf{B}) = K$.

Using this matrix equality and the expression for $\mathbf{\Delta}$, we have

$$\begin{aligned}
\mathbf{\Sigma}_{k\ell}(t)^{-1} &= \left(\mathbf{I}_{p,q}\mathbf{\Delta}^{-1}\mathbf{X_B}^\top\mathbf{\Pi}\mathbf{\Lambda}_{k\ell}(t)\mathbf{X_B}\mathbf{\Delta}^{-1}\mathbf{I}_{p,q}\right)^{-1} \\
&= \mathbf{I}_{p,q}\mathbf{X_B}^\top\mathbf{\Pi}\mathbf{X_B}\left(\mathbf{X_B}^\top\mathbf{\Pi}\mathbf{\Lambda}_{k\ell}(t)\mathbf{X_B}\right)^{-1}\mathbf{X_B}^\top\mathbf{\Pi}\mathbf{X_B}\mathbf{I}_{p,q} \\
&= \mathbf{I}_{p,q}\mathbf{X_B}^\top\mathbf{\Pi}\mathbf{\Lambda}_{k\ell}(t)^{-1}\mathbf{X_B}\mathbf{I}_{p,q}
\end{aligned}$$

Substituting this expression into the objective function in the size-adjusted Chernoff information gives

$$\begin{aligned}
&\left\{(\mathbf{X_B})_k - (\mathbf{X_B})_\ell\right\}^\top \mathbf{\Sigma}_{k\ell}(t)^{-1}\left\{(\mathbf{X_B})_k - (\mathbf{X_B})_\ell\right\} \\
&= (\mathbf{e}_k - \mathbf{e}_\ell)^\top\mathbf{X_B}\mathbf{I}_{p,q}\mathbf{X_B}^\top\mathbf{\Pi}\mathbf{\Lambda}_{k\ell}(t)^{-1}\mathbf{X_B}\mathbf{I}_{p,q}\mathbf{X_B}^\top(\mathbf{e}_k - \mathbf{e}_\ell) \\
&= (\mathbf{e}_k - \mathbf{e}_\ell)^\top\mathbf{B}\mathbf{\Pi}\mathbf{\Lambda}_{k\ell}(t)^{-1}\mathbf{B}(\mathbf{e}_k - \mathbf{e}_\ell),
\end{aligned}$$

using the expressions $(\mathbf{X_B})_k - (\mathbf{X_B})_\ell = \mathbf{X_B}^\top(\mathbf{e}_k - \mathbf{e}_\ell)$ and $\mathbf{B} = \mathbf{X_B}\mathbf{I}_{p,q}\mathbf{X_B}^\top$. □

If the block mean matrix $\mathbf{B}$ is not full rank, then the matrix inequality effectively passes through the whole argument in the above proof,

$$\mathbf{\Sigma}_{k\ell}(t)^{-1} \preceq \mathbf{I}_{p,q}\mathbf{X}_{\mathbf{B}}^{\top}\mathbf{\Pi}\mathbf{\Lambda}_{k\ell}(t)^{-1}\mathbf{X}_{\mathbf{B}}\mathbf{I}_{p,q},$$

which becomes a regular inequality when we compute the quadratic form. We get that, for $\text{rank}(\mathbf{B}) \leq K$,

$$C \leq \min_{k\neq\ell} \sup_{t\in(0,1)} \left[\frac{t(1-t)}{2}\left\{(\mathbf{e}_k - \mathbf{e}_\ell)^{\top}\mathbf{B}\mathbf{\Pi}\mathbf{\Lambda}_{k\ell}(t)^{-1}\mathbf{B}(\mathbf{e}_k - \mathbf{e}_\ell)\right\}\right].$$

with equality when $\mathbf{B}$ is full rank.

Example 22 (Adjacency versus Laplacian spectral embedding)**.** In this example we use size-adjusted Chernoff information to compare the adjacency spectral embedding and Laplacian spectral embedding of a two-community stochastic block model. Both approaches have their advantages [57] and depending on the model parameters one embedding may separate the two communities better than the other [68, 10].

Consider a stochastic block model with two communities with block mean matrix

$$\mathbf{B} = \begin{pmatrix} 0.20 & 0.05 \\ 0.05 & 0.05 \end{pmatrix},$$

and a distribution F that assigns equal probability to the two communities; $\pi_1 = 1/2$ and $\pi_2 = 1/2$. This is called a *core-periphery network* where the nodes in community 1 form the core of the network and the nodes in community 2 the periphery.

We generate a stochastic block model with $n = 1000$ nodes and Figure 5.2 shows the adjacency spectral embedding and Laplacian spectral embedding both into $d = 2$ dimensions. Points are coloured according to their community assignment Z_i. The ellipses show the 95% contours of the asymptotic Gaussian components using the asymptotic results by Rubin-Delanchy et al. [60] for Laplacian spectral embedding.

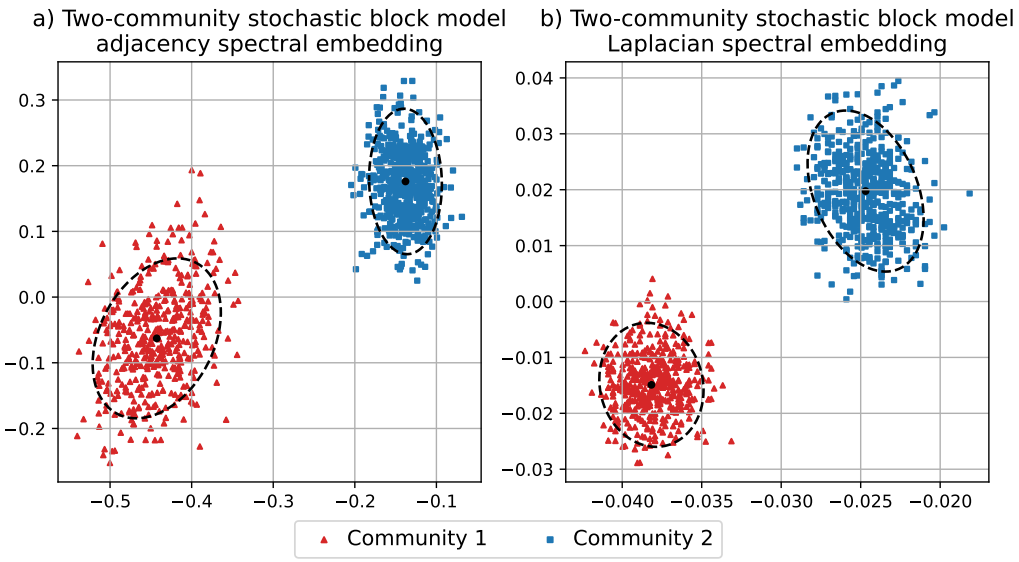

Figure 5.2: The embedding $\hat{\mathbf{X}}$ of a two-community weighted stochastic block model into $d = 2$ dimensions for a) adjacency spectral embedding and b) Laplacian spectral embedding. The points are coloured according to their true community assignment Z_i. The ellipses show the 95% contours of the asymptotic Gaussian components.

In this example Laplacian spectral embedding performs slightly better with size-adjusted Chernoff information $C = 1.63 \times 10^{-2}$ compared to adjacency spectral embedding with size-adjusted Chernoff information $C = 1.47 \times 10^{-2}$. A deeper analysis showing when the adjacency spectral embedding or Laplacian spectral embedding is preferred for core-periphery networks is perform in Tang and Priebe [68]. ◁

In the following two examples we consider two types of stochastic block models with extra structure where it is possible to calculate the size-adjusted Chernoff information analytically. These calculations will be useful when we consider models motivated by real-world networks in Example 26 and Section 5.3.

Example 23 (Homogeneous weighted stochastic block model)**.** Consider a K-community weighted stochastic block model where matrix entries are distributed according to the symmetric family of distributions $\{H(z_1, z_2) : z_1, z_2 \in \mathcal{Z}\}$ given by

$$H(z_1, z_2) = \begin{cases} F_1 & \text{if } z_1 = z_2, \\ F_2 & \text{if } z_1 \neq z_2. \end{cases}$$

An edge weight is distributed as F_1 if both endpoints are in the same community, and as F_2 if the endpoints are in different communities. Further, assume that the community assignment distribution F is such that a node is equally likely to be assigned to any of the K communities, $\pi_k = 1/K$. This is known as a homogeneous stochastic block model [35], although it is also known as a symmetric stochastic block model in other circles [1]. All the nodes behave in the same way regarding links between and within community, one such example being the two-community Poisson distribution stochastic block model in Example 13.

Since every community behaves the same way, we only need to consider one pair of communities for the size-adjusted Chernoff information; without loss of generality, let $k = 1, \ell = 2$. Denote the entries in the block mean and variance matrices $\mathbf{B}$ and $\mathbf{C}$ as, for $X \sim F_i$, $b_i = \mathbb{E}(X)$ and $c_i = \text{var}(X)$, leading to the expressions

$$\begin{aligned} \mathbf{B}(\mathbf{e}_1 - \mathbf{e}_2) &= (b_1 - b_2, b_2 - b_1, 0, \ldots, 0)^\top, \\ \mathbf{\Lambda}_{12}(t) &= \text{diag}[(1-t)c_1 + tc_2, tc_1 + (1-t)c_2, 1, \ldots, 1]. \end{aligned}$$

Substituting these into the formula for the size-adjusted Chernoff information, after some matrix manipulation, gives

$$C = \sup_{t \in (0,1)} \left[\frac{(b_1 - b_2)^2}{2K} \left\{ \frac{t(1-t)}{(1-t)c_1 + tc_2} + \frac{t(1-t)}{tc_1 + (1-t)c_2} \right\} \right].$$

It can be shown that for $c_1, c_2 > 0$, for all $t \in (0, 1)$,

$$\frac{t(1-t)}{(1-t)c_1 + tc_2} + \frac{t(1-t)}{tc_1 + (1-t)c_2} \leq \frac{1}{c_1 + c_2},$$

where equality is achieved when $t = 1/2$. Therefore, the size-adjusted Chernoff information is given by

$$C = \frac{(b_1 - b_2)^2}{2K(c_1 + c_2)}.$$

In the Poisson distribution stochastic block model in Example 13 where $b_1 = c_1 = 0.5$ and $b_2 = c_2 = 0.6$, the size-adjusted Chernoff information is equal to $C = 2.27 \times 10^{-3}$. ◁

Example 24 (Anomalous two-community weighted stochastic block model)**.** Consider a two-community weighted stochastic block model where matrix entries are distributed according to the symmetric family of distributions $\{H(z_1, z_2) : z_1, z_2 \in \mathcal{Z}\}$ given by

$$H(k, \ell) = \begin{cases} F_1 & \text{if } k = 1, \ell = 1, \\ F_2 & \text{otherwise.} \end{cases}$$

An edge weight is distributed as F_1 if both endpoints are in the community 1, and F_2 otherwise. We think of community 1 as an anomalous community in the network. Let π_1 be the probability a node is an anomaly corresponding to community 1, and π_2 the probability it is behaving normally corresponding to community 2.

This time there are only two communities to consider in the Chernoff information calculation, so let $k = 1$ and $\ell = 2$. Denote the entries in the block mean and variance matrices $\mathbf{B}$ and $\mathbf{C}$ as, for $X \sim F_i$, $b_i = \mathbb{E}(X)$ and $c_i = \text{var}(X)$, leading to the expressions

$$\begin{aligned} \mathbf{B}(\mathbf{e}_1 - \mathbf{e}_2) &= (b_1 - b_2, 0)^\top, \\ \mathbf{\Lambda}_{12}(t) &= \text{diag}[(1-t)c_1 + tc_2, c_2]. \end{aligned}$$

Substituting these into the formula for the size-adjusted Chernoff information, after some matrix manipulation, gives

$$C = \sup_{t\in(0,1)} \left[\frac{\pi_1(b_1 - b_2)^2}{2} \left\{ \frac{t(1-t)}{(1-t)c_1 + tc_2} \right\} \right].$$

In this case, we can find the optimal $t^* \in (0,1)$ by differentiating the function in the curly brackets with respect to t, although it is easier to minimise the reciprocal of this quantity rather than maximise the quantity directly. The optimum is achieved by $t^* = \frac{\sqrt{c_1}}{\sqrt{c_1}+\sqrt{c_2}}$ and the Chernoff information is given by

$$C = \frac{\pi_1(b_1 - b_2)^2}{2(\sqrt{c_1} + \sqrt{c_2})^2}. \qquad \triangleleft$$

5.2 Edge weight representation

Definition 15 (Entry-wise transformation)**.** Given a weighted adjacency matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and a real-valued function g, the *entry-wise transformation* of $\mathbf{A}$ is the weighted adjacency matrix $\mathbf{A}' \in \mathbb{R}^{n\times n}$ with entries $\mathbf{A}'_{ij} = g(\mathbf{A}_{ij})$.

An entry-wise transformation of a weighted generalised random dot product graph changes the symmetric family of distributions $\{H(z_1, z_2) : z_1, z_2 \in \mathcal{Z}\}$, but not the latent positions Z_i. Therefore, a weighted stochastic block model will remain a weighted stochastic block model after an entry-wise transformation, and likewise for the weighted mixed membership stochastic block model and zero-inflated stochastic block model if $g(0) = 0$.

When the weighted adjacency matrix $\mathbf{A}$ has block mean and variance matrices $\mathbf{B}$ and $\mathbf{C}$ respectively, we use $\mathbf{B}'$ and $\mathbf{C}'$ to represent the block mean and variance matrices of the entry-wise transformation $\mathbf{A}'$. Be aware that the matrices $\mathbf{B}'$ and $\mathbf{C}'$ are not the entry-wise transformations of $\mathbf{B}$ and $\mathbf{C}$, in general $\mathbf{B}'_{k\ell} \neq g(\mathbf{B}_{k\ell})$ and $\mathbf{C}'_{k\ell} \neq g(\mathbf{C}_{k\ell})$.

Example 25 (Two-community Gaussian distribution stochastic block model with equal means continued)**.** In Example 18 we computed the asymptotic distribution for the embedding $\hat{\mathbf{X}}$ of a Gaussian distribution stochastic block model with equal means for all pairs of communities. In this example, we showed that both communities have the same latent position, for example, $(\mathbf{X_B})_1^\top = (\mathbf{X_B})_2^\top = 1$, leading to a size-adjusted Chernoff information equal to zero. In this case, there is no longer a linear term in n in the Chernoff information so it does not increase as the number of nodes in the network tends to infinite.

By performing an entry-wise transformation of the adjacency matrix, it is possible to distinguish the two communities. We create a new stochastic block model $\mathbf{A}'$ using the entry-wise transformation $\mathbf{A}'_{ij} = \mathbf{A}^2_{ij}$ resulting in edge weights with noncentral χ^2 distributions. Figure 5.3 shows the asymptotic distribution for the resulting embedding $\hat{\mathbf{X}}$ of the transformed network into $d = 2$ dimensions. Points are coloured according to their community assignment Z_i. The ellipses show the 95% contours of the asymptotic Gaussian components. It is now possible to distinguish between the two communities with size-adjusted Chernoff information $C = 6.01 \times 10^{-3}$. ◁

Example 26 (Two-community Poisson versus Bernoulli distribution stochastic block model)**.** Recall the two-community Poisson distribution stochastic block model first introduced in Example 13 where the symmetric family of distributions $\{H(z_1, z_2) : z_1, z_2 \in \mathcal{Z}\}$ is given by

$$H(z_1, z_2) = \begin{cases} \text{Poisson}(\lambda_1) & \text{if } z_1 = z_2, \\ \text{Poisson}(\lambda_2) & \text{if } z_1 \neq z_2, \end{cases}$$

where $\lambda_1 \neq \lambda_2$ and a community assignment distribution F such that $\pi_1 = 1/2$ and $\pi_2 = 1/2$. This is an example of a homogeneous stochastic block model introduced in Example 23 and using the size-adjusted Chernoff information calculated in that example, this stochastic block model has size-

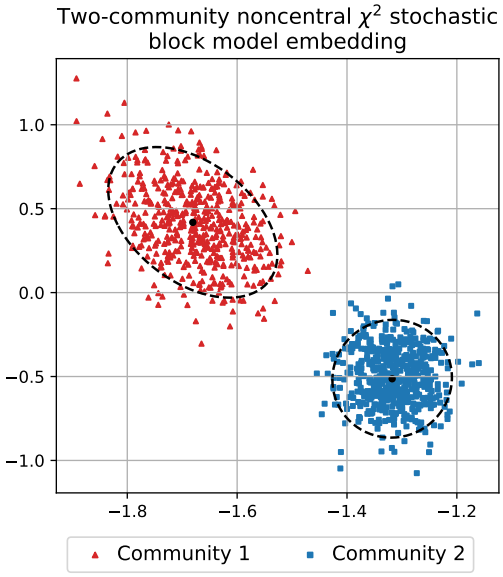

Figure 5.3: The embedding $\hat{\mathbf{X}}$ of a two-community weighted stochastic block model into $d = 2$ dimensions. The points are coloured according to their true community assignment Z_i. The ellipses show the 95% contours of the asymptotic Gaussian components.

adjusted Chernoff information given by

$$C_{\mathrm{P}} = \frac{(\lambda_1 - \lambda_2)^2}{4(\lambda_1 + \lambda_2)}.$$

However, rather than using a Poisson edge weight to model the number of events occurring between two nodes, we could instead track whether any event happened between those two nodes. We create a new stochastic block model $\mathbf{A}'$ using the entry-wise transformation $\mathbf{A}'_{ij} = \mathbb{I}(\mathbf{A}_{ij} > 0)$. If the counts are distributed by $X \sim \text{Poisson}(\lambda)$, then a presence event is distributed as $\mathbb{I}(X > 0) \sim \text{Bernoulli}(1 - e^{-\lambda})$. This creates a symmetric Bernoulli distribution stochastic block model which has size-adjusted Chernoff information

given by

$$C_{\mathrm{B}} = \frac{(e^{-\lambda_1} - e^{-\lambda_2})^2}{4\left\{e^{-\lambda_1}(1 - e^{-\lambda_1}) + e^{-\lambda_2}(1 - e^{-\lambda_2})\right\}}.$$

Depending on the values λ_1, λ_2, either the Poisson or Bernoulli representation may result in a better size-adjusted Chernoff information and, hence, a better separation of the two communities in the adjacency spectral embedding. Figure 5.4 compares the two size-adjusted Chernoff information values by showing a heat map of the ratio $C_{\mathrm{P}}/C_{\mathrm{B}}$ for $\lambda_1, \lambda_2 \in (0, 2]$. For most

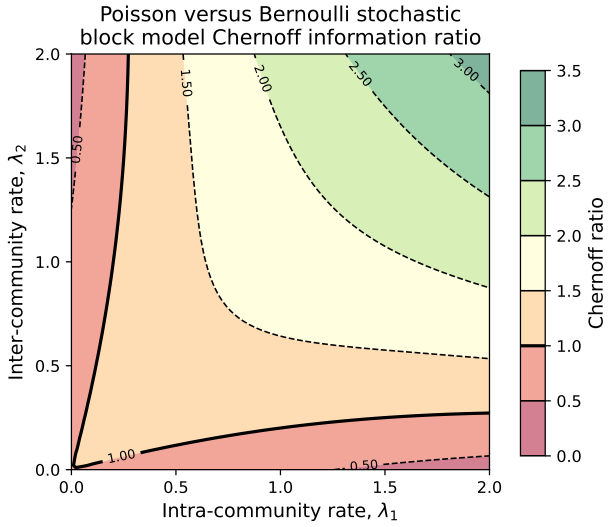

Figure 5.4: Chernoff ratio comparing the embedding of a two-community Poisson distribution stochastic block model $\mathbf{A}$ and the entry-wise transformed Bernoulli distribution stochastic block model $\mathbf{A}' = \mathbb{I}(\mathbf{A} > 0)$. The solid black line shows the boundary $C_{\mathrm{P}} = C_{\mathrm{B}}$.

pairs of Poisson distribution parameters, it is better to embed the network of count values $\mathbf{A}$ rather than just the network of presence events $\mathbf{A}'$. However, when λ_1 is small and λ_2 is relatively large (or *vice versa*), then we get better community separation by ignoring the count values in the matrix $\mathbf{A}$ losing information about the network.

Returning to Example 13 where $\lambda_1 = 0.5$ and $\lambda_2 = 0.6$, Figure 5.5 shows the embedding $\hat{\mathbf{X}}$ for the entry-wise transformed stochastic block model $\mathbf{A}'$. Points are coloured according to their community assignment Z_i. The ellipses show the 95% contours of the asymptotic Gaussian components computed using Theorem 3 and 4 to get the asymptotic distribution for the adjacency spectral embedding.

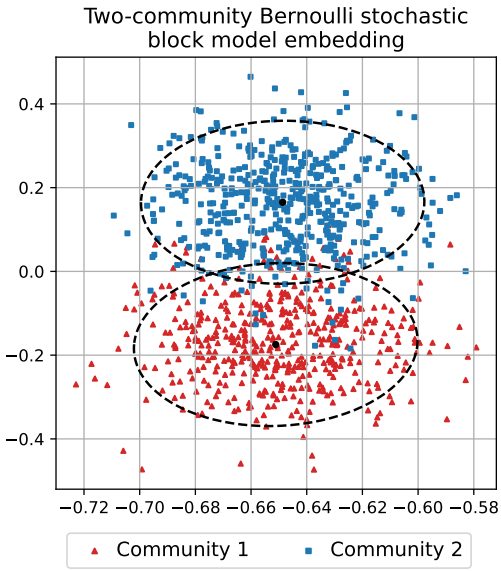

Figure 5.5: The embedding $\hat{\mathbf{X}}$ of a two-community entry-wise transformed stochastic block model into $d = 2$ dimensions. The points are coloured according to their true community assignment Z_i. The ellipses show the 95% contours of the asymptotic Gaussian components.

Compared to the embedding of Poisson distribution stochastic block model in Figure 4.4, the two communities appear more mixed together and harder to distinguish. The two ellipses showing the 95% contours of the asymptotic Gaussian components overlap more for the Bernoulli distribution stochastic block model. This is confirmed by the corresponding size-adjusted Chernoff information values; for the Bernoulli presence event embedding we have

$C_{\mathrm{B}} = 1.71 \times 10^{-3}$ compared to the Poisson counts embedding having $C_{\mathrm{P}} = 2.27 \times 10^{-3}$ computed at the end of Example 23. ◁

5.2.1 Affine transformations

The following lemma is a key result from Gallagher et al. [21] showing that an affine entry-wise transformation of a weighted adjacency matrix does not effect the size-adjusted Chernoff information.

Lemma 8. Let $\mathbf{A}$ be a weighted stochastic block model with full rank block mean matrix $\mathbf{B}$ and size-adjusted Chernoff information C. For the affine entry-wise transformation $\mathbf{A}'_{ij} = a\mathbf{A}_{ij} + b$, if the block mean matrix $\mathbf{B}'$ is full rank, then $\mathbf{A}'$ has size-adjusted Chernoff information C.

Proof. By assumption $\mathbf{A}$ and $\mathbf{A}'$ have full rank block mean matrices, therefore, by Lemma 7, both stochastic block models have the same Chernoff information if they have the same value for

$$(\mathbf{e}_k - \mathbf{e}_\ell)^\top \mathbf{B}\mathbf{\Pi}\mathbf{\Lambda}_{k\ell}(t)^{-1}\mathbf{B}(\mathbf{e}_k - \mathbf{e}_\ell).$$

For an affine entry-wise transformation, the entries of the block mean and variance matrices are given by $\mathbf{B}'_{k\ell} = a\mathbf{B}_{k\ell} + b$ and $\mathbf{C}'_{k\ell} = a^2\mathbf{C}_{k\ell}$. Therefore,

$$\begin{aligned}\mathbf{B}'(\mathbf{e}_k - \mathbf{e}_\ell) &= a\mathbf{B}(\mathbf{e}_k - \mathbf{e}_\ell),\\ \mathbf{\Lambda}'_{k\ell}(t)^{-1} &= a^2\mathbf{\Lambda}_{k\ell}(t)^{-1},\end{aligned}$$

where $\mathbf{\Lambda}'_{k\ell}(t)^{-1}$ is the equivalent version of $\mathbf{\Lambda}_{k\ell}(t)^{-1}$ in the entry-wise transformed stochastic block model. The contribution from a cancel and there is no contribution from b, meaning the size-adjusted Chernoff information is unaffected by affine transformation. □

Insisting that the block mean matrix $\mathbf{B}'$ is full rank prevents the degenerate case $a = 0$ and all the matrix entries are the same. If $a \neq 0$, then

it is highly likely that the block mean matrix $\mathbf{B}'$ is full rank if $\mathbf{B}$ is itself full rank. Without loss of generality, consider an affine transformation with $a > 0, b > 0$ on a block mean matrix $\mathbf{B}$ with signature (p, q). Analysis for other affine transformations leads to slight variations of the following results.

Suppose that the ordered eigenvalues of the block mean matrix $\mathbf{B}$ are given by $\lambda_1(\mathbf{B}) < \ldots < \lambda_K(\mathbf{B})$ where the first q eigenvalues are less than zero and the last p eigenvalues are greater than zero in order to have signature (p, q). The block mean matrix for the affine transformation is given by $\mathbf{B}' = a\mathbf{B}+b\mathbf{1}\mathbf{1}^\top$ where $\mathbf{1} \in \mathbb{R}^K$ is the all-one vector. This is a rank-one perturbation of the scaled matrix $a\mathbf{B}$, therefore, by Horn and Johnson [32], Corollary 4.3.9, the eigenvalues of $\mathbf{B}'$ interweave with those of $\mathbf{B}$,

$$\lambda_k(\mathbf{B}) \leq \lambda_k(\mathbf{B}') \leq \lambda_{k+1}(\mathbf{B}) \leq \lambda_{k+1}(\mathbf{B}').$$

Figure 5.6 shows an example of this eigenvalue behaviour. The only eigenvalue that can change sign after affine transformation is $\lambda_q(\mathbf{B}')$. In the example shown, $\lambda_q(\mathbf{B}') > 0$, meaning the signature of the stochastic block model after affine transformation is $(p + 1, q - 1)$. While the Chernoff information remains unchanged for full-rank affine transformations, the signature of the stochastic block model can be altered. However, if $\lambda_q(\mathbf{B}') < 0$, then the signature would have remained (p, q).

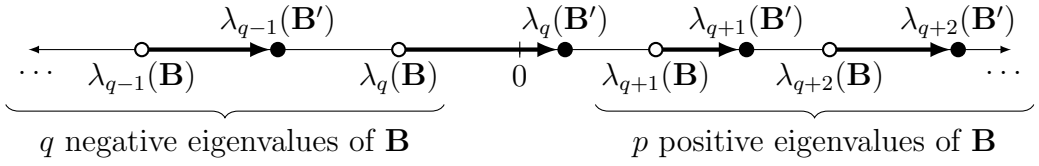

Figure 5.6: Number line showing eigenvalues of mean block matrices close to the origin. White nodes represent eigenvalues of $\mathbf{B}$ with full rank d and signature (p, q), black nodes represent eigenvalues of $\mathbf{B}' = a\mathbf{B} + b\mathbf{1}\mathbf{1}^\top$ with signature $(p + 1, q - 1)$.

The only remaining possibility is that $\lambda_q(\mathbf{B}') = 0$, meaning that $\mathbf{B}'$ is not full rank. In this case, the affine transformation stochastic block model would instead need to be embedded into a lower dimension and the conditions of Lemma 8 do not hold. Since $\mathbf{B}'$ is rank-deficient, the size-adjusted Chernoff information of the affine transformation stochastic block model is less than or equal to that of the original stochastic block model due to the comment made after the proof of Lemma 7.

Lemma 8 has some interesting consequences for common data transformations. For example, given an unweighted stochastic block model, any two distinct values could be used to represent edges and non-edges. In particular, a stochastic block model graph and its complement have the same size-adjusted Chernoff information.

Another consequence regards the edge-wise regularisation of the adjacency matrix where a small value is added to every edge to aid clustering [4]. Usually the extra weight is scaled by the number of nodes n in the network to create the regularised adjacency matrix,

$$\mathbf{A}' = \mathbf{A} + \frac{\tau}{n}\mathbf{1}\mathbf{1}^\top.$$

This is an affine transformation with $a = 1$ and $b = \tau/n$, so Lemma 8 states that applying this transformation to a stochastic block model does not change the size-adjusted Chernoff information. Therefore, for a stochastic block model, there is no reason to use the regularised adjacency matrix according to the size-adjusted Chernoff information.

5.2.2 Optimal transformations

Since different entry-wise transformations of stochastic block models lead to better community separation than others this raises the question: What is the optimal transformation of an adjacency matrix? Given the stochastic block model structure and the edge weight distributions, it is possible to find

the function g that maximises the size-adjusted Chernoff information. While it may be unrealistic to assume that we know the details of a stochastic block model before analysing it, knowing the forms of some optimal transformations gives insight into what data transformations may be useful.

Example 27 (Homogeneous weighted stochastic block model continued)**.** Consider the homogeneous weighted stochastic block model described in Example 23. Using the size-adjusted Chernoff information, the aim is to find the function g that maximises

$$C[g] = \frac{\{\mathbb{E}(g(X_1)) - \mathbb{E}(g(X_2))\}^2}{2K\{\mathrm{var}(g(X_1)) + \mathrm{var}(g(X_2))\}}.$$

where $X_i \sim F_i$.

Without loss of generality, we can choose a transformation such that $\mathbb{E}(g(X_i)) = \mathbb{E}(X_i)$. We can take an affine transformation of the function g to map the means to any value, since by Lemma 8 this does not affect the size-adjusted Chernoff information. Therefore, maximising $C[g]$ is equivalent to minimising the sum of the variances subject to the means being unchanged. This optimisation task can be expressed using Lagrange multipliers,

$$\begin{aligned}
&\mathrm{var}(g(X_1)) + \mathrm{var}(g(X_2)) + \lambda_1\{\mathbb{E}(X_1) - \mathbb{E}(g(X_1))\} + \lambda_2\{\mathbb{E}(X_2) - \mathbb{E}(g(X_2))\} \\
&= \int \left[g(x)^2\{f_1(x) + f_2(x)\} + \lambda_1\{x - g(x)\}f_1(x) + \lambda_2\{x - g(x)\}f_2(x)\right] \mathrm{d}x \\
&= \int L(x, g(x))\,\mathrm{d}x.
\end{aligned}$$

This can be minimised over g using the Euler-Lagrange equation of the calculus of variations, although things are simpler as the function L does not depend on the derivative $g'(x)$. Solving $\frac{\partial L}{\partial g} = 0$ gives the solution

$$g(x) = \frac{\lambda_1 f_1(x) + \lambda_2 f_2(x)}{2(f_1(x) + f_2(x))},$$

where λ_1, λ_2 are chosen so that the conditions $\mathbb{E}(g(X_i)) = \mathbb{E}(X_i)$ are satisfied. However, once again the same argument about affine transformations can be

used and the same size-adjusted Chernoff information is achieved by the entry-wise transformation

$$h(x) = \frac{f_1(x)}{f_1(x) + f_2(x)},$$

removing the need to compute the terms λ_1 and λ_2. The transformation h is a sigmoid function which depends on the ratio of the two probability density functions. $h(x) \to 1$ as $f_1(x)/f_2(x) \to \infty$ meaning it is getting more likely that the observation x comes from the distribution F_1. Conversely, $h(x) \to 0$ as $f_1(x)/f_2(x) \to 0$ meaning it is getting more likely that the observation x comes from the distribution F_2.

In Example 26 with Poisson distribution edge weights, converting edge weights using the function $\mathbb{I}(\mathbf{A}_{ij} > 0)$ can be seen as an approximation of the optimal function h that does not require knowledge of the Poisson distribution parameters. ◁

5.3 Example: Pairwise p-value data

In this example, we describe a synthetic example designed to show the importance of data representation for analysing a computer network. This work was first published in Gallagher et al. [21].

Consider a computer network where packets are sent along connections between computers and suppose we have a test statistic t for the traffic between a pair of computers in the network. For any edge, the p-value is the probability of obtaining a test statistic more extreme than the observed statistic under some model,

$$\mathbf{A}_{ij} = \mathbb{P}(T \geq t(\text{traffic between computers } i \text{ and } j)).$$

A low p-value may occur if the amount of traffic between a pair of computers differs wildly compared to historic behaviour [28], occur over an unusual port [26], or happens at an unusual time of day [56].

The p-values are modelled using an anomalous two-community weighted stochastic block model introduced in Example 24. A proportion π_1 of computers are considered to be acting anomalously and are assigned to community 1. Interactions between those computers tend to generate smaller p-values modelled by a Beta$(\alpha, 1)$ distribution with $\alpha < 1$. The remaining proportion π_2 of computers are considered to be acting normally and are assigned to community 2. For any interaction involving one of these normal computers, the p-value is modelled by a uniform distribution (equivalently, a Beta$(1, 1)$ distribution) corresponding to no anomalous activity.

In a practical application, not every edge in the network can be assigned a p-value. For example, there may not be any observed traffic between two computers, or an initial triage may suggest it is not worth running a full test for anomalous behaviour. To model this behaviour, we assume there is a probability ρ that any edge in the network will be present and treat missing edges as having a p-value equal to one. All together this gives a two-community weighted stochastic block model $\mathbf{A}$ of p-values where matrix entries are distributed according to the symmetric family of distributions $\{H(z_1, z_2) : z_1, z_2 \in \mathcal{Z}\}$ given by

$$H(k, \ell) = \begin{cases} (1-\rho)\delta_1 + \rho \,\text{Beta}(\alpha, 1) & \text{if } k = 1, \ell = 1, \\ (1-\rho)\delta_1 + \rho \,\text{Beta}(1, 1) & \text{otherwise,} \end{cases}$$

where δ_1 is the delta distribution that places all its probability mass at one.

In this example, we discuss three different representations of this network of p-values. The first is the affine entry-wise transformation $\mathbf{A}^{\text{P}}_{ij} = 1 - \mathbf{A}_{ij}$. By Lemma 8, this transformation does not affect the size-adjusted Chernoff information. It has the practical advantage of assigning the missing edges in the network to zero, which makes it easier to use sparse versions of the singular value decomposition algorithm for the adjacency spectral embedding calculation. Entries in $\mathbf{A}^{\text{P}}_{ij}$ follow a zero-inflated beta distribution since, if

$X \sim \text{Beta}(\alpha, \beta)$, then $1 - X \sim \text{Beta}(\beta, \alpha)$.

Alternatively, based on Fisher's method for combining p-values [20], one could consider embedding the matrix of log p-values $\mathbf{A}^{\text{L}}_{ij} = -\log(\mathbf{A}_{ij})$. This is the uniformly most powerful method of combining the p-values if they have a $\text{Beta}(\alpha, 1)$ distribution with $\alpha < 1$ under the alternative hypothesis [27]. Entries in $\mathbf{A}^{\text{L}}_{ij}$ follow a zero-inflated exponential distribution since, if $X \sim \text{Beta}(\alpha, 1)$, then $-\log X \sim \text{Exp}(\alpha)$.

Finally, one could decide to only consider interesting p-values less than a chosen threshold $\tau \in (0, 1)$ [28]. This corresponds to the matrix with entries $\mathbf{A}^{\text{T}}_{ij} = \mathbb{I}(\mathbf{A}_{ij} \leq \tau)$. Entries in $\mathbf{A}^{\text{T}}_{ij}$ follow a Bernoulli distribution since, if $X \sim \text{Beta}(\alpha, 1)$, then $\mathbb{I}(X \leq \tau) \sim \text{Bernoulli}(\tau^\alpha)$, and the zero-inflation parameter ρ can be incorporated into the Bernoulli distribution.

Table 5.1 gives a summary of the original p-values data and the three transformations outlined above. The edge weight distributions for the anomalous weighted stochastic block models when both nodes are in the anomalous community is shown in the right-most column. The edge weight distributions when at least one node is behaving normally can be found be setting $\alpha = 1$.

Data representation	Matrix entries	Anomalous distribution ($\alpha < 1$)
p-values	$\mathbf{A}_{ij}$	$(1-\rho)\delta_1 + \rho\,\text{Beta}(\alpha, 1)$
1−p-values	$\mathbf{A}^{\text{P}}_{ij} = 1 - \mathbf{A}_{ij}$	$(1-\rho)\delta_0 + \rho\,\text{Beta}(1, \alpha)$
log p-values	$\mathbf{A}^{\text{L}}_{ij} = -\log(\mathbf{A}_{ij})$	$(1-\rho)\delta_0 + \rho\,\text{Exp}(\alpha)$
Threshold p-values	$\mathbf{A}^{\text{T}}_{ij} = \mathbb{I}(\mathbf{A}_{ij} \leq \tau)$	$\text{Bernoulli}(\rho\tau^\alpha)$

Table 5.1: The four different representations of the p-values network data and the corresponding edge weight distributions for the anomalous weighted stochastic block models when both nodes are in the anomalous community.

To show the different embeddings, we generate a weighted stochastic block

model with $n = 1000$ nodes with beta distribution parameter $\alpha = 0.25$, zero-inflation probability $\rho = 0.25$, probability a node exhibits anomalous behaviour $\pi_1 = 0.25$, and interesting p-value threshold $\tau = 0.1$. Figure 5.7 shows the asymptotic distribution for the four different embeddings $\hat{\mathbf{X}}$ of the p-values matrix $\mathbf{A}$, the $1-$p-values matrix $\mathbf{A}^{\mathrm{P}}_{ij}$, the log p-values matrix $\mathbf{A}^{\mathrm{L}}_{ij}$ and the threshold p-value matrix $\mathbf{A}^{\mathrm{T}}_{ij}$. Points are coloured according to their community assignment Z_i with the anomalous nodes being represented by red triangles.

In Figures 5.7a and 5.7b the embeddings for the p-values matrix and the $1-$p-values matrix appear very similar, almost mirror images of each other. It is not surprising they have a similar overlap between the two communities as they must have the same size-adjusted Chernoff information. Both stochastic block models have size-adjusted Chernoff information $C = 1.76 \times 10^{-3}$.

In Figures 5.7c and 5.7d the embeddings for the log p-values matrix and the threshold p-values matrix visually have better separation of the two communities compared to the p-values and the $1-$p-values matrices. The log p-values and threshold p-values stochastic block models have size-adjusted Chernoff information $C = 6.43 \times 10^{-3}$ and $C = 6.58 \times 10^{-3}$ respectively, so the threshold p-values is only slightly preferable in this example.

This example raises the question as to whether embedding the log p-values or threshold p-values will always give better community separation than embedding the p-values, for any choice of π_1, α, ρ and τ. Each of the transformations results in an anomalous two-community weighted stochastic block model meaning we can use the formula for the size-adjusted Chernoff information from Example 24,

$$C = \frac{\pi_1(b_1 - b_2)^2}{2(\sqrt{c_1} + \sqrt{c_2})^2},$$

where, for $X \sim F_i$, $b_i = \mathbb{E}(X)$ and $c_i = \mathrm{var}(X)$.

The size-adjusted Chernoff information is a scalar multiple of π_1 so this

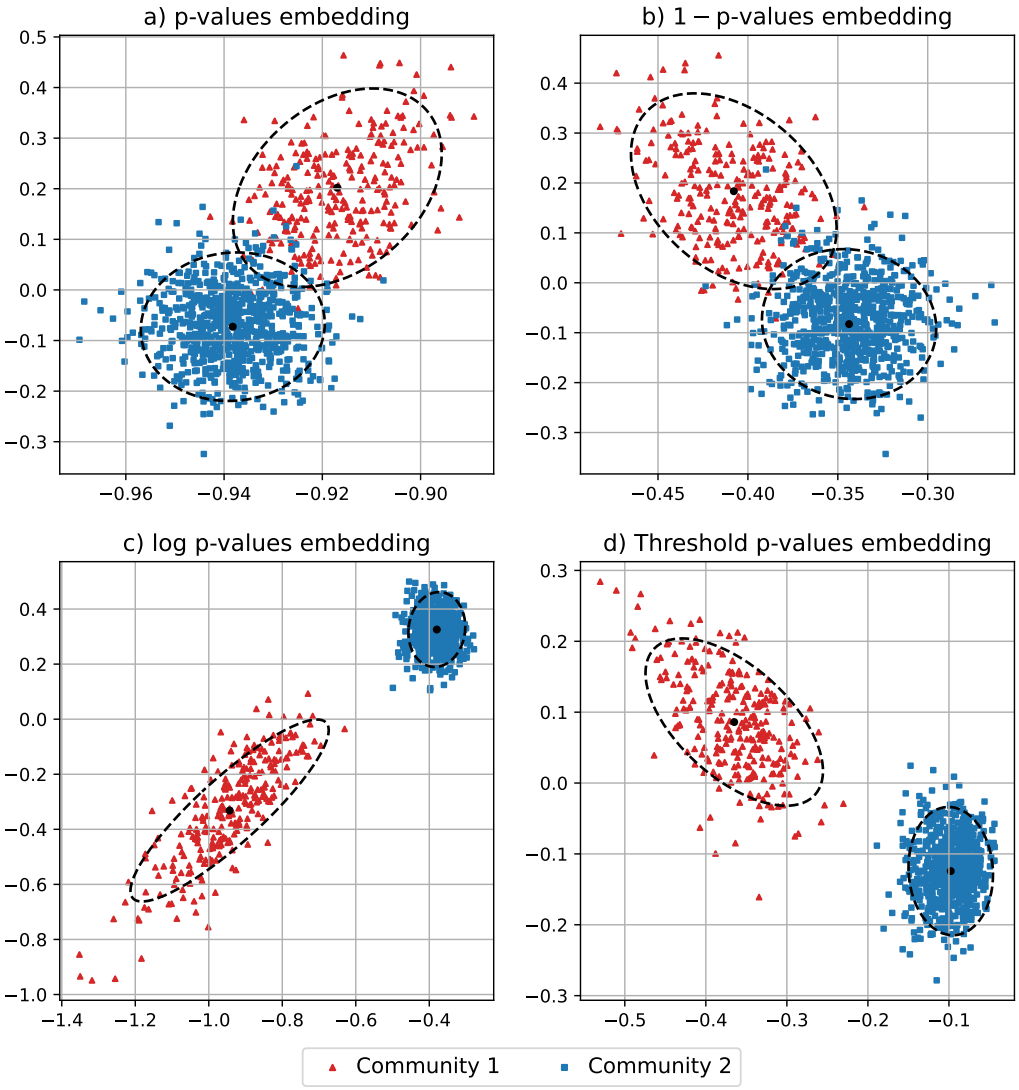

Figure 5.7: The embeddings $\hat{\mathbf{X}}$ of the two-community entry-wise transformed stochastic block model into $d = 2$ dimensions of a) the p-values matrix $\mathbf{A}$, b) the 1−p-values matrix $\mathbf{A}^{\mathrm{P}}$, c) the log p-values matrix $\mathbf{A}^{\mathrm{L}}$ and d) the threshold p-value matrix $\mathbf{A}^{\mathrm{T}}$. The points are coloured according to their true community assignment Z_i. The ellipses show the 95% contours of the asymptotic Gaussian components.

parameter does not effect which entry-wise transformation gives the best size-adjusted Chernoff information. There are no parameter choices where embedding the p-values is preferable to both log p-values and threshold p-values embedding according to the size-adjusted Chernoff information; Figure 5.7 shows that it is often noticeably worse.

Having dismissed embedding p-values directly, Figure 5.8 divides the parameter space $\alpha \in (0, 1)$, $\rho \in (0, 1)$ into two regions showing whether using log p-values or threshold p-values is the better approach. In the white region, embedding log p-values gives larger size-adjusted Chernoff information than embedding the matrix of threshold p-values for any choice of threshold τ. This includes many difficult cases where the beta distribution parameter α is close to one and it is hard to distinguish anomalous p-values from those drawn from the uniform distribution.

In the coloured region, embedding threshold p-values achieves a higher size-adjusted Chernoff information, and the colour indicates the optimal threshold τ. Using threshold p-values is preferable when the zero-inflation parameter for the network ρ is close to zero, although the optimal threshold could be as high as $\tau = 0.15$ depending on the beta distribution parameter α.

In conclusion, neither matrix representation completely dominates in this example, but rather they complement each other in the two cases where it is most difficult to find the anomalous nodes in the network. When it is hard to distinguish anomalous edges from those with no signal, log p-values should be used, and if the zero-inflation parameter for the network is small, threshold p-values should be used.

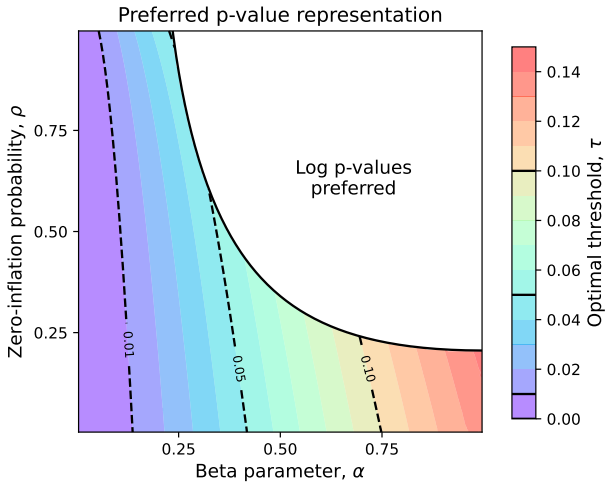

Figure 5.8: Detection of an anomalous network cluster. The parameter $\rho \in (0, 1)$ controls the zero-inflation of the network (low is sparse), and $\alpha \in (0, 1)$ controls the strength of the signal in the p-values (low is strong). In the white region, embedding log p-values is preferred, while in the coloured region, embedding threshold p-values is preferred, where the colour indicates the threshold τ achieving maximal size-adjusted Chernoff information.

Chapter 6

Dynamic networks

In this chapter, rather than embedding a single network, suppose instead we have a time series of adjacency matrices. The goal is to create an embedding for every node at every time point in a statistically consistent way. We propose two desirable properties for such an embedding; cross-sectional stability where nodes behaving similarly at a particular time have comparable embeddings, and longitudinal stability where a node behaving similarly at different time points have comparable embeddings. We show that an existing technique called unfolded adjacency spectral embedding has both of these stability properties. The work was first published in Gallagher et al. [22].

6.1 Unfolded adjacency spectral embedding

Definition 16 (Unfolded adjacency spectral embedding)**.** Given a sequence of T adjacency matrices $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(T)} \in \mathbb{R}^{n \times n}$, compute the d-truncated singular value decomposition of the *unfolded adjacency matrix*,

$$\mathbf{A} = \left(\mathbf{A}^{(1)} \mid \cdots \mid \mathbf{A}^{(T)}\right) = \mathbf{U}_{\mathbf{A}} \mathbf{\Sigma}_{\mathbf{A}} \mathbf{V}_{\mathbf{A}}^{\top} + \mathbf{U}_{\mathbf{A},\perp} \mathbf{\Sigma}_{\mathbf{A},\perp} \mathbf{V}_{\mathbf{A},\perp}^{\top},$$

where $\mathbf{U}_{\mathbf{A}} \in \mathbb{O}(n \times d)$, $\mathbf{V}_{\mathbf{A}} \in \mathbb{O}(nT \times d)$ and $\mathbf{\Sigma}_{\mathbf{A}} \in \mathbb{R}^{d \times d}$ is the diagonal matrix comprising the d largest singular values of $\mathbf{A}$. Denote the *left unfolded adjacency spectral embedding* of $\mathbf{A}$,

$$\hat{\mathbf{X}} = \mathbf{U}_{\mathbf{A}} \mathbf{\Sigma}_{\mathbf{A}}^{1/2} \in \mathbb{R}^{n \times d},$$

and the *right unfolded adjacency spectral embedding* of $\mathbf{A}$,

$$\hat{\mathbf{Y}} = \left(\hat{\mathbf{Y}}^{(1)} \mid \cdots \mid \hat{\mathbf{Y}}^{(T)}\right) = \mathbf{V}_{\mathbf{A}}\mathbf{\Sigma}_{\mathbf{A}}^{1/2} \in \mathbb{R}^{nT \times d}.$$

We divide the left adjacency spectral embedding into rows to represent the global spectral embedding for each node. By writing $\hat{\mathbf{X}} = (\hat{X}_1 \mid \cdots \mid \hat{X}_n)^\top$, $\hat{X}_i$ is the global spectral embedding representation for node i across all the graphs. We divide the right unfolded adjacency spectral embedding into rows to represent the local spectral embedding for each node for each time period. By writing

$$\hat{\mathbf{Y}}^{(t)} = \left(\hat{Y}_1^{(t)} \mid \cdots \mid \hat{Y}_n^{(t)}\right)^\top,$$

$\hat{Y}_i^{(t)} \in \mathbb{R}^d$ is the unfolded adjacency spectral embedding of node i at time t.

We denote the two unfolded adjacency spectral embeddings of an arbitrary matrix $\mathbf{M}$ as $\mathbf{X}_{\mathbf{M}} = \mathbf{U}_{\mathbf{M}}\mathbf{\Sigma}_{\mathbf{M}}^{1/2}$ and $\mathbf{Y}_{\mathbf{M}} = \mathbf{V}_{\mathbf{M}}\mathbf{\Sigma}_{\mathbf{M}}^{1/2}$ employing notation consistent with Definition 16.

Example 28 (Dynamic two-community stochastic block model)**.** Consider a two-community dynamic stochastic block model over two time periods where nodes have a fixed community assignment Z_i for $t = 1$ and $t = 2$. For the two stochastic block models, the block mean matrices are given by

$$\mathbf{B}^{(1)} = \begin{pmatrix} 0.10 & 0.20 \\ 0.20 & 0.05 \end{pmatrix}, \quad \mathbf{B}^{(2)} = \begin{pmatrix} 0.25 & 0.20 \\ 0.20 & 0.05 \end{pmatrix}.$$

The difference between the two block mean matrices is that the probability of an edge between two nodes assigned to community 1 increases for the second stochastic block model. The nodes assigned to community 2 behave the same in both time periods.

We generate a dynamic stochastic block model with $n = 500$ nodes with community assignment $\pi_1 = 1/2$ and $\pi_2 = 1/2$ and perform adjacency spectral embedding on the individual adjacency matrices $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$. Figure 6.1 shows the two adjacency spectral embeddings into $d = 2$ dimensions.

Points are coloured according to their community assignment Z_i. The ellipses show the 95% contours of the asymptotic Gaussian components. In both plots, the two communities can are easily distinguished, but it is not obvious that nodes in community 2 are behaving the same way for $t = 1$ and $t = 2$.

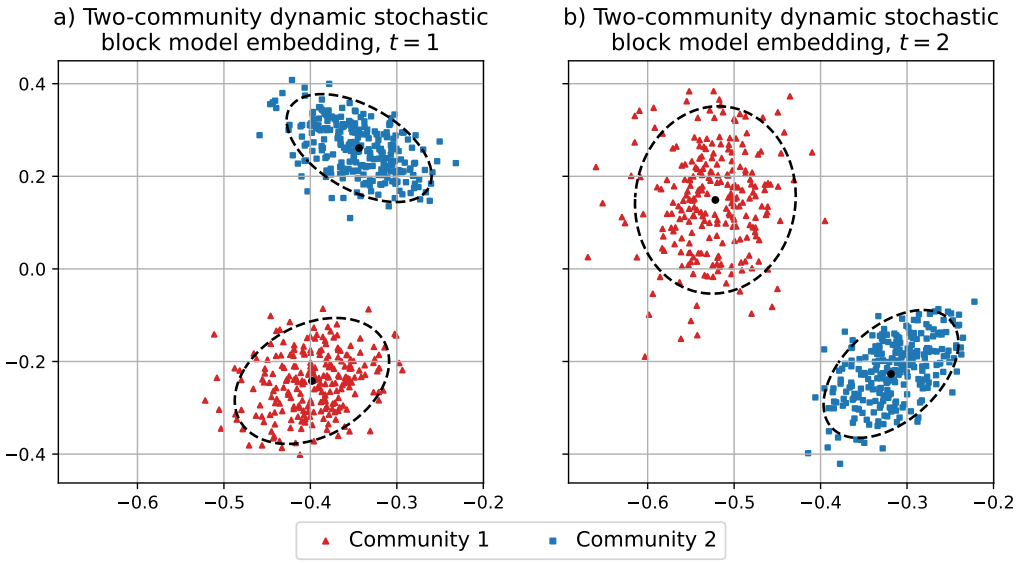

Figure 6.1: The individual adjacency spectral embeddings $\hat{\mathbf{X}}$ of the two-community dynamic stochastic block model into $d = 2$ dimensions for a) $t = 1$ and b) $t = 2$. The points are coloured according to their true community assignment Z_i. The ellipses show the 95% contours of the asymptotic Gaussian components.

Figure 6.2 shows the unfolded adjacency spectral embeddings into $d = 2$ dimensions, where points are coloured according to their community assignment Z_i. The ellipses show the 95% contours of the asymptotic Gaussian components derived in Section 6.4.1. The two communities are distinguishable in both time periods, but now the embedding for community 2 has the same distribution for both $t = 1$ and $t = 2$. This is an example of longitudinal

stability, which is described in Section 6.5. ◁

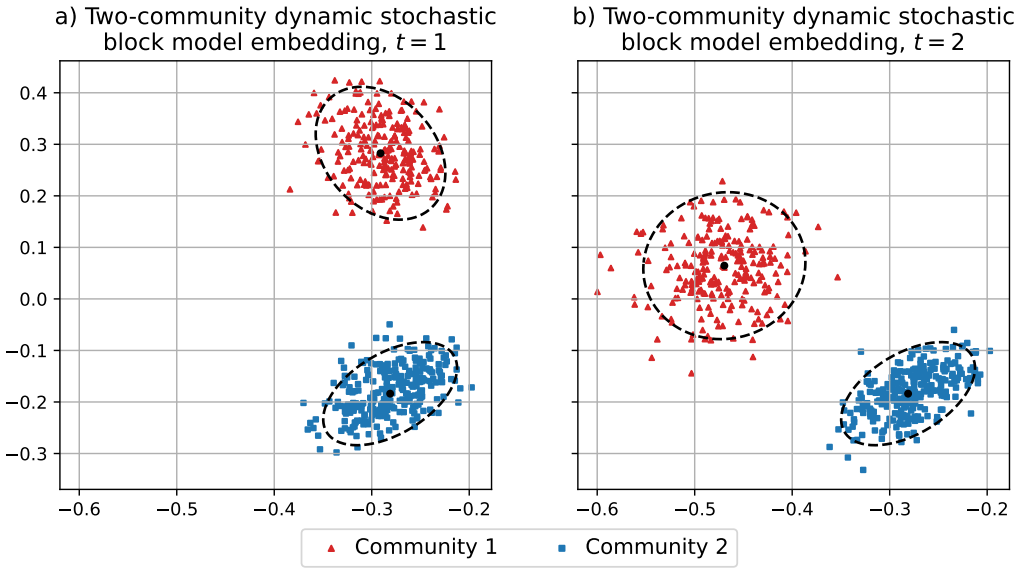

Figure 6.2: The right unfolded adjacency spectral embeddings $\hat{\mathbf{Y}}^{(t)}$ of the two-community dynamic stochastic block model into $d = 2$ dimensions for a) $t = 1$ and b) $t = 2$. The points are coloured according to their true community assignment Z_i. The ellipses show the 95% contours of the asymptotic Gaussian components.

6.2 Dynamic latent position model

We introduce the following model as a way to generate dynamic networks.

Definition 17 (Dynamic latent position model)**.** Let $\mathcal{Z}$ be a sample space and define the trajectory of a node as a sequence of latent positions,

$$Z_i = \left(Z_i^{(1)}, \ldots, Z_i^{(T)}\right) \in \mathcal{Z}^T$$

with distribution F. Given a symmetric family of distributions $\{H(z_1, z_2) = \text{Bernoulli}(f(z_1, z_2)) : z_1, z_2 \in \mathcal{Z}\}$ with function $f : \mathcal{Z} \times \mathcal{Z} \to [0, 1]$, the

symmetric matrices $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(T)} \in \mathbb{R}^{n \times n}$ are distributed as a *dynamic latent position model*, if $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} F$ and, for $i < j$,

$$\mathbf{A}_{ij}^{(t)} \mid Z_i, Z_j \overset{\text{ind}}{\sim} H(Z_i^{(t)}, Z_j^{(t)}).$$

We restrict ourselves to Bernoulli random variables so that the output is a sequence of unweighted graphs, but the model obviously extends to weighted graphs for more general families of distributions.

The dynamic latent position model covers a wide range of models many of them covered in recent reviews [39, 70]. For example, it includes dynamic versions of the stochastic block model [75, 74, 47, 54, 38], the mixed membership stochastic block model [73, 30] and the degree-corrected stochastic block model [45].

It appears the dynamic latent position model could be made more general by allowing the probability function f to also depend on the time t. Given a set of symmetric families of distributions $\{H_t(z_1, z_2) = \text{Bernoulli}(f_t(z_1, z_2)) : z_1, z_2 \in \mathcal{Z}\}$, the symmetric matrices $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(T)} \in \mathbb{R}^{n \times n}$ are distributed as

$$\mathbf{A}_{ij}^{(t)} \mid Z_i, Z_j \overset{\text{ind}}{\sim} H_t(Z_i^{(t)}, Z_j^{(t)}).$$

However, Definition 17 already includes this possibility. The different functions f_t can be combined into a single f by creating a new sample space $\mathcal{Z}'$ where the latent positions for each t are distinct and the function f is equal to f_t in the corresponding regions of $\mathcal{Z}'$.

An important feature of the dynamic latent position model is that the latent positions $Z_i \in \mathcal{Z}^T$ are independent. The latent positions $Z_i^{(t)} \in Z$ that make up the entries of Z_i are allowed to depend on each other. In fact, a common model for dynamic networks assumes a Markov process for the dynamics of $Z_i^{(t)}$ [64, 66]. The assumption that the latent positions Z_i are independent says that a node cannot change its behaviour based on another node, which may be unrealistic for real-world networks. For example, in a

computer network a virus may infect new computers by sending malware across the network. The unfolded adjacency spectral embedding can be used to test this independence assumption by finding causal links between the embeddings $\hat{Y}_i^{(t)}$ and $\hat{Y}_j^{(t)}$ over time.

Example 29 (Two-community dynamic latent position model)**.** Consider a dynamic latent position model consisting of two stochastic block models each with two-communities and mean block matrices,

$$\mathbf{B}^{(1)} = \begin{pmatrix} 0.10 & 0.20 \\ 0.20 & 0.05 \end{pmatrix}, \quad \mathbf{B}^{(2)} = \begin{pmatrix} 0.20 & 0.05 \\ 0.05 & 0.10 \end{pmatrix}.$$

Unlike Example 28, nodes are allowed to change community assignment over time according to a Markov model. At time $t = 1$, nodes are randomly assigned to the two communities with probabilities $\mathbb{P}(Z^{(1)} = 1) = 2/3$ and $\mathbb{P}(Z^{(1)} = 2) = 1/3$. At time $t = 2$, a node in community 1 with equal probability stays in the same community or switches to community 2, while a node in community 2 does not change community,

$$\begin{aligned} &\mathbb{P}(Z^{(2)} = 1 \mid Z^{(1)} = 1) = 1/2, \quad \mathbb{P}(Z^{(2)} = 1 \mid Z^{(1)} = 2) = 0, \\ &\mathbb{P}(Z^{(2)} = 2 \mid Z^{(1)} = 1) = 1/2, \quad \mathbb{P}(Z^{(2)} = 2 \mid Z^{(1)} = 2) = 1. \end{aligned}$$

The adjacency matrices $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$ are generated using this dynamic latent position model with $n = 500$ nodes. Figure 6.3 shows the unfolded adjacency spectral embeddings into $d = 2$ dimensions, where points are coloured according to their community assignment $Z_i^{(t)}$. The ellipses show the 95% contours of the asymptotic Gaussian components derived in Section 6.4.1.

The faint red triangles in Figure 6.3a and faint blue squares in Figure 6.3b correspond to the nodes that change community over time, $Z_i^{(1)} = 1$ and $Z_i^{(2)} = 2$. These nodes have the same asymptotic distribution as the other nodes in that community at that time rather than appearing as a third cluster

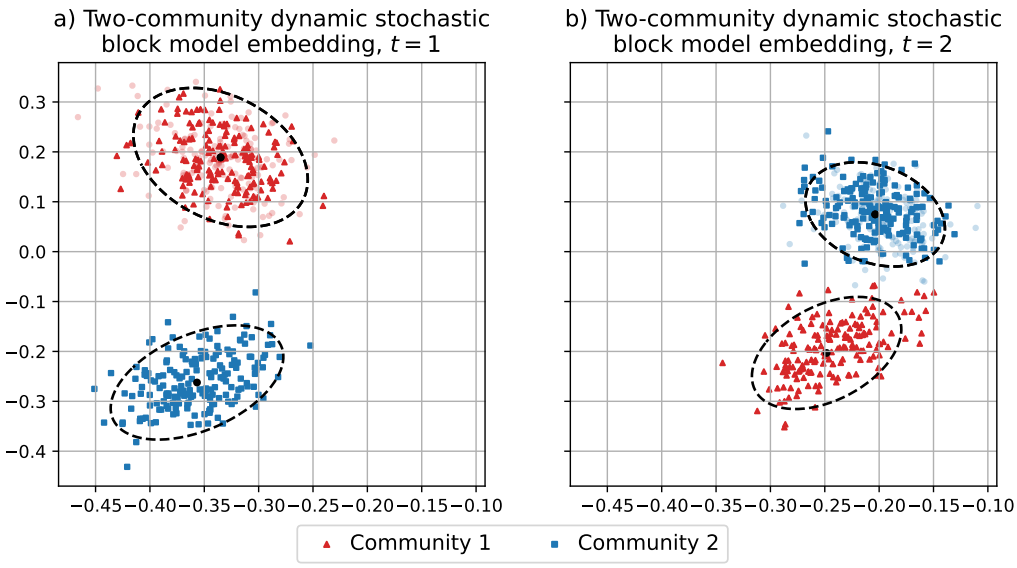

Figure 6.3: The right unfolded adjacency spectral embeddings $\hat{\mathbf{Y}}^{(t)}$ of the two-community dynamic stochastic block model into $d = 2$ dimensions for a) $t = 1$ and b) $t = 2$. The points are coloured according to their true community assignment Z_i, where nodes that change community between time periods are less opaque. The ellipses show the 95% contours of the asymptotic Gaussian components.

in the embedding. This is an example of cross-sectional stability, which is described in Section 6.5.

It is possible to rewrite this network model as a dynamic stochastic block model with three communities that are fixed over both time periods by adding a third community for nodes that switch between communities 1 and 2 in the original model.

$$Z' = \begin{cases} 1 & \text{if } Z^{(1)} = 1, Z^{(2)} = 1, \\ 2 & \text{if } Z^{(1)} = 2, Z^{(2)} = 2, \\ 3 & \text{if } Z^{(1)} = 1, Z^{(2)} = 2, \end{cases}$$

with community assignment probability $\mathbb{P}(Z' = k) = 1/3$ for each of the three communities. These new communities correspond to the possible trajectories through the space of latent positions $\mathcal{Z}^T$ with mean block matrices,

$$\mathbf{B}'^{(1)} = \begin{pmatrix} 0.10 & 0.20 & 0.10 \\ 0.20 & 0.05 & 0.20 \\ 0.10 & 0.20 & 0.10 \end{pmatrix}, \quad \mathbf{B}'^{(2)} = \begin{pmatrix} 0.20 & 0.05 & 0.05 \\ 0.05 & 0.10 & 0.10 \\ 0.05 & 0.10 & 0.10 \end{pmatrix}. \qquad \triangleleft$$

6.3 Multilayer random dot product graph

In this section we introduce the multilayer random dot product graph [36]. This extends the generalised random dot product graph to analyse multiple graphs as long as they share a common set of nodes. This allows for the joint analysis of different types of graphs together, such as directed and undirected graphs, or the rectangular adjacency matrices of bipartite graphs.

In this work, we are interested in the simplest possible scenario where all the graphs are undirected resulting in symmetric adjacency matrices. In this case the definition of a multilayer random dot product graph can be simplified.

Definition 18 (Multilayer random dot product graph)**.** Given a set of matrices $\mathbf{\Lambda}^{(t)} \in \mathbb{R}^{d \times d_t}$, let $\mathcal{X}$ be a subset of $\mathbb{R}^d$ and $\mathcal{Y}_t$ be a subset of $\mathbb{R}^{d_t}$ such that $x^\top \mathbf{\Lambda}^{(t)} y_t \in [0, 1]$ for all $x \in \mathcal{X}, y_t \in \mathcal{Y}_t$, with probability distribution F over $\mathcal{X} \times \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_T$. The symmetric matrices $\mathbf{A}^{(t)} \in \mathbb{R}^{n \times n}$ are distributed as a *multilayer random dot product graph*, if $\left(X_i, Y_i^{(1)}, \ldots, Y_i^{(T)}\right) \overset{\text{iid}}{\sim} F$ for $i = 1, \ldots, n$ and, for $i < j$,

$$\mathbf{A}_{ij}^{(t)} \mid X_i, Y_j^{(t)} \overset{\text{ind}}{\sim} \text{Bernoulli}(X_i^\top \mathbf{\Lambda}^{(t)} Y_j^{(t)}).$$

We refer to $\mathbf{X} = (X_1 \mid \cdots \mid X_n)^\top \in \mathbb{R}^{n \times d}$ as the global latent positions of the multilayer random dot product graph, and similarly, $\mathbf{Y}^{(t)} = (Y_1^{(t)} \mid \cdots \mid$

$Y_n^{(t)})^\top \in \mathbb{R}^{n \times d}$ the local latent positions at time t. The choice of notation here intentionally mirrors that used for the unfolded adjacency spectral embedding. The latent positions $\mathbf{X}$ and $\mathbf{Y}^{(t)}$ are related to the left and right unfolded adjacency spectral embedding $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}^{(t)}$ where Jones and Rubin-Delanchy [36] proves the asymptotic behaviour of both embeddings. For dynamic networks we are primarily interested in how the right embedding $\hat{\mathbf{Y}}^{(t)}$ changes over time.

To be able to use the accompanying asymptotic theory, we must show that the dynamic latent position model in Definition 17 is an example of the multilayer random dot product graph. In order for this to be the case, we have to make some assumptions about the function f.

Assumption 6 (Finite eigendecomposition)**.** The function f has finite eigendecomposition

$$f(x, y) = \sum_{i=1}^{D} \lambda_i u_i(x) u_i(y),$$

with non-zero eigenvalues λ_i and corresponding eigenfunctions u_i.

The assumption that $D < \infty$ is essentially stating that the adjacency matrix $\mathbf{A}$ has low rank even as the number of nodes in the network tends to infinity, a phenomenon often seen in real-world data [71]. The dynamic versions of the stochastic block model, mixed membership stochastic block model and degree-corrected stochastic block model all satisfy Assumption 6.

Lemma 9. The dynamic latent position model satisfying Assumption 6 is an instance of the multilayer random dot product graph.

Sketch proof. The aim is to find maps $\psi : \mathcal{Z}^T \to \mathbb{R}^d$ and $\psi_t : \mathcal{Z} \to \mathbb{R}^{d_t}$ for some d and d_t such that, for the sequence of latent positions $Z_i = \left(Z_i^{(1)}, \ldots, Z_i^{(T)}\right) \in \mathcal{Z}^T$,

$$f(Z_i^{(t)}, Z_j^{(t)}) = \psi(Z_i)^\top \mathbf{\Lambda}^{(t)} \psi_t(Z_j^{(t)}),$$

for $\mathbf{\Lambda}^{(t)} \in \mathbb{R}^{d \times d_t}$, then setting $X_i = \psi(Z_i)$ and $Y_j^{(t)} = \psi_t(Z_j^{(t)})$ in Definition 18 for the multilayer random dot product graph.

Given Assumption 6, define the map $\phi : \mathcal{Z} \to \mathbb{R}^D$ component-wise using the eigendecomposition of the function f for all $z \in \mathcal{Z}$,

$$\phi(z) = \left(|\lambda_1|^{1/2} u_1(z), \ldots, |\lambda_D|^{1/2} u_D(z)\right),$$

so that for $z_1, z_2 \in Z$, $f(z_1, z_2) = \phi(z_1)^\top \mathbf{I}_{p,q} \phi(z_2)$, where $\mathbf{I}_{p,q} \in \mathbb{R}^{D \times D}$ is the diagonal matrix consisting of the signs of the eigenvalues.

This would be enough to complete the proof except, for the random variable $Z = \left(Z^{(1)}, \ldots, Z^{(T)}\right) \sim F$, the support of $\left(\phi(Z^{(1)}), \ldots, \phi(Z^{(T)})\right)$ will be a d-dimensional subspace of $\mathbb{R}^{DT}$ and the support of $\phi(Z^{(t)})$ a d_t-dimensional subspace of $\mathbb{R}^T$. The asymptotic results regarding the multilayer random dot product model require a minimal dimensionality assumption similar to Assumptions 1 and 3 so some redundancy needs to be removed.

There exists a linear map $L : \mathbb{R}^{TD} \to \mathbb{R}^d$ such that, for $z = (z^{(1)}, \ldots, z^{(T)}) \in \mathcal{Z}^T$, the map $\psi : \mathcal{Z}^T \to \mathbb{R}^d$ defined by

$$\psi(z) = L(\{\phi(z^{(1)}) \mid \cdots \mid \phi(z^{(T)})\})$$

is full rank. Similarly, there exists linear maps $L_t : \mathbb{R}^T \to \mathbb{R}^{d_t}$ such that the maps $\psi_t : \mathcal{Z} \to \mathbb{R}^{d_t}$ defined by

$$\psi_t(z^{(t)}) = L_t(\phi(z^{(t)}))$$

are also full rank. Using these linear maps and the eigendecomposition of the function f, we can construct $\mathbf{\Lambda}^{(t)} \in \mathbb{R}^{d \times d_t}$ such that

$$f(Z_i^{(t)}, Z_j^{(t)}) = \psi(Z_i)^\top \mathbf{\Lambda}^{(t)} \psi_t(Z_j^{(t)}).$$

More details can be found in the Appendix of Gallagher et al. [22]. □

6.4 Asymptotic results

Since the dynamic latent position network model can be written as a multilayer random dot product graph, Theorems 2 and 3 from Jones and Rubin-Delanchy [36] can be applied to the dynamic latent position model to create versions of uniform consistency and a central limit theorem for the unfolded adjacency spectral embedding.

These theorems differ slightly in flavour to Theorem 1 and 2 for the generalised random dot product graph. Rather than relating $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ directly, the comparison is made through an intermediate embedding. Let $\mathbf{P} = \left(\mathbf{P}^{(1)} \mid \cdots \mid \mathbf{P}^{(T)}\right) \in [0,1]^{nT \times n}$ be the unfolded matrix of probabilities that two nodes in the network form an edge at different times,

$$\mathbf{P}_{ij}^{(t)} = \mathbb{E}(\mathbf{A}_{ij}^{(t)}) = f(Z_i^{(t)}, Z_j^{(t)}).$$

The unfolded adjacency matrix $\mathbf{A}$ is a noisy observation of the unfolded probability matrix $\mathbf{P}$.

We denote the noise-free embedding $\tilde{Y}_i^{(t)} \in \mathbb{R}^d$ for each node and every time period using the right unfolded spectral embedding of $\mathbf{P}$,

$$\mathbf{Y}_{\mathbf{P}}^{(t)} = \left(\tilde{Y}_1^{(t)} \mid \cdots \mid \tilde{Y}_n^{(t)}\right)^{\top}.$$

The noise-free embeddings $\tilde{Y}_i^{(t)}$ are a linear transformation of the latent positions of the multilayer random dot product graph latent positions $Y_i^{(t)} = \psi_t(Z_i^{(t)})$ defined using the latent positions of the dynamic latent position model in the proof of Lemma 9 [22].

The following uniform consistency and central limit theorem connects the spectral embedding $\hat{Y}_i^{(t)}$ to the noise-free embedding $\tilde{Y}_i^{(t)}$.

Theorem 5 (Unfolded adjacency spectral embedding uniform consistency)**.** Let $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(T)}$ be an instance of a dynamic latent position model satisfying Assumption 6. There exists a sequence of matrices $\tilde{\mathbf{W}}_n \in \mathbb{O}(d)$ such

that

$$\max_{i\in\{1,\ldots,n\}} \left\| \tilde{\mathbf{W}}_n \hat{Y}_i^{(t)} - \tilde{Y}_i^{(t)} \right\| \xrightarrow{\mathbb{P}} 0.$$

Note that the sequence $\tilde{\mathbf{W}}_n$ is the same for each embedding. This highlights that the connection between the embeddings for the different time periods, an idea first discovered for the generalised random dot product graph [2]. After applying a second orthogonal transformation, the error converges in distribution to a fixed Gaussian distribution.

Theorem 6 (Unfolded adjacency spectral embedding central limit theorem)**.** Let $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(T)}$ be an instance of a dynamic latent position model satisfying Assumption 6. Given $z \in \mathcal{Z}$, define the covariance-valued function

$$\mathbf{\Sigma}^{(t)}(z) = \mathbb{E}\left[f(z, Z^{(t)})(1 - f(z, Z^{(t)}))\psi(Z)\psi(Z)^\top \right],$$

where $Z = \left(Z^{(1)}, \ldots, Z^{(T)}\right) \sim F$. Then, there exists a deterministic matrix $\mathbf{R} \in \mathbb{R}^{d\times d}$ and a sequence of matrices $\mathbf{W}_n, \tilde{\mathbf{W}}_n \in \mathbb{O}(d)$ such that, for all $\mathbf{y} \in \mathbb{R}^d$,

$$\mathbb{P}\left(n^{1/2}\mathbf{W}_n \left(\tilde{\mathbf{W}}_n \hat{Y}_i^{(t)} - \tilde{Y}_i^{(t)} \right) \leq \mathbf{y} \mid Z_i^{(t)} = z \right) \to \Phi\left(\mathbf{y}, \mathbf{R}\mathbf{\Sigma}^{(t)}(z)\mathbf{R}^\top\right).$$

In Theorem 2 for the generalised random dot product graph, the covariance function depended on the second moment matrix $\mathbf{\Delta} = \mathbb{E}(\psi(Z)\psi(Z)^\top)$. In this version of the central limit theorem, the corresponding second moment matrix is being absorbed into the deterministic matrix $\mathbf{R}$.

6.4.1 Dynamic stochastic block model asymptotic distribution

To give one application of the asymptotic theory, we consider a dynamic K-community stochastic block model where nodes are assigned to one community for all time periods with probability $\pi_k = \mathbb{P}(Z = k)$ but with dynamic

block mean matrices $\mathbf{B}^{(t)}$ (and dynamic block variance matrices $\mathbf{C}^{(t)}$) that change over time. One such example was outlined in Example 28, and in Example 29 we showed a dynamic stochastic block model where the community assignments changing over time can be redefined to fit this description. In general, to do this one needs K^T new communities for every possible latent position sequence $Z = \left(Z^{(1)}, \ldots, Z^{(T)}\right)$ but it is unlikely every possible sequence would be present.

Using the unfolded block mean matrix $\mathbf{B} = (\mathbf{B}^{(1)} \mid \cdots \mid \mathbf{B}^{(T)}) \in [0,1]^{K \times KT}$, we construct latent positions for the multilayer random dot product graph using its unfolded adjacency spectral similar to the method in Section 3.2.1 for a single stochastic block model,

$$\begin{aligned} X_i &= \psi(Z_i) = (\mathbf{X_B})^\top_{Z_i}, \\ Y_i^{(t)} &= \psi_t(Z_i) = (\mathbf{Y_B}^{(t)})^\top_{Z_i}. \end{aligned}$$

In this case, the matrices $\mathbf{\Lambda}^{(t)}$ from the multilayer random dot product graph are equal to the identity matrix for all t since $\mathbf{B}^{(t)} = \mathbf{X_B}\mathbf{Y_B}^{(t)\top}$. Under this construction, there are only K possible values for the covariance-valued function $\mathbf{\Sigma}^{(t)}(z)$ corresponding to the local latent positions $z = (\mathbf{Y_B}^{(t)})^\top_k$ which we denote $\mathbf{\Sigma}^{(t)}_k$. For random variable $Z = \left(Z^{(1)}, \ldots, Z^{(T)}\right) \sim F$

$$\begin{aligned} \mathbf{\Sigma}^{(t)}_k &= \mathbf{\Sigma}^{(t)}((\mathbf{Y_B}^{(t)})^\top_k) \\ &= \mathbb{E}\left[f(k, Z^{(t)})(1 - f(k, Z^{(t)}))\psi(Z)\psi(Z)^\top\right] \\ &= \sum_{\ell=1}^{K} \pi_\ell \mathbf{B}^{(t)}_{k\ell} \left(1 - \mathbf{B}^{(t)}_{k\ell}\right) (\mathbf{X_B})^\top_\ell \, (\mathbf{X_B})_\ell \\ &= \mathbf{X}_\mathbf{B}^\top \mathbf{\Pi} \mathbf{\Lambda}^{(t)}_k \mathbf{X_B}, \end{aligned}$$

where $\mathbf{\Pi} = \operatorname{diag}(\boldsymbol{\pi}) \in \mathbb{R}^{K \times K}$ is the diagonal matrix consisting of the community assignment probabilities and $\mathbf{\Lambda}^{(t)}_k = \operatorname{diag}((\mathbf{C}^{(t)})_k) \in \mathbb{R}^{K \times K}$ is the diagonal matrix consisting of the row of the block variance matrix corresponding to community k at time t. The second moment matrix, which is

part of the $\mathbf{R}$ term in Theorem 6, is given by

$$\begin{aligned}\mathbf{\Delta} &= \mathbb{E}(\psi(Z)\psi(Z)^\top) \\ &= \sum_{k=1}^{K} \pi_k \left(\mathbf{X_B}\right)_k^\top \left(\mathbf{X_B}\right)_k \\ &= \mathbf{X_B^\top \Pi X_B}.\end{aligned}$$

Writing things similarly to Section 3.3.1, there exists a sequence of matrices $\mathbf{Q}_n \in \mathbb{O}(d)$ such that the points $\mathbf{Q}_n \hat{Y}_i^{(t)}$ have an approximate Gaussian mixture model distribution with components Normal$((\mathbf{Y_B}^{(t)})_k^\top, \mathbf{\Delta}^{-1}\mathbf{\Sigma}_k^{(t)}\mathbf{\Delta}^{-1}/n)$ with components probability π_k. Since $\mathbf{Q}_n$ is a linear transformation, the unfolded adjacency spectral embedding $\hat{\mathbf{Y}}^{(t)}$ also has an approximate Gaussian mixture model distribution. This is how the asymptotic distributions shown in Figure 6.2 and 6.3 are calculated.

6.4.2 Dynamic stochastic block model Chernoff information

Given the asymptotic distribution for a dynamic stochastic block model derived in the previous section, we can calculate the size-adjusted Chernoff information for each time period. We now denote the time period by s to distinguish from the variable $t \in (0, 1)$ in the size-adjusted Chernoff information

$$C^{(s)} = \min_{k \neq \ell} \sup_{t \in (0,1)} \left[\frac{t(1-t)}{2} \left\{ (\mathbf{Y_B}^{(s)})_k - (\mathbf{Y_B}^{(s)})_\ell \right\}^\top \mathbf{\Sigma}_{k\ell}^{(s)}(t)^{-1} \left\{ (\mathbf{Y_B}^{(s)})_k - (\mathbf{Y_B}^{(s)})_\ell \right\} \right],$$

where $\mathbf{\Sigma}_{k\ell}^{(s)}(t) = (1-t)\mathbf{\Sigma}_k^{(s)} + t\mathbf{\Sigma}_\ell^{(s)}$.

For simplicity, assume that the right embedding $\mathbf{X_B}$ is full rank, although a similar calculation can be done when it is not. Substituting the mean and covariances from the approximate Gaussian mixture model distribution into

the objective function in the size-adjusted Chernoff information gives

$$\begin{aligned}
&\left\{(\mathbf{Y}_{\mathbf{B}}^{(s)})_k - (\mathbf{Y}_{\mathbf{B}}^{(s)})_\ell\right\}^\top \boldsymbol{\Sigma}_{k\ell}^{(s)}(t)^{-1}\left\{(\mathbf{Y}_{\mathbf{B}}^{(s)})_k - (\mathbf{Y}_{\mathbf{B}}^{(s)})_\ell\right\} \\
&= (\mathbf{e}_k - \mathbf{e}_\ell)^\top \mathbf{Y}_{\mathbf{B}}^{(s)}(\mathbf{X}_{\mathbf{B}}^\top \boldsymbol{\Pi}\mathbf{X}_{\mathbf{B}})(\mathbf{X}_{\mathbf{B}}^\top \boldsymbol{\Pi}\boldsymbol{\Lambda}_{k\ell}^{(s)}(t)\mathbf{X}_{\mathbf{B}})^{-1}(\mathbf{X}_{\mathbf{B}}^\top \boldsymbol{\Pi}\mathbf{X}_{\mathbf{B}})\mathbf{Y}_{\mathbf{B}}^{(s)\top}(\mathbf{e}_k - \mathbf{e}_\ell) \\
&= (\mathbf{e}_k - \mathbf{e}_\ell)^\top \mathbf{B}^{(s)}\boldsymbol{\Pi}\mathbf{X}_{\mathbf{B}}(\mathbf{X}_{\mathbf{B}}^\top \boldsymbol{\Pi}\boldsymbol{\Lambda}_{k\ell}^{(s)}(t)\mathbf{X}_{\mathbf{B}})^{-1}\mathbf{X}_{\mathbf{B}}^\top \boldsymbol{\Pi}\mathbf{B}^{(s)}(\mathbf{e}_k - \mathbf{e}_\ell) \\
&= (\mathbf{e}_k - \mathbf{e}_\ell)^\top \mathbf{B}^{(s)}\boldsymbol{\Pi}\boldsymbol{\Lambda}_{k\ell}^{(s)}(t)^{-1}\mathbf{B}^{(s)}(\mathbf{e}_k - \mathbf{e}_\ell),
\end{aligned}$$

where $\boldsymbol{\Lambda}_{k\ell}^{(s)}(t) = (1-t)\boldsymbol{\Lambda}_k^{(s)} + t\boldsymbol{\Lambda}_\ell^{(s)}$ and using the first matrix equality from the proof of Lemma 7 to make the simplification in the final equation.

Comparing this expression to Lemma 7, this is the same as the size-adjusted Chernoff information of the standard stochastic block model $\mathbf{A}^{(s)}$. There is no benefit to analysing the dynamic stochastic block model adjacency matrices separately in terms of the individual size-adjusted Chernoff information. Essentially, the unfolded adjacency spectral embedding is aligning the separate embeddings so they are consistent over time, as shown in Example 28 and 29, which does not affect the size-adjusted Chernoff information.

6.5 Dynamic embedding stability

We now provide a formal definition for two desirable properties of a dynamic embedding. Recall that $f(z_1, z_2), z_1, z_2 \in \mathcal{Z}$ denotes the edge probability used in the dynamic latent position model given by Definition 17.

Definition 19 (Dynamic embedding stability)**.** Given a dynamic latent position model, let the output of a generic dynamic network embedding be denoted as $\hat{Z}_i^{(t)}$. For $Z = \left(Z^{(1)}, \ldots, Z^{(T)}\right) \sim F$, define the following stability properties:

1. *Cross-sectional stability:* For $z, z' \in \mathcal{Z}$, if $f(z, Z^{(t)}) = f(z', Z^{(t)})$ with

probability one, then $\hat{Z}_i^{(t)}$ and $\hat{Z}_j^{(t)}$ are asymptotically equal, with identical error distribution, conditional on $Z_i^{(t)} = z$ and $Z_j^{(t)} = z'$.

2. *Longitudinal stability:* For $z \in \mathcal{Z}$, if $f(z, Z^{(t)}) = f(z, Z^{(t')})$ with probability one, then $\hat{Z}_i^{(t)}$ and $\hat{Z}_i^{(t')}$ are asymptotically equal, with identical error distribution, conditional on $Z_i^{(t)} = z$ and $Z_i^{(t')} = z$.

Example 28 demonstrates an example of longitudinal stability. The probability a node in community 2 forms an edge with the other communities is the same for both time points since the bottom rows of $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ are identical. The asymptotic distribution of community 2 appear the same in Figure 6.2.

Example 29 shows an example of cross-sectional stability. The nodes with new community labels $Z' = 1$ and $Z' = 3$ have the same asymptotic distribution in Figure 6.3 at time $t = 1$ as both correspond to community label $Z^{(1)} = 1$. The fact that they have different behaviour at time $t = 2$ does not alter their distribution for time $t = 1$.

6.5.1 Unfolded adjacency spectral embedding stability

The following lemma is from Gallagher et al. [22] showing that unfolded adjacency spectral embedding has both desirable stability properties.

Lemma 10. Unfolded adjacency spectral embedding demonstrates both cross-sectional and longitudinal stability.

Proof. Since $\mathbf{P}^{(t)} = \mathbf{X_P} \mathbf{Y}_\mathbf{P}^{(t)\top}$, we have the following expression for the noise-free embedding,

$$\tilde{Y}_i^{(t)} = \left(\mathbf{X_P} \mathbf{X}_\mathbf{P}^\top\right)^{-1} \mathbf{X}_\mathbf{P}^\top (\mathbf{P}^{(t)})_i.$$

For $Z = \left(Z^{(1)}, \ldots, Z^{(T)}\right) \sim F$, suppose that $f(z, Z^{(t)}) = f(z', Z^{(t')})$ with probability one for $Z_i^{(t)} = z$ and $Z_j^{(t')} = z'$. This implies that $(\mathbf{P}^{(t)})_i = (\mathbf{P}^{(t')})_j$

and, hence, the means of the two asymptotic distributions are equal, $\tilde{Y}_i^{(t)} = \tilde{Y}_j^{(t')}$. From the covariance function from Theorem 6, we have $\mathbf{\Sigma}^{(t)}(z) = \mathbf{\Sigma}^{(t')}(z')$. Since the matrices $\mathbf{W}_n, \tilde{\mathbf{W}}_n$ and $\mathbf{R}$ are independent of t, the covariances of the two asymptotic distributions are also equal. Therefore, the unfolded adjacency spectral embeddings $\hat{Y}_i^{(t)}$ and $\hat{Y}_j^{(t')}$ are asymptotically equal, with identical error distribution.

Cross-sectional stability corresponds to the case $t = t'$, longitudinal stability to the case $i = j$. □

6.5.2 Other dynamic network embedding stability

In this section we investigate the stability properties of other dynamic network embedding algorithms. For the alternatives to unfolded adjacency spectral embedding, we analyse the stability of the embeddings of the probability matrices $\mathbf{P}^{(1)}, \ldots, \mathbf{P}^{(T)}$. A method found to be unstable in a noise-free condition is not expected to be stable when analysing the adjacency matrices $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(T)}$. This argument says if an embedding algorithm is unstable but without knowing the asymptotic distribution for an embedding, it is not possible to say whether it is cross-sectional or longitudinal stable.

Table 6.1 gives an overview of the stability of a number of classes of dynamic network embedding algorithms by saying where those algorithm are provably unstable. The following subsections go through all the new algorithms in turn, but the unfolded adjacency spectral embedding is the only dynamic network embedding technique that we are aware of with both cross-sectional and longitudinal stability.

Independent adjacency spectral embedding

Independent adjacency spectral embedding computes the spectral embeddings of the matrices $\mathbf{A}^{(t)}$ separately. Embedding the individual adjacency

Embedding method	Description	Instability
Unfolded adjacency	Embed $\mathbf{A} = (\mathbf{A}^{(1)}\|\cdots\|\mathbf{A}^{(T)})$	None
Independent adjacency	Embed $\mathbf{A}^{(t)}$	Longitudinal
Omnibus	Embed $\tilde{\mathbf{A}}$; $\tilde{\mathbf{A}}_{st} = (\mathbf{A}^{(s)} + \mathbf{A}^{(t)})/2$	Cross-sectional
Separate embedding	Embed $\bar{\mathbf{A}}^{(t)} = \sum_k w_k \mathbf{A}^{(t-k)}$	Both
Joint embedding	See main text	Both

Table 6.1: Classes of dynamic network embedding algorithms with their cross-sectional and longitudinal stability properties.

matrices $\mathbf{A}^{(t)}$ are subject to different indefinite orthogonal transformations $\mathbf{Q}^{(t)} \in \mathbb{O}(p, q)$. Therefore, independent adjacency spectral embedding does not have longitudinal stability, which can be seen in Figure 6.2 from Example 28. However, the algorithm does exhibit cross-sectional stability. If the rows $(\mathbf{P}^{(t)})_i$ and $(\mathbf{P}^{(t)})_j$ are equal, then they will be affected by $\mathbf{Q}^{(t)}$ in the same way.

The independent adjacency spectral embeddings can be combined over time, for example, using a singular value decomposition of the combined embeddings [5]. However, these approaches produce a single embedding for each node representing its global behaviour similar to the left unfolded adjacency spectral embedding $\hat{\mathbf{X}}$ rather than a set of local embeddings.

Omnibus spectral embedding

The omnibus method [42] computes the spectral embedding of the matrix $\tilde{\mathbf{A}} \in \{0, 1/2, 1\}^{nT \times nT}$ where the blocks of the omnibus matrix are given by

$$\tilde{\mathbf{A}}_{st} = \frac{1}{2}(\mathbf{A}^{(s)} + \mathbf{A}^{(t)}) \in \mathbb{R}^{n \times n}.$$

The algorithm was designed to test if multiple graphs are identically distributed rather than producing a dynamic embedding, but it has interesting

properties making it worthy of inclusion.

Let $\tilde{\mathbf{P}} \in [0,1]^{nT \times nT}$ be the noise-free version of $\tilde{\mathbf{A}}$ with corresponding blocks $\tilde{\mathbf{P}}_{st}$. If a node has the same probability of forming edges at times t and t', then the rows $(\mathbf{P}^{(t)})_i$ and $(\mathbf{P}^{(t')})_i$ are equal. This means that the rows of $(\tilde{\mathbf{P}}_{st})_i$ and $(\tilde{\mathbf{P}}_{st'})_i$ are also the same for all s, which means the two rows of $\tilde{\mathbf{P}}$ corresponding to this node at times t and t' are the same. Therefore, the noise-free omnibus embedding for the node are the same demonstrating longitudinal stability, but we cannot say with certainty that the noisy omnibus embedding has longitudinal stability.

However, if two nodes have the same probability of forming edges at time t, meaning the rows $(\mathbf{P}^{(t)})_i$ and $(\mathbf{P}^{(t)})_j$ are equal, this does not imply that the rows of $(\tilde{\mathbf{P}}_{st})_i$ and $(\tilde{\mathbf{P}}_{st})_j$ are also the same for all s. Therefore, omnibus embedding does not exhibit cross-sectional stability.

Separate spectral embedding

Separate embedding covers a collection of embedding techniques separately applied to time-averaged matrices, $\bar{\mathbf{A}}^{(t)} = \sum_k w_k \mathbf{A}^{(t-k)}$ where w_k are non-negative weights. This covers a wide range of possibilities; average adjacency matrices $w_k = 1/t$ [65, 8], exponential forgetting factors $w_k = (1-\lambda)^k$ [17, 38], and more complex time series models [62].

Let $\bar{\mathbf{P}}^{(t)} = \sum_k w_k \mathbf{P}^{(t-k)}$ be the noise-free version of the the time-averaged adjacency matrices $\bar{\mathbf{A}}^{(t)}$. If two nodes are behaving identically at time t then $(\mathbf{P}^{(t)})_i = (\mathbf{P}^{(t)})_j$ but, in general, $(\bar{\mathbf{P}}^{(t)})_i \neq (\bar{\mathbf{P}}^{(t)})_j$ as they depend on the past behaviour of the two nodes which are likely different. Therefore, the methods do not have cross-sectional stability. One counterexample is the trivial case with $w_0 = 1$ and $w_k = 0$ elsewhere, reducing to the independent adjacency spectral embedding algorithm.

Embedding the time-averaged adjacency matrices separately runs into the same problem as independent adjacency spectral embedding. Alignment

issues between the different time periods means that the techniques also do not have longitudinal stability.

Joint spectral embedding

Joint embedding covers another collection of techniques where the aim is to find an embedding that fits the observed adjacency matrices $\mathbf{A}^{(t)}$, but also has smoothly transitioning embeddings between time steps. Let $\hat{\mathbf{A}}^{(t)}$ be the estimated adjacency matrix from a generic dynamic embedding $\hat{\mathbf{Z}}^{(t)}$, for example, using the dynamic latent position model, $\hat{\mathbf{A}}_{ij}^{(t)} = f(\hat{\mathbf{Z}}_i^{(t)}, \hat{\mathbf{Z}}_j^{(t)})$. The goal is to minimise the regularised objective function balancing these two costs for some $\alpha \in [0, 1]$,

$$\underset{\hat{\mathbf{Z}}^{(1)},\ldots,\hat{\mathbf{Z}}^{(T)}}{\arg\min} \left\{ \alpha \sum_{t=1}^{T} \|\mathbf{A}^{(t)} - \hat{\mathbf{A}}^{(t)}\|_{\mathrm{F}}^2 + (1-\alpha) \sum_{t=1}^{T-1} \|\hat{\mathbf{Z}}^{(t+1)} - \hat{\mathbf{Z}}^{(t)}\|_{\mathrm{F}}^2 \right\},$$

subject to a low rank constraint on the embeddings $\hat{\mathbf{Z}}^{(1)}, \ldots, \hat{\mathbf{Z}}^{(T)}$. Many algorithms take this kind of approach for fitting models to dynamic networks albeit with different versions of the two cost functions [15, 79, 13, 44].

Under this model, changing the behaviour of just a node at one time point changes the embedding of that node for all other time points, which in turn affects the embedding of all the other nodes. The smoothing over time prevents a dynamic embedding fitted to this type of objective function from having cross-sectional or longitudinal stability. The exception again being the trivial case with $\alpha = 1$, which reduces to the independent adjacency spectral embedding algorithm.

All the techniques described above focus on embedding the adjacency matrix but they can be extended to work on other representations of the network, such as the symmetric Laplacian matrix. For example, we could have performed independent Laplacian spectral embedding, which would also result in cross-sectional stability for the same reason as independent adja-

cency spectral embedding. The time-averaged adjacency matrices from separate spectral embedding can be replaced with the time-averaged symmetric Laplacian matrices [38] and the objective function in joint spectral embedding can try the model the symmetric Laplacian matrices [44].

It is currently further work to create versions of unfolded spectral embedding algorithms for other matrices such as the symmetric Laplacian matrix that have both cross-sectional and longitudinal stability.

6.6 Example: Primary school interactions

The Lyon primary school data set shows the social interactions at a French primary school over two days in October 2009 [67]. The school consisted of 10 teachers and 232 participating students from five school years, each year divided into two classes. Face-to-face interactions were detected when radio-frequency identification devices worn by participants were in close proximity over an interval of 20 seconds and recorded as a pair of anonymous identifiers together with a timestamp. The data is available for download from the Network Repository website[1] [58].

A time series of networks was created by dividing the data into ten hour-long windows on the two days. If at least one interaction was observed between two people in a particular time window, an edge was created connecting the two nodes in the corresponding network. This results in a time series of graphs $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(20)}$ each with $n = 242$ nodes.

Figure 6.4 shows the first two dimensions of the unfolded adjacency spectral embedding $\hat{Y}^{(1)}, \ldots, \hat{Y}^{(20)}$ embedded into $\hat{d} = 10$ dimensions obtained using profile likelihood [80]. This feels reasonable as there are ten classes in the school.

For time windows corresponding to classroom time, such as 09:00–10:00

[1] `https://networkrepository.com`

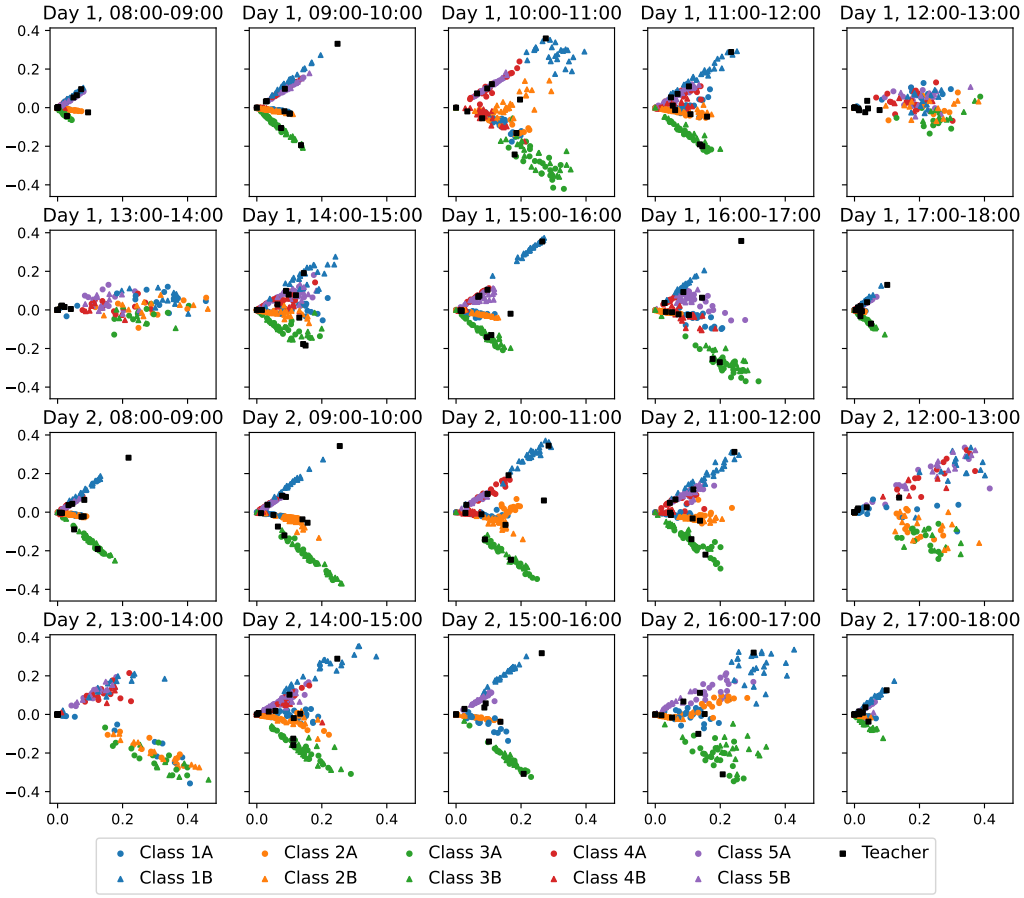

Figure 6.4: The first two dimensions of the right unfolded adjacency spectral embeddings of the Lyon primary school data set $\hat{\mathbf{Y}}^{(t)}$.

and 15:00–16:00, the embedding forms rays of points each corresponding to a single class. For those time periods the network follows a degree-corrected stochastic block model with approximately the same structure for those classroom periods, evidence of longitudinal stability. However, not all time windows exhibit this structure, for instance, the different classes mix more during lunchtimes between 12:00–14:00.

Following recommendations regarding community detection under a degree-corrected stochastic block model [53], the unfolded adjacency spectral embedding $\hat{\mathbf{Y}}^{(t)}$ is converted into spherical coordinates $\hat{\boldsymbol{\theta}}^{(t)}$ by projecting the points onto the unit sphere. Figure 6.5 shows the first two dimensions of this projection, where we can see the rays for each class being projected into clusters of similar points.

To analyse this output, we combine the embeddings from each time period into a single point cloud and fit a Gaussian mixture model. Since unfolded adjacency spectral embedding has cross-sectional and longitudinal stability, this will detect people returning to a previous behaviour in the dynamic network. We fit a Gaussian mixture model with unequal and non-spherical covariances for 30 clusters chosen using the Bayesian Information Criterion and assign each student in each time period to its maximum a posteriori cluster. Figure 6.6 shows how students in the ten different classes move between these clusters over time.

Each class has one or two clusters unique to it, for example, the majority of students in class 1A spend their classroom time assigned to cluster 15 or cluster 20. This highlights the importance of longitudinal stability in unfolded adjacency spectral embedding, as we are detecting points in the embedding returning to the same part of latent space over time. The two clusters could represent two different styles of teaching in those time periods.

There are also instances of multiple school classes being assigned the same cluster at the same time period, for example, on the morning of day 1,

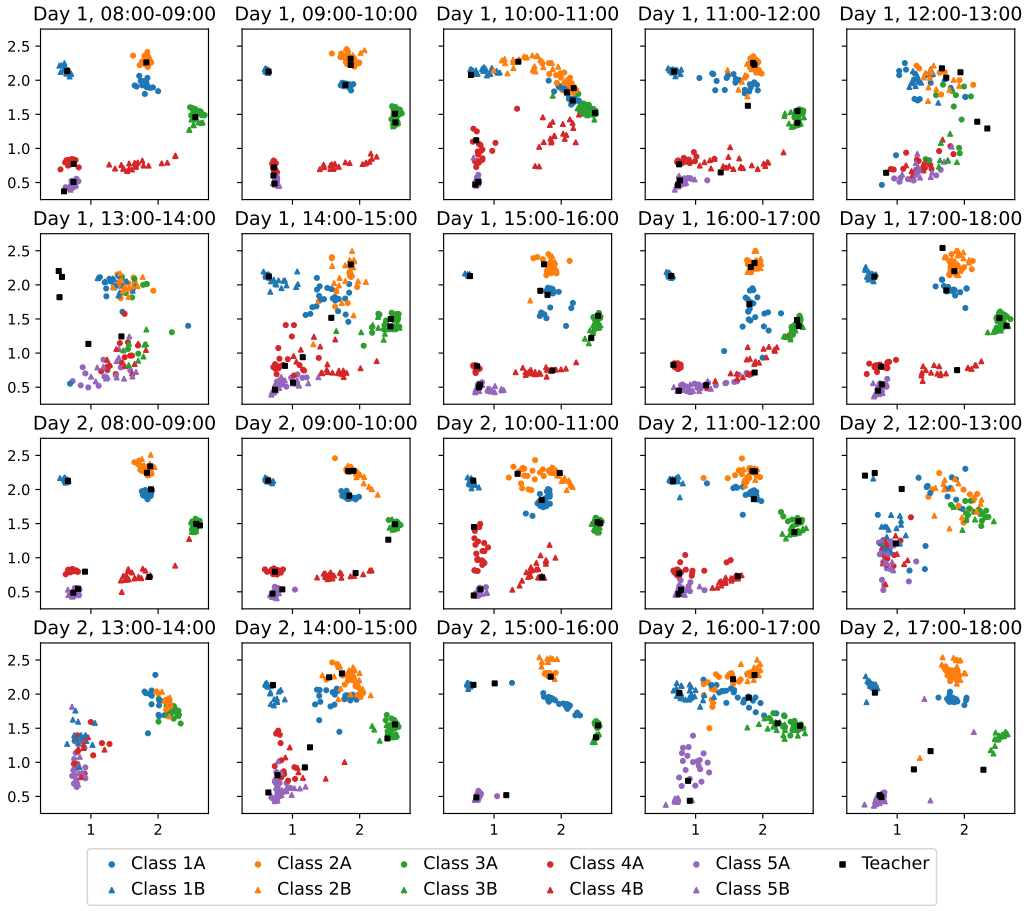

Figure 6.5: The first two dimensions of the right unfolded adjacency spectral embedding of the Lyon primary school data set projected into spherical coordinates $\hat{\boldsymbol{\theta}}^{(t)}$.

Figure 6.6: Bar chart showing the Gaussian cluster assignment of each school class over time. The height of each coloured bar represents the proportion of students, in that class and at that time, assigned to the corresponding Gaussian cluster.

classes 5A and 5B are mainly assigned to cluster 26, perhaps suggesting some form of joint lesson. This shows the importance of cross-sectional stability in unfolded adjacency spectral embedding. It allows the detection of nodes behaving similarly in a specific time window, irrespective of their behaviour across the other time periods.

Bibliography

[1] Abbe, E. (2017). Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531.

[2] Agterberg, J., Tang, M., and Priebe, C. E. (2020). On two distinct sources of nonidentifiability in latent position random graph models. *arXiv preprint arXiv:2003.14250*.

[3] Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of machine learning research*.

[4] Amini, A. A., Chen, A., Bickel, P. J., and Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122.

[5] Arroyo, J., Athreya, A., Cape, J., Chen, G., Priebe, C. E., and Vogelstein, J. T. (2021). Inference for multiple heterogeneous networks with a common invariant subspace. *Journal of Machine Learning Research*, 22(142):1–49.

[6] Athreya, A., Fishkind, D. E., Tang, M., Priebe, C. E., Park, Y., Vogelstein, J. T., Levin, K., Lyzinski, V., and Qin, Y. (2017). Statistical inference on random dot product graphs: a survey. *The Journal of Machine Learning Research*, 18(1):8393–8484.

[7] Bell, R., Koren, Y., and Volinsky, C. (2007). Modeling relationships at multiple scales to improve accuracy of large recommender systems. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 95–104.

[8] Bhattacharyya, S. and Chatterjee, S. (2018). Spectral clustering for multiple sparse networks: I. *arXiv preprint arXiv:1805.10594*.

[9] Cai, H., Zheng, V. W., and Chang, K. C.-C. (2018). A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1616–1637.

[10] Cape, J., Tang, M., and Priebe, C. E. (2019a). On spectral embedding performance and elucidating network structure in stochastic blockmodel graphs. *Network Science*, 7(3):269–291.

[11] Cape, J., Tang, M., and Priebe, C. E. (2019b). The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *The Annals of Statistics*, 47(5):2405–2439.

[12] Chen, F., Roch, S., Rohe, K., and Yu, S. (2021). Estimating graph dimension with cross-validated eigenvalues. *arXiv preprint arXiv:2108.03336*.

[13] Chen, H. and Li, J. (2018). Exploiting structural and temporal evolution in dynamic link prediction. In *Proceedings of the 27th ACM International conference on information and knowledge management*, pages 427–436.

[14] Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507.

[15] Deng, D., Shahabi, C., Demiryurek, U., Zhu, L., Yu, R., and Liu, Y. (2016). Latent space model for road networks to predict time-varying traffic. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1525–1534.

[16] Devroye, L., Györfi, L., and Lugosi, G. (2013). *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media.

[17] Dunlavy, D. M., Kolda, T. G., and Acar, E. (2011). Temporal link prediction using matrix and tensor factorizations. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(2):1–27.

[18] Edgington, E. S. (1972). An additive method for combining probability values from independent experiments. *The Journal of Psychology*, 80(2):351–363.

[19] Erdős, P., Rényi, A., et al. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60.

[20] Fisher, R. A. (1992). Statistical methods for research workers. In *Breakthroughs in statistics*, pages 66–70. Springer.

[21] Gallagher, I., Jones, A., Bertiger, A., Priebe, C., and Rubin-Delanchy, P. (2019). Spectral embedding of weighted graphs. *arXiv preprint arXiv:1910.05534*.

[22] Gallagher, I., Jones, A., and Rubin-Delanchy, P. (2021). Spectral embedding for dynamic networks with stability guarantees. *Advances in Neural Information Processing Systems*, 34.

[23] Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826.

[24] Goyal, P. and Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151:78–94.

[25] Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

[26] Heard, N. and Rubin-Delanchy, P. (2016). Network-wide anomaly detection via the dirichlet process. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 220–224. IEEE.

[27] Heard, N. A. and Rubin-Delanchy, P. (2018). Choosing between methods of combining-values. *Biometrika*, 105(1):239–246.

[28] Heard, N. A., Weston, D. J., Platanioti, K., and Hand, D. J. (2010). Bayesian anomaly detection methods for social networks. *The Annals of Applied Statistics*, 4(2):645–662.

[29] Heimlicher, S., Lelarge, M., and Massoulié, L. (2012). Community detection in the labelled stochastic block model. *arXiv preprint arXiv:1209.2910*.

[30] Ho, Q., Song, L., and Xing, E. (2011). Evolving cluster mixed-membership blockmodel for time-evolving networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 342–350. JMLR Workshop and Conference Proceedings.

[31] Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137.

[32] Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

[33] Hugo, V. (1863). *Les misérables*. C. Lassalle.

[34] Jin, J. (2015). Fast community detection by score. *The Annals of Statistics*, 43(1):57–89.

[35] Jog, V. and Loh, P.-L. (2015). Information-theoretic bounds for exact recovery in weighted stochastic block models using the renyi divergence. *arXiv preprint arXiv:1509.06418*.

[36] Jones, A. and Rubin-Delanchy, P. (2020). The multilayer random dot product graph. *arXiv preprint arXiv:2007.10455*.

[37] Karrer, B. and Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107.

[38] Keriven, N. and Vaiter, S. (2020). Sparse and smooth: improved guarantees for spectral clustering in the dynamic stochastic block model. *arXiv preprint arXiv:2002.02892*.

[39] Kim, B., Lee, K. H., Xue, L., and Niu, X. (2018). A review of dynamic network models with latent variables. *Statistics surveys*, 12:105.

[40] Knuth, D. E. (1993). *The Stanford GraphBase: a platform for combinatorial computing*. ACM.

[41] Lelarge, M., Massoulié, L., and Xu, J. (2013). Reconstruction in the labeled stochastic block model. In *2013 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE.

[42] Levin, K., Athreya, A., Tang, M., Lyzinski, V., Park, Y., and Priebe, C. E. (2017). A central limit theorem for an omnibus embedding of multiple random graphs and implications for multiscale network inference. *arXiv preprint arXiv:1705.09355*.

[43] Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27.

[44] Liu, F., Choi, D., Xie, L., and Roeder, K. (2018). Global spectral clustering in dynamic networks. *Proceedings of the National Academy of Sciences*, 115(5):927–932.

[45] Liu, S., Wang, S., and Krishnan, R. (2014). Persistent community detection in dynamic social networks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 78–89. Springer.

[46] Marshall, A. W. and Olkin, I. (1990). Matrix versions of the Cauchy and Kantorovich inequalities. *Aequationes Mathematicae*, 40(1):89–93.

[47] Matias, C. and Miele, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1119–1141.

[48] Mercer, J. (1909). Functions ofpositive and negativetypeand theircommection with the theory ofintegral equations. *Philos. Trinsdictions Rogyal Soc*, 209:4–415.

[49] Modell, A., Gallagher, I., Cape, J., and Rubin-Delanchy, P. (2022). Spectral embedding and the latent geometry of multipartite networks. *arXiv preprint arXiv:2202.03945*.

[50] Modell, A. and Rubin-Delanchy, P. (2021). Spectral clustering under degree heterogeneity: a case for the random walk laplacian. *arXiv preprint arXiv:2105.00987*.

[51] Nickel, C. L. M. (2008). *Random dot product graphs a model for social networks*. PhD thesis, Johns Hopkins University.

[52] Nielsen, F. (2011). Chernoff information of exponential families. *arXiv preprint arXiv:1102.2684*.

[53] Passino, F. S., Heard, N. A., and Rubin-Delanchy, P. (2020). Spectral clustering on spherical coordinates under the degree-corrected stochastic blockmodel. *arXiv preprint arXiv:2011.04558*.

[54] Pensky, M. and Zhang, T. (2019). Spectral clustering in the dynamic stochastic block model. *Electronic Journal of Statistics*, 13(1):678–709.

[55] Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710.

[56] Price-Williams, M., Turcotte, M., and Heard, N. (2018). Time of day anomaly detection. In *2018 European Intelligence and Security Informatics Conference (EISIC)*, pages 1–6. IEEE.

[57] Priebe, C. E., Park, Y., Vogelstein, J. T., Conroy, J. M., Lyzinski, V., Tang, M., Athreya, A., Cape, J., and Bridgeford, E. (2019). On a two-truths phenomenon in spectral graph clustering. *Proceedings of the National Academy of Sciences*, 116(13):5995–6000.

[58] Rossi, R. A. and Ahmed, N. K. (2015). The network data repository with interactive graph analytics and visualization. In *AAAI*.

[59] Rubin-Delanchy, P. (2020). Manifold structure in graph embeddings. *Advances in Neural Information Processing Systems*, 33:11687–11699.

[60] Rubin-Delanchy, P., Cape, J., Tang, M., and Priebe, C. E. (2017a). A statistical interpretation of spectral embedding: the generalised random dot product graph. *arXiv preprint arXiv:1709.05506*.

[61] Rubin-Delanchy, P., Priebe, C. E., and Tang, M. (2017b). Consistency of adjacency spectral embedding for the mixed membership stochastic blockmodel. *arXiv preprint arXiv:1705.04518*.

[62] Sanna Passino, F., Bertiger, A. S., Neil, J. C., and Heard, N. A. (2021). Link prediction in dynamic networks using random dot product graphs. *Data Mining and Knowledge Discovery*, 35(5):2168–2199.

[63] Sarkar, P. and Bickel, P. J. (2015). Role of normalization in spectral clustering for stochastic blockmodels. *The Annals of Statistics*, 43(3):962–990.

[64] Sarkar, P. and Moore, A. (2005). Dynamic social network analysis using latent space models. *Advances in Neural Information Processing Systems*, 18.

[65] Scheinerman, E. R. and Tucker, K. (2010). Modeling graphs using dot product representations. *Computational statistics*, 25(1):1–16.

[66] Sewell, D. K. and Chen, Y. (2015). Latent space models for dynamic networks. *Journal of the American Statistical Association*, 110(512):1646–1657.

[67] Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Isella, L., Pinton, J.-F., Quaggiotto, M., Van den Broeck, W., Régis, C., Lina, B., et al. (2011). High-resolution measurements of face-to-face contact patterns in a primary school. *PloS one*, 6(8):e23176.

[68] Tang, M. and Priebe, C. E. (2018). Limit theorems for eigenvectors of the normalized laplacian for random graphs. *The Annals of Statistics*, 46(5):2360–2415.

[69] Trendafilov, N. T. and Lippert, R. A. (2002). The multimode procrustes problem. *Linear algebra and its applications*, 349(1-3):245–264.

[70] Turnbull, K. R. (2020). *Advancements in latent space network modelling*. Lancaster University (United Kingdom).

[71] Udell, M. and Townsend, A. (2019). Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160.

[72] Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

[73] Xing, E. P., Fu, W., and Song, L. (2010). A state-space mixed membership blockmodel for dynamic network tomography. *The Annals of Applied Statistics*, 4(2):535–566.

[74] Xu, K. (2015). Stochastic block transition models for dynamic networks. In *Artificial Intelligence and Statistics*, pages 1079–1087. PMLR.

[75] Xu, K. S. and Hero, A. O. (2014). Dynamic stochastic blockmodels for time-evolving social networks. *IEEE Journal of Selected Topics in Signal Processing*, 8(4):552–562.

[76] Xu, M., Jog, V., and Loh, P.-L. (2020). Optimal rates for community estimation in the weighted stochastic block model. *The Annals of Statistics*, 48(1):183–204.

[77] Young, S. J. and Scheinerman, E. R. (2007). Random dot product graph models for social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 138–149. Springer.

[78] Zhang, Y. and Tang, M. (2021). Consistency of random-walk based network embedding algorithms. *arXiv preprint arXiv:2101.07354*.

[79] Zhu, L., Guo, D., Yin, J., Ver Steeg, G., and Galstyan, A. (2016). Scalable temporal latent space inference for link prediction in dynamic social networks. *IEEE Transactions on Knowledge and Data Engineering*, 28(10):2765–2777.

[80] Zhu, M. and Ghodsi, A. (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51(2):918–930.