The Multilayer Random Dot Product Graph

Andrew Jones and Patrick Rubin-Delanchy

University of Bristol, U.K.

July 22, 2020

Abstract

We present an extension of the latent position network model known as the generalised
random dot product graph to accommodate multiple graphs with a common node structure,
based on a matrix representation of the natural third-order tensor created from the adjacency
matrices of these graphs. Theoretical results concerning the asymptotic behaviour of the node
representations obtained by spectral embedding are established, showing that after the appli-
cation of a linear transformation these converge uniformly in the Euclidean norm to the latent
positions with a Gaussian error. The flexibility of the model is demonstrated through appli-
cation to the tasks of latent position recovery and two-graph hypothesis testing, in which it
performs favourably compared to existing models. Empirical improvements in link prediction
over single graph embeddings are exhibited in a cyber-security example.

1. Introduction

Networks permeate the world in which we live, and so developing an accurate understanding of
them is a matter of great interest to many branches of academia and industry, with applications
as diverse as identifying patterns in brain scans [12] and the detection of fraudulent behaviour in
the financial services sector [1]. The mathematical study of random graphs has its roots in the
work of E. N. Gilbert [13] and, contemporaneously, Erdős and Rényi [10], who considered graphs
in which edges between nodes occur independently according to Bernoulli random variables with
a fixed probability $p$, in what can be considered the simplest probabilistic model of a naturally
occurring network (this type of graph now being universally referred to as an Erdős-Rényi graph).

Among the more modern statistical treatments of networks is the concept of a latent position
model [15], in which the $i$th node of a graph is mapped to a vector $X_i$ in some underlying latent
space $X \subseteq \mathbb{R}^d$, and (conditional on this choice of latent positions) the $i$th and $j$th nodes connect
independently with probability $\kappa(X_i, X_j)$, where the kernel function $\kappa : X \times X \to [0,1]$. The random
dot product graph (RDPG, [5]), which uses the kernel function $\kappa(x, y) = x^\top y$, and its generalisation
(GRDPG, [28]), using the kernel function $\kappa(x, y) = x^\top I_{p,q} y$ (where $I_{p,q}$ is the diagonal matrix whose
entries are $p$ ones followed by $q$ minus ones, where $p + q = d$) are of particular interest here due to
their computational tractability and associated statistical estimation theory: spectral embedding
[37] the observed graph based on largest eigenvalues (respectively, largest-in-magnitude) produces
uniformly consistent estimates of the latent positions $X_i$ of the RDPG (respectively, GRDPG)
model up to orthogonal (respectively, indefinite orthogonal) transformation, with asymptotically
Gaussian error [31, 4, 23, 8, 6, 32, 28, 5]. The GRDPG can be used to effectively model networks
which exhibit disassortative connectivity behaviour [18] in which dissimilar nodes are the more
likely to connect to each other, and is therefore the preferred model for studying biological or
technological networks, which typically exhibit such behaviour [24].

Recently, attention has turned to the joint study of multiple graphs. Often, the graphs of interest
share a common set of nodes but have different edges, and such a collection of graphs is known
as a multilayer network [19]. This framework is of interest in the study of dynamic networks, in
which each of the graphs may represent a “snapshot” of a network at a given point in time, and has
been used, for instance, for link prediction in cyber-security applications [27]. Alternatively, one
may be interested in detecting differences in node behaviour between graphs, and this approach
was used to identify regions of the brain associated with schizophrenia by comparing brain scans of
both healthy and schizophrenic patients [21].

Latent position models readily extend to the study of multiple graphs by allowing the kernel func-
tions \(\kappa_r\) to vary, while retaining a common set of latent positions \(X_i\) across the graphs, and in
particular there exist several RDPG-based methods for working with multiple graphs. If each
graph is drawn from the same distribution (that is, \(\kappa_r(x, y) = x'y\) for each \(r\)) then one can
consider the mean embedding [34] by spectrally embedding the average of the adjacency matrices
\(\bar{A} = \frac{1}{m} \sum_{r=1}^m A^{(r)}\), or the omnibus embedding [21], in which each graph is assigned a different em-
embedding in a common latent space. The mean embedding is known to perform well asymptotically
at the task of estimating the latent positions [34], while the omnibus embedding is particularly
suited to the task of testing whether the graphs are drawn from the same underlying distribution.

Other RDPG-based methods are more general, allowing different kernel functions \(\kappa_r\) across the
graphs. In the multiple random eigengraph (MREG, [38]) model, \(\kappa_r(x, y) = x'\Lambda_y y\), where the
matrices \(\Lambda_r \in \mathbb{R}^{d \times d}\) are diagonal with non-negative entries. The multiple random dot product
graph (multi-RDPG, [25]) loosens these restrictions by allowing the matrices \(\Lambda_r\) to be non-diagonal
(but symmetric and positive definite), while requiring that the matrix of latent positions \(X\) have
orthonormal columns. Expanding on this is the common subspace independent edge (COSIE, [3])
graph model which allows the matrices \(\Lambda_r\) to have positive and negative eigenvalues. Each of these
models is proposed along with a spectral embedding technique for latent position estimation. Under
the COSIE model, the adjacency matrix of each component graph is embedded separately, into a
dimension \(d_r\), say. A second, joint, spectral decomposition is then applied to the point clouds, to
find a common embedding of dimension \(d\). The approach requires estimation of each \(d_r\) as well as
\(d\), for which a generic method based on the ‘elbow’ of the scree-plot is suggested [40].

Statistical discourse on network embedding is often inspired by the stochastic block model [16], in
which an unknown partition of the nodes exists so that the nodes of any group (or community) are
statistically indistinguishable. Under this model, a network embedding procedure can reasonably
be expected to ascribe identical positions to the nodes of one group, up to statistical error, and
different embedding techniques can therefore be compared through the theoretical performance of
an appropriate clustering algorithm at recovering the communities [29, 32, 7].

Of the approaches referred to above, only the COSIE model allows estimation of a generic multilayer
stochastic block model [16]. For example, if one graph has assortative community structure (“birds of a feather flock together”) and the other disassortative (“opposites attract”), then the mean embedding can evidently eradicate all community structure visible in either individual graph.

The Multilayer Random Dot Product Graph model (MRDPG), presented here, is equivalent to the COSIE model in terms of its likelihood given latent positions, but the latent positions are themselves defined differently. The spectral embedding method to which this leads is materially different and simpler, while estimation performance is apparently superior (by numerical experiments).

We retain the use of the kernel functions \( \kappa_r(x, y) = x^\top \Lambda_r y \) for each graph, but now only require that the matrices \( \Lambda_r \) be symmetric, with no restriction of orthogonality imposed on the latent positions \( X \). Given a collection of adjacency matrices \( A^{(1)}, \ldots, A^{(m)} \in \{0, 1\}^{n \times n} \), those latent positions are then estimated by the following procedure: the matrix \( A \in \{0, 1\}^{n \times mn} \) is formed by adjoining the matrices \( A^{(r)} \), after which left and right spectral embeddings of \( A \) are obtained by scaling its left and right singular vectors by the square roots of the corresponding singular values. The matrix \( A \), sometimes known as an unfolding, is a natural object to study, as it is one of the standard matrix representations of the third-order tensor \( A \in \{0, 1\}^{n \times n \times m} \) formed from the \( A^{(r)} \), and (up to symmetry) the only such representation to accurately capture the node-to-node behaviour from the corresponding graphs. The left-sided embedding \( X_A \) is our proposed estimate of \( X \), that is, a single embedding of the nodes that is common to all graphs; the right-sided embedding \( Y_A \) can be split into \( m \) distinct embeddings \( Y_{A,r} \), from which the hypothesis that the latent positions are fixed across graphs can be tested.

We allow some of the matrices \( \Lambda_r \) to be of non-maximal rank, requiring instead that the matrix \( \Lambda = [\Lambda_1 | \ldots | \Lambda_m] \) be of maximal rank. This allows for the situation in which information about the latent positions can be obscured in individual graphs. As a simple example, suppose we have a three-party political system in which members always vote along party lines, and that any given graph represents the outcome of a vote on a particular motion (in which members can either support or oppose the motion, and two members are linked if they vote in the same way). Suppose further that there are no coalitions (that is, every pair of parties has at least one motion on which they vote differently). Then any individual vote will only highlight two groups (those who support and those who oppose that particular motion) and it is only with knowledge of multiple votes that one can correctly identify the individual parties. In our method, no intermediate estimate of the rank \( d_r \) of \( \Lambda_r \) is required, in contrast to the COSIE-based approach.

We investigate the asymptotic behaviour of the left- and right-sided spectral embeddings of the unfolding \( A \) under an MRDPG model. It is shown that, up to linear transformation, the rows of each embedding converge uniformly in the Euclidean norm to the latent positions \( X_i \) (Theorem 1) and, through the derivation of a central limit theorem (Theorem 2), that these rows are distributed around their corresponding latent positions according to a Gaussian mixture model. These distributional results show that, in particular, if the graphs are identically distributed then the transformed rows of the left-sided embedding have the same limiting distribution as those of the mean embedding (Corollaries 3 and 4). Consequently, if multiple graphs are identically drawn according to a stochastic block model [16] then joint embedding will always be more effective at the task of cluster separation than any individual graph embedding (Proposition 9), where we evaluate this effectiveness via the Chernoff information [14] of the limiting Gaussian distributions.
of the embeddings. The Chernoff information belongs to the class of $f$-divergences [2, 9] and is therefore invariant under invertible linear transformations [22], an important requirement here since distributional results hold only up to such transformation.

Establishing the asymptotic results relating to the MRDPG model requires more care than for the standard RDPG, with the asymmetry of the unfolding $A$ producing an additional layer of complexity. While one might have hoped that the left- and right-sided embeddings could somehow be considered in isolation, in fact knowledge of the asymptotic behaviour of the singular vectors from both sides is needed before any results can be proved. Over the course of the proofs of our main results we are required to derive several novel asymptotic bounds to accommodate the lack of complete symmetry in the matrix unfolding, which should pave the way for similar results for arbitrary asymmetric adjacency matrices.

Finally, we assess the effectiveness of unfolded adjacency spectral embedding at the inference tasks of recovery of latent positions, estimation of the common invariant subspaces, estimation of the underlying probabilistic model and two-graph hypothesis testing in simulated data. We demonstrate that performance at the estimation tasks is often better than that of the multiple adjacency spectral embedding (the method proposed in [3] to embed multiple graphs distributed according to a COSIE model, which demonstrably yields state of the art performance at such tasks), while its performance at the latter task is comparable with that of the omnibus embedding for reasonably-sized graphs (those with at least 500 nodes). We also apply the UASE to the task of link prediction, using connectivity data from the Los Alamos National Laboratory computer network [36] to predict connections between computers across the entire network as an example of a dynamic link prediction inference task, before restricting our attention to connections occurring through specific ports, demonstrating that the majority of the time the UASE yields greater accuracy than individual adjacency spectral embeddings.

2. The multilayer random dot product graph

Definition 1. (The multilayer random dot product graph model).

Let $A_1, \ldots, A_m \in \mathbb{R}^{d \times d}$ be symmetric matrices (not necessarily of maximal rank) with signature $(p_r, q_r)$, where $p_r \geq 1$, such that the matrix $A = [A_1] \cdots [A_m]$ is of rank $d$. Suppose that $\mathcal{X}$ is a bounded subset of $\mathbb{R}^d$ satisfying $xA_r y^\top \in [0, 1]$ for all $x, y \in \mathcal{X}$ and each $r \in \{1, \ldots, m\}$, and let $F$ be a distribution on $\mathcal{X}$. Draw vectors $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$, define $X = [X_1 | \cdots | X_n]^\top \in \mathbb{R}^{n \times d}$ and, for each $r \in \{1, \ldots, m\}$, set $P^{(r)} = X A_r X^\top \in [0, 1]^{n \times n}$. Finally, let $A^{(1)}, \ldots, A^{(m)} \in \{0, 1\}^{n \times n}$ be symmetric matrices such that, conditional on $X_1, \ldots, X_n$, we have $A^{(r)}_{ij} \sim \text{Bern}(P^{(r)}_{ij})$ for all $i \leq j$ and all $r \in \{1, \ldots, m\}$.

If the above hold, then we say that $A^{(1)}, \ldots, A^{(m)}$ are distributed as a multilayer random dot product graph and write $(A^{(1)}, \ldots, A^{(m)}, X) \sim \text{MRDPG}(F, A_1, \ldots, A_m)$.

If the matrices $A_r$ are all equal to some fixed $I_{p,q}$ with $p + q = d$, we say that $A^{(1)}, \ldots, A^{(m)}$ are identically distributed as a MRDPG, and write $(A^{(1)}, \ldots, A^{(m)}, X) \sim \text{MRDPG}(F, I_{p,q})$.

For ease of notation, we will define the unfoldings $A = [A^{(1)} | \cdots | A^{(m)}]$ and $P = [P^{(1)} | \cdots | P^{(m)}]$. 4
We note that there is a degree of ambiguity in the choice of latent positions $X_i$, in the sense that if $(A^{(1)}, \ldots, A^{(m)}, X) \sim \text{MRDPG}(F, \Lambda_1, \ldots, \Lambda_m)$, then for any linear transformation $H \in \text{GL}(d)$, we also have $(A^{(1)}, \ldots, A^{(m)}, XH^\top) \sim \text{MRDPG}(H \cdot F, H^{-\top} \Lambda_1 H^{-1}, \ldots, H^{-\top} \Lambda_m H^{-1})$. Thus any measure we might use to evaluate the effectiveness of the MRDPG must be invariant under invertible linear transformations.

Key to our study of the MRDPG will be the spectral embeddings of the unfolding $A$. Unlike the GRDPG, in which the left and right singular vectors of the symmetric matrices $\Lambda^{(r)}$ coincide, we obtain distinct embeddings by considering each side of $A$:

**Definition 2.** (Unfolded adjacency spectral embeddings (UASE)).

Let $(A^{(1)}, \ldots, A^{(m)}, X) \sim \text{MRDPG}(F, \Lambda_1, \ldots, \Lambda_m)$, and let $A$ and $P$ admit singular value decompositions

$$A = U_A \Sigma_A V_A^\top + U_A, \perp \Sigma_A, \perp V_A, \perp^\top, \quad P = U_P \Sigma_P V_P^\top,$$

where $U_A, U_P \in \mathbb{O}(n \times d)$, $V_A, V_P \in \mathbb{O}(mn \times d)$, and $\Sigma_A, \Sigma_P \in \mathbb{R}^{d \times d}$ are diagonal containing the largest singular values of $A$ and $P$ respectively. We then define the **left unfolded adjacency spectral embedding** $X_A \in \mathbb{R}^{n \times d}$ by $X_A = U_A \Sigma_A^{1/2}$. Moreover, setting $Y_A = V_A \Sigma_A^{1/2}$, we define the $r$th **right unfolded adjacency spectral embedding** $Y_{A,r} \in \mathbb{R}^{n \times d}$ for $r \in \{1, \ldots, m\}$ to be the $r$th $n \times d$ block of $Y_A$ (equivalently, $Y_{A,r} = V_{A,r} \Sigma_A^{1/2}$, where $V_{A,r}$ is the $r$th $n \times d$ block of $V_A$). We define $X_P$, $Y_P$ and $Y_{P,r}$ analogously.

We note at this juncture that many of the asymptotic bounds we will establish are probabilistic in nature; to accommodate this, we adopt the following convention: given a random variable $X$ and a real valued function $f$, we say that $|X| = O(f(n))$ almost surely if, for any $\rho > 0$, there exist constants $N \in \mathbb{N}$ and $c \in \mathbb{R}$ (possibly dependent on $\rho$) such that $|X| \leq cf(n)$ for all $n \geq N$ with probability at least $1 - n^{-\rho}$. In addition, we write that sequences $a_n = \omega(b_n)$ when there exist a positive constant $C$ and an integer $n_0$ such that $a_n \geq C b_n$ for all $n \geq n_0$ and the ratio $a_n/b_n \to \infty$.

### 2.1. Theoretical results

The main aim of this paper is to accurately describe the asymptotic behaviour of the UASE, which we do by establishing two key results. In each of these results, we allow some control over how regularly new connections may appear as the graphs in the MRDPG model grow by introducing a **sparsity factor** $\rho_n$. This is a real-valued function of $n$ which is either equal to 1 or tends to zero (corresponding to dense and sparse regimes respectively) such that the latent positions $X_i$ can each be written as $X_i = \rho_n^{1/2} \xi_i$, where the $\xi_i$ are identically distributed according to $F$. We will typically use the notation $F_\rho^n$ to represent this scaled distribution.

The first of our main results shows that, after a linear transformation, the latent positions $X_i$ are well-approximated by the rows of the left embedding (and those of the maximal rank right restricted embeddings), in the sense that the maximum error vanishes. This is stated using the two-to-infinity norm [6] of the associated error matrix, which is the maximum Euclidean norm of any of its rows.
Theorem 1 (Consistency in the two-to-infinity norm). Let $F$ be a distribution on a subset $\mathcal{X}$ of $\mathbb{R}^d$, and let $(A^{(1)}, \ldots, A^{(m)}, X) \sim \text{MRDPG}(F^n, \Lambda_1, \ldots, \Lambda_m)$. Then there exists a sequence of matrices $L_{(n)} \in \text{GL}(d)$ such that

$$\|X_A L_{(n)} - X\|_{2 \to \infty} = O \left( \frac{m^{13/4} \log^{3/2}(mn)}{n^{1/2} \rho_n} \right)$$

almost surely. Moreover, for each $r \in \{1, \ldots, m\}$ with $\text{rank}(\Lambda_r) = d$, there exists a sequence $L_{(n)}^{(r)} \in \text{GL}(d)$ such that

$$\|Y_{\Lambda_r} L_{(n)}^{(r)} - X\|_{2 \to \infty} = O \left( \frac{m^{13/4} \log^{3/2}(mn)}{n^{1/2} \rho_n} \right)$$

almost surely.

(We note this bound can be reduced by a factor of $m^{1/4}$ if every matrix $\Lambda_r$ is of maximal rank.)

The matrices $L_{(n)}$ and $L_{(n)}^{(r)}$ can be described explicitly, and are constructed by a two-step process; first using a modified Procrustes-style argument to simultaneously align $X_A$ with $X_P$ and $Y_A$ with $Y_P$ via an orthogonal transformation, and then applying a second linear transformation which maps $X_P$ or $Y_{P,r}$ directly to $X$. For the first step, let $U_P^T U_A + V_P^T V_A$ admit the singular value decomposition

$$U_P^T U_A + V_P^T V_A = W_1 \Sigma W_2^T,$$

and let $W = W_1 W_2^T$. The matrix $W$ solves the one mode orthogonal Procrustes problem

$$W = \arg\min_{Q \in O(d)} \|U_A - U_P Q\|_F^2 + \|V_A - V_P Q\|_F^2,$$

and we use $W^T$ to align $X_A$ with $X_P$.

For the second step, in Proposition 15 we construct matrices $L_X$ and $L_{Y,r} \in \text{GL}(d)$ which satisfy $X_P = X_L X$ and $Y_{P,r} = X L_{Y,r}$ (and consequently we find that $L_X L_{Y,r}^T = \Lambda_r$) whose inverses are shown in Proposition 16 to have bounded spectral norm. The matrices $L_{(n)}$ and $L_{(n)}^{(r)}$ are then given by $L_{(n)} = W^T L_1^{-1}$ and $L_{(n)}^{(r)} = W^T L_{Y,r}^{-1}$.

The second of our main results centres on the error distribution of a latent position estimate, and shows that conditional on the true position it is asymptotically Gaussian:

Theorem 2 (Central Limit Theorems). With the same notation as Theorem 1, let $\xi_1, \ldots, \xi_n, \xi \sim F$, $\Delta = \mathbb{E}[\xi \xi^T] \in \mathbb{R}^{d \times d}$, $\hat{\Delta} = I_m \otimes \Delta \in \mathbb{R}^{md \times md}$, and suppose that the sparsity factor $\rho_n$ satisfies $\rho_n = \omega \left( \frac{\log^{3/2}(n)}{n^{1/2}} \right)$. Given $x \in \mathcal{X}$ and $r \in \{1, \ldots, m\}$, let

$$\Sigma_r(x) = \begin{cases} 
\mathbb{E} \left[ x^T \Lambda_r \xi \left( 1 - x^T \Lambda_r \xi \right) \xi \xi^T \right] & \text{if } \rho_n = 1 \\
\mathbb{E} \left[ (x^T \Lambda_r \xi) (\xi \xi^T) \right] & \text{if } \rho_n \to 0
\end{cases}$$
and let $\hat{\Sigma}(x) = \text{diag}(\Sigma_1(x), \ldots, \Sigma_m(x))$. Then, for all $z \in \mathbb{R}^d$ and for any fixed $i$,

$$
P \left( n^{1/2} \left( X_{A(\hat{L}_n)} - X \right)^\top \leq z \mid \xi_i = x \right) \rightarrow \Phi \left( z, \Delta_A \hat{\Sigma}(x) \Delta_A^\top \right)
$$

almost surely and, if $\text{rank}(\Lambda_r) = d$,

$$
P \left( n^{1/2} \left( Y_{A,r}{\hat{L}_n} - X \right)^\top \leq z \mid \xi_i = x \right) \rightarrow \Phi \left( z, \Delta_r \Sigma_r(x) \Delta_r^\top \right),
$$

almost surely, where $\Delta_A = (A\Delta A^\top)^{-1}A$ and $\Delta_r = (A_r\Delta A_r^\top)^{-1}A_r$.

The theorems, of which single-graph analogs were derived in [31, 4, 23, 8, 6, 32, 28, 5], also have analogous methodological implications. Under a multilayer stochastic block model, discussed in Section 4, the left UASE asymptotically follows a Gaussian mixture model with non-circular components. Fitting this model is preferable to using $K$-means, which is implicitly fitting circular components. Apart from shape considerations, note that (as with the GRDPG) latent positions under the MRDPG are only identifiable up to a distance-distorting transformation (here invertible linear, there indefinite orthogonal) to which partitions obtained using a Gaussian mixture model are invariant but those obtained using $K$-means are not. Consistency in the two-to-infinity norm should imply the consistency of many subsequent statistical analyses: usually, if a method is consistent for data $X$, it is consistent under a perturbation of vanishing maximal error; one need only then worry about the effect of an unidentifiable linear transformation on conclusions; if the estimand is invariant, this effect will often vanish on account of the transformation having bounded spectral norm.

In the special case in which the graphs $A^{(1)}, \ldots, A^{(m)}$ are identically distributed, Theorems 1 and 2 can be made more precise:

**Corollary 3.** Let $F$ be a distribution on a subset $\mathcal{X}$ of $\mathbb{R}^d$, and let $(A^{(1)}, \ldots, A^{(m)}, X) \sim \text{MRDPG}(F^n, I_{p,q})$. Then there exists a sequence of matrices $\hat{L}_n \in \text{GL}(d)$ such that

$$
\|X_{A(\hat{L}_n)} - X\|_{2 \rightarrow \infty} = O \left( \frac{m^3 \log^{3/2} (mn)}{n^{1/2} \rho_n} \right)
$$

and, for each $r \in \{1, \ldots, m\}$,

$$
\|Y_{A,r}(\hat{L}_n) - X\|_{2 \rightarrow \infty} = O \left( \frac{m^3 \log^{3/2} (mn)}{n^{1/2} \rho_n} \right)
$$

almost surely, where $\hat{L}_n \hat{L}_n^\top = I_{p,q}$.

As was mentioned previously, a slightly tighter bound than Theorem 1 is possible as the matrix $I_{p,q}$ is of full rank (which forces the singular values of $P$ to satisfy $\sigma_1(P) = \omega(m^{1/2}n^{1/2})$).
Corollary 4. With the same notation as Corollary 3, let $\xi_1, \ldots, \xi_n, \xi \sim F$, $\Delta = \mathbb{E}[\xi\xi^T] \in \mathbb{R}^{d \times d}$, $\tilde{\Delta} = \mathbf{I}_m \otimes \Delta \in \mathbb{R}^{md \times md}$, and suppose that the sparsity factor $\rho_n$ satisfies $\rho_n = \omega \left( \frac{\log^{3/2}(n)}{n^{1/2}} \right)$. Given $x \in \mathcal{X}$, let

$$\Sigma(x) = \begin{cases} 
\mathbf{I}_{p,q} \Delta^{-1} \mathbb{E} \left[ x^T \mathbf{I}_{p,q} \xi \xi^T \right] \Delta^{-1} \mathbf{I}_{p,q} & \text{if } \rho_n = 1 \\
\mathbf{I}_{p,q} \Delta^{-1} \mathbb{E} \left[ (x^T \mathbf{I}_{p,q} \xi) \xi \xi^T \right] \Delta^{-1} \mathbf{I}_{p,q} & \text{if } \rho_n \to 0
\end{cases} \mathbf{I}_{p,q}$$

Then, for all $z \in \mathbb{R}^d$ and for any fixed $i$,

$$\mathbb{P} \left( n^{1/2} \left( X_{A,\tilde{L}(n)} - X \right)_{i}^T \leq z \mid \xi_i = x \right) \to \Phi \left( z, \frac{1}{m} \Sigma(x) \right)$$

and

$$\mathbb{P} \left( n^{1/2} \left( Y_{A,\tilde{L}(n)} - X \right)_{i}^T \leq z \mid \xi_i = x \right) \to \Phi \left( z, \Sigma(x) \right),$$

almost surely.

In this case the limiting distribution for the rows UASE are the same as that for the ASE stated in [28], scaled by a factor of $\frac{1}{m}$, and in particular coincides with that of ASE(mean) (see for example [34]).

In Corollaries 3 and 4 the matrices $\tilde{L}(n)$ are common across graphs, allowing a direct comparison of their right-sided embeddings. By contrast, independently embedded graphs first need to be aligned before a meaningful comparison is possible. In [33] this is achieved using Procrustes, the appropriate method of alignment under an RDPG model where latent positions are identifiable only up to orthogonal transformation. Under the GRDPG, finding a best indefinite orthogonal alignment is less straightforward. Computational issues aside, alignment comes at a statistical cost [21] when testing whether two point clouds differ statistically, and the empirical performance of using right-sided unfolded adjacency spectral embeddings for two-graph testing are investigated in Section 5.4.

3. Connection to the higher-order singular value decomposition

Any collection of matrices $M^{(1)}, \ldots, M^{(m)} \in \mathbb{R}^{n \times n}$ can instead be thought of as a 3-tensor $\mathcal{M} \in \mathbb{R}^{n \times n \times m}$, where $\mathcal{M}_{ijr} = M^{(r)}_{ij}$, and so it is natural to consider applying tensor-based techniques to study multiple graphs. Perhaps the most obvious candidate would be to use the higher order singular value decomposition (see [20]) a generalisation of the matrix singular value decomposition, which we shall briefly discuss.

Any 3-tensor $\mathcal{M} = (m_{ijk}) \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ has 3 matrix representations (known as unfoldings) $\mathcal{M}_{(1)} \in \mathbb{R}^{n_1 \times n_2 n_3}$, $\mathcal{M}_{(2)} \in \mathbb{R}^{n_2 \times n_1 n_3}$ and $\mathcal{M}_{(3)} \in \mathbb{R}^{n_3 \times n_1 n_2}$ which contain the element $m_{ijk}$ in position $(i, (j-1)n_3 + k), (j, (k-1)n_1 + i)$ and $(k, (i-1)n_2 + j)$ respectively.
To visualize this, consider a 3-tensor $\mathcal{M} \in \mathbb{R}^{3 \times 3 \times 2}$. Then the unfoldings defined above are

\[
\mathcal{M}(1) = \begin{pmatrix}
m_{111} & m_{112} & m_{121} & m_{122} & m_{131} & m_{132} \\
m_{211} & m_{212} & m_{221} & m_{222} & m_{231} & m_{232} \\
m_{311} & m_{312} & m_{321} & m_{322} & m_{331} & m_{332}
\end{pmatrix},
\]

\[
\mathcal{M}(2) = \begin{pmatrix}
m_{111} & m_{121} & m_{131} & m_{112} & m_{122} & m_{132} \\
m_{211} & m_{221} & m_{231} & m_{212} & m_{222} & m_{232} \\
m_{311} & m_{321} & m_{331} & m_{312} & m_{322} & m_{332}
\end{pmatrix},
\]

and

\[
\mathcal{M}(3) = \begin{pmatrix}
m_{111} & m_{121} & m_{131} & m_{112} & m_{122} & m_{132} \\
m_{211} & m_{221} & m_{231} & m_{212} & m_{222} & m_{232} \\
m_{311} & m_{321} & m_{331} & m_{312} & m_{322} & m_{332}
\end{pmatrix}.
\]

Given matrices $U \in \mathbb{R}^{m_1 \times n_1}$, $V \in \mathbb{R}^{m_2 \times n_2}$ and $W \in \mathbb{R}^{m_3 \times n_3}$, we define the 1-, 2- and 3-mode products $\mathcal{M} \times_1 U \in \mathbb{R}^{m_1 \times n_2 \times n_3}$, $\mathcal{M} \times_2 V \in \mathbb{R}^{n_1 \times m_2 \times n_3}$ and $\mathcal{M} \times_3 W \in \mathbb{R}^{n_1 \times n_2 \times m_3}$ by

\[
(\mathcal{M} \times_1 U)_{rjk} = \sum_{i=1}^{n_1} a_{ijk}u_{ri}, \quad (\mathcal{M} \times_2 V)_{isk} = \sum_{j=1}^{n_2} a_{ijk}v_{sj}, \quad (\mathcal{M} \times_3 W)_{ijt} = \sum_{k=1}^{n_3} a_{ijk}w_{tk}.
\]

Perhaps more intuitively, if we let $\mathcal{M}_{sjk} \in \mathbb{R}^{n_1}$ denote the vector obtained by fixing indices $j$ and $k$ (and similarly for other indices), then

\[
(\mathcal{M} \times_1 U)_{sjk} = U \cdot \mathcal{M}_{sjk}, \quad (\mathcal{M} \times_2 V)_{isk} = V \cdot \mathcal{M}_{isk}, \quad (\mathcal{M} \times_3 W)_{ijt} = W \cdot \mathcal{M}_{ijt}.
\]

The $n$-mode products satisfy the properties

\[
\mathcal{M} \times_m A \times_n B = (\mathcal{M} \times_m A) \times_n B = (\mathcal{M} \times_n B) \times_m A \text{ if } m \neq n
\]

and

\[
(\mathcal{M} \times_n A) \times_n B = \mathcal{M} \times_m BA,
\]

where $A$ and $B$ are appropriately sized matrices.

The higher order singular value decomposition is then given by the following result (see [20], Theorem 2):

**Theorem 5** (3rd order singular value decomposition). Every 3-tensor $\mathcal{M} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ can be written as a product

\[
\mathcal{M} = S \times_1 U^{(1)} \times_2 U^{(2)} \times_3 U^{(3)}
\]

in which

- $U^{(1)} \in \mathbb{O}(n_1), U^{(2)} \in \mathbb{O}(n_2)$ and $U^{(3)} \in \mathbb{O}(n_3)$; and
• $S \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is a tensor whose sub-tensors $S_{i * *}$ (obtained by fixing the index $i$) satisfy $\langle S_{1 * *}, S_{2 * *} \rangle_F = 0$ and $\|S_{1 * *}\|_F \geq \ldots \geq \|S_{n_1 * *}\|_F \geq 0$ (and similarly for the sub-tensors obtained by fixing $j$ and $k$).

The Frobenius norms $\|S_{i * *}\|_F$, symbolized by $\sigma_i^{(n)}$ are $n$-mode singular values of $M$ and the columns of the matrix $U^{(n)}$ are the $n$-mode singular vectors of $M$. The $n$-rank of $M$ is the highest index $i$ for which $\sigma_i^{(n)}$ is non-zero.

The higher order singular value decomposition is intrinsically linked with the matrix singular value decompositions of the unfoldings $M$; the $n$-mode singular values $\sigma_i^{(n)}$ are simply the singular values of $M_{(n)}$ (and thus the $n$-rank of $M$ is simply the rank of $M_{(n)}$) while the $n$-mode singular vectors are the left singular vectors of $M_{(n)}$.

Certain concepts relating to the matrix singular value decomposition extend readily to the tensor case; in particular, the notions of spectral embeddings and low-rank approximations. Let $\|M\|_F$ denote the Frobenius norm of $M$, defined by

$$\|M\|_F = \left( \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} m_{ijk}^2 \right)^{1/2},$$

which clearly satisfies $\|M\|_F = \|M_{(n)}\|_F$ for any $n$, and for any unit norm vector $x$, define the $n$-mode oriented energy of $M$ in the direction of $x$ to be $\|x^\top M_{(n)}\|_F^2$. Then ([20], Property 9) the directions of extremal $n$-mode oriented energy of $M$ correspond to the $n$-mode singular vectors, with oriented energy equal to the square of the corresponding $n$-mode singular value. Thus we have a natural extension of the concept of the spectral embedding of a matrix to tensors by defining (for any given rank $d$) the $n$-mode spectral embedding of $M$ to be the spectral embedding $X_{M_{(n)}}$ of $M_{(n)}$, the columns of which are therefore the $d$ directions of greatest energy among the $n$-mode vectors, scaled according to their contribution.

Similarly, just as the matrix singular value decomposition gives us a means to approximate a matrix by discarding small singular values, one can obtain a reasonable (although not necessarily the best) approximation to $M$ by discarding small $n$-mode singular values ([20], Property 10):

**Theorem 6.** Let the $n$-rank of $M$ be $r_n$ for $n = 1, 2, 3$, and choose integers $d_n < r_n$ for each $n$. Define a tensor $\widehat{M}$ by discarding the $n$-mode singular values $\sigma_{d_n+1}^{(n)}, \ldots, \sigma_{r_n}^{(n)}$ for each $n$ (that is, by setting the corresponding parts of $S$ to zero - equivalently we consider the restriction of $S$ to $\mathbb{R}^{d_1 \times d_2 \times d_3}$ and the corresponding restrictions of the matrices $U^{(n)}$ to $\mathcal{O}(n \times d_n)$). Then

$$\|M - \widehat{M}\|_F^2 \leq \sum_{i=d_1+1}^{r_1} (\sigma_i^{(1)})^2 + \sum_{j=d_2+1}^{r_2} (\sigma_j^{(2)})^2 + \sum_{k=d_3+1}^{r_3} (\sigma_k^{(3)})^2.$$ 

We argue that the MRDPG is a natural matrix-based method to study the higher order singular value decomposition. Define tensors $P \in [0,1]^{n \times n \times m}$, $A \in \{0,1\}^{n \times n \times m}$ and $L \in \mathbb{R}^{d \times d \times m}$ by $A_{ijr} = A_{ijr}^{(r)}$, $P_{ijr} = P_{ijr}^{(r)}$ and $L_{ijr} = (A_{r})_{ij}$. If $m = 1$ (that is, if we are working with the standard GRDPG) we can recover the latent positions $X$ from $P$ (up to an indefinite orthogonal
transformation) via the 1- and 2-mode spectral embeddings \( X_{\mathcal{P}(1)} = X_{\mathcal{P}(2)} = X_P \), and similarly the 1- and 2-mode spectral embeddings \( X_{\mathcal{A}(1)} = X_{\mathcal{A}(2)} = X_A \) approximate \( X \) up to an indefinite orthogonal transformation.

For general \( m \), therefore, it is natural to consider these same spectral embeddings. But due to the symmetry of the matrices \( P^{(r)} \), the second unfoldings \( \mathcal{A}(2) \) and \( \mathcal{P}(2) \) are precisely the matrices \( A \) and \( P \) given in the definition of the MRDPG, while the first unfoldings \( \mathcal{A}(1) \) and \( \mathcal{P}(1) \) are simply column permutations of \( A \) and \( P \). In particular, the 1- and 2-mode singular vectors (and similarly the 1- and 2-mode singular values) of \( A \) and \( P \) coincide, and thus so too do the 1- and 2-mode spectral embeddings, which are identical to the left embeddings \( X_A \) and \( X_P \) defined in the MRDPG. We note too that by Theorem 6 the tensor \( \hat{A} \) obtained by discarding the most significant 1- and 2-mode singular values \( \sigma_i^{(1)} \) and \( \sigma_i^{(2)} \) for \( i > d \) provides a good approximation to \( A \), and that \( \hat{A} \) in turn provides a reasonable approximation to \( P \), in the sense that \( \| \hat{A} - P \|_F \leq \sqrt{d} \| A - P \| = O(\sqrt{n \log(n)}) \).

We can use knowledge of the MRDPG to give information about the higher order singular value decomposition of the tensor \( P \). To begin with, define a matrix \( F \in \mathbb{R}^{d \times d} \) by \( F_{ij} = (P^{(i)}, P^{(j)})_F \), and note that \( F = P^{(3)} P^{(3)}_F \), so that in particular the 3-mode singular vectors of \( P \) are given by the columns of \( U_F \), where \( F \) admits the singular value decomposition \( F = U_F \Sigma_F V_F^\top \). Note that the rank of \( F \) (and thus the 3-rank of \( P \)) is the dimension of the subspace of \( \mathbb{R}^{d^2} \) spanned by the \( P^{(r)} \), which is equal to the dimension of the subspace of \( \mathbb{R}^{d^2} \) spanned by the \( \Lambda_r \). Since each matrix \( \Lambda_r \) is symmetric, we see that:

**Corollary 7.** The 3-rank of the tensor \( P \) defined above is at most \( \frac{d(d+1)}{2} \).

Note that we have the tensor identity \( P = \mathcal{L} \times_1 X \times_2 X \times_3 X \), where \( \mathcal{L} \in \mathbb{R}^{d \times d \times m} \) is defined by \( \mathcal{L}_{ijr} = (\Lambda_r)_{ij} \) (and so we can think of \( P \) as a core tensor \( \mathcal{L} \) whose 1- and 2-spaces are transformed simultaneously by \( X \)). Also, some manipulation of the terms in the 3rd order singular value decomposition yields:

\[
\mathcal{P} = S \times_1 U_P \times_2 U_P \times_3 U_F \\
= S \times_1 X_P \Sigma_P^{-1/2} \times_2 X_P \Sigma_P^{-1/2} \times_3 U_F \\
= \left( S \times_1 \Sigma_P^{-1/2} \times_2 \Sigma_P^{-1/2} \times_3 U_F \right) \times_1 X_P \times_2 X_P,
\]

(note that, since \( P \) has 1- and 2-rank \( d \), the core tensor \( S \in \mathbb{R}^{n \times n \times m} \) can be considered to belong to \( \mathbb{R}^{d \times d \times m} \) by setting \( S_{ijr} = 0 \) if \( i > d \) or \( j > d \). By Theorem 6 this does not change the tensor \( \mathcal{P} \)).

Noting that \( X_P = XL_X \) for some matrix \( L_X \in \text{GL}(d) \), we see that

\[
\mathcal{P} = \left( S \times_1 \Sigma_P^{-1/2} \times_2 \Sigma_P^{-1/2} \times_3 U_F \right) \times_1 XL_X \times_2 XL_X \\
= \left( S \times_1 L_X \Sigma_P^{-1/2} \times_2 L_X \Sigma_P^{-1/2} \times_3 U_F \right) \times_1 X \times_2 X,
\]

and so (by taking the 1- and 2-mode products with \( (X^\top X)^{-1} X \), which is well-defined as \( X \) is of rank \( d \) for \( n \) large enough) we obtain the following result:
Corollary 8. The tensor $\mathcal{P}$ admits a 3rd-order singular value decomposition

$$\mathcal{P} = S \times_1 U_1 \times_2 U_2 \times_3 U_3,$$

where the tensor $S \in \mathbb{R}^{d \times d \times m}$ is given by

$$S = L \times_1 L^{-1} \Sigma^{1/2}_1 \times_2 L^{-1} \Sigma^{1/2}_2 \times_3 U_3^\top.$$

4. Examples

4.1. The multilayer stochastic blockmodel

We say that undirected graphs $G^{(1)}, \ldots, G^{(m)}$ with adjacency matrices $A^{(1)}, \ldots, A^{(m)} \in \{0, 1\}^{n \times n}$ follow a $K$-community multilayer stochastic blockmodel [16] if there is a common set of $n$ vertices and a partition of these vertices into $K$ disjoint communities (with a vertex belonging to the $i$th community with probability $\pi_i$), together with a set of $K$ symmetric matrices $B^{(1)}, \ldots, B^{(K)} \in [0, 1]^{K \times K}$ such that, conditional on our partition, $A^{(r)}_{ij} \sim \text{Bern} \left( B^{(r)}_{z_i z_j} \right)$, where $z_i \in \{1, \ldots, K\}$ is an index denoting the community membership of the $i$th vertex.

Just as the GRDPG can model a $K$-community stochastic blockmodel [28], the MRDPG can model a $K$-community multilayer stochastic blockmodel, with the only significant difference being the lack of a canonical choice of latent positions $X$. In practice, we use one of the following choices:

- We can draw the $X_i$ from the set of standard basis vectors of $\mathbb{R}^K$, so that $X$ is simply the community membership matrix for the stochastic blockmodel, and $\Lambda_r = B^{(r)}$.

- Let $B = U \Sigma V^\top$ be the singular value decomposition of $B = [B^{(1)}] \cdots [B^{(m)}]$ and let $v_i$ be the $i$th row of $U \Sigma^{1/2}$. Then we can draw the $X_i$ from the set $\{v_1, \ldots, v_K\}$, in which case $\Lambda_r = U \Sigma^{-1/2} B^{(r)} \Sigma^{-1/2} U^\top$. This choice reduces to the model given in [28] when $m = 1$.

We now consider two examples. For the first example, we take pair of graphs with adjacency matrices $A^{(1)}$ and $A^{(2)}$ generated according to a multilayer stochastic blockmodel with parameters

$$B^{(i)} = \begin{pmatrix} 0.42 & 0.42 \\ 0.42 & 0.5 \end{pmatrix}, \quad \pi = (0.6, 0.4)$$

(the Laplacian spectral embedding of a single graph generated with these parameters was studied in [32]). Figure 1 plots the estimated latent positions for the ASE of the matrix $A^{(1)}$ (first row) and the UASE (second row) for $n = 1000, 2000$ and $4000$. Also displayed are the 95% level curves of the empirical distributions (dashed curves) and the theoretical distributions specified by Theorem 2 (solid curves).

For the second example, we take pair of graphs with adjacency matrices $A^{(1)}$ and $A^{(2)}$ generated according to a multilayer stochastic blockmodel with parameters

$$B^{(1)} = \begin{pmatrix} 0.58 & 0.58 \\ 0.58 & 0.5 \end{pmatrix}, \quad B^{(2)} = \begin{pmatrix} 0.42 & 0.42 \\ 0.42 & 0.5 \end{pmatrix}, \quad \pi = (0.6, 0.4).$$
Since the matrices $B^{(1)}$ and $B^{(2)}$ have signatures $(2, 0)$ and $(1, 1)$ respectively, they exhibit markedly different assortativity behaviours. Figure 2 plots the estimated latent positions for the ASE of the matrix $A^{(1)}$ (first row) and the UASE (second row) for $n = 1000, 2000$ and $4000$, with the 95% level curves displayed as in 1.

In both examples, the UASE demonstrates greater cluster separation over the ASE. To test this empirically, each experiment was repeated 100 times for $n \in \{250, 500, 750, \ldots, 2000\}$ and nodes assigned to the most likely cluster predicted by the Gaussian mixture model obtained via the MCLUST algorithm (see [30]). These were then compared against the known cluster assignments given by the latent positions $X_i$, and the average error rate calculated across all samples of a given size $n$ (for the ASE, the average error rate of the two embeddings was used). Figure 3 displays these error rates in the case of the identical and mixed parameter examples, as well as the average error rates for the mean embedding ASE(mean) - that is, the spectral embedding of the matrix $\bar{A} = \frac{1}{2}(A^{(1)} + A^{(2)})$ - and, in the identically distributed case, the omnibus embedding OMNI (see [21]).

As expected, the UASE (black line) clearly outperforms the ASE (red line) in both cases. In the first case, it is comparable with both ASE(mean) and OMNI; slightly outperforming OMNI (green line) while in turn being slightly outperformed by ASE(mean) (blue line). When the two graphs exhibit different assortativity behaviours, however, ASE(mean) lags significantly behind both the
ASE and the UASE, plateauing at an error rate of roughly 42\%, indicating a clear improvement in the UASE over ASE(mean) at solving community detection problems when dealing with multiple graphs with a suspected mixture of assortativity behaviours. This is not particularly surprising: if the adjacency matrices of different graphs have different signatures, then it is entirely possible the matrix $\mathbf{P}$ has non-maximal rank, causing some of the information in the system to be lost when we spectrally embed the matrix $\mathbf{A}$. Conversely, one finds that the embedding $\mathbf{X}_A$ is the same as that of the positive-definite square root of the matrix $\sum_{r=1}^{m}(\mathbf{A}^{(r)})^2$, which will always be of maximal rank.

4.2. Non-maximal rank stochastic blockmodels

One advantage of studying the MRDPG is that we do not require the matrices $\mathbf{A}_r$ are to have maximal rank; this can lead to situations in which information about latent positions is obscured in individual graphs, but becomes apparent when considering the joint embedding. As an example, consider a network which can be modeled at any given time $t$ as a two-community stochastic blockmodel with matrix of probabilities $\mathbf{B}_t = \begin{pmatrix} p(t) & q(t) \\ q(t) & r(t) \end{pmatrix}$, where $p, q, r : \mathbb{R}_{\geq 0} \to [0, 1]$, and suppose that at some time $T$ a number of members move from one community to the other, updating their communication preferences accordingly.

This situation can be modeled by considering these members as a independent third community;
if we view the networks at times $T_1$ and $T_2$ with $T_1 < T < T_2$ then the pre- and post-movement probability matrices

$$B_1 = \begin{pmatrix} p_1 & p_1 & q_1 \\ p_1 & p_1 & q_1 \\ q_1 & q_1 & r_1 \end{pmatrix}, \quad B_2 = \begin{pmatrix} p_2 & q_2 & q_2 \\ q_2 & r_2 & r_2 \\ q_2 & r_2 & r_2 \end{pmatrix}$$

(where for clarity we suppress $T$ from our notation) are both of non-maximal rank, but $B = [B_1|B_2]$ has maximal rank, provided that the $p_i, q_i$ and $r_i$ are distinct.

We demonstrate this with an example. Let $n = 4000$ and suppose that we have two fixed communities, each containing 1750 nodes, with the remaining 500 nodes representing members who move between communities. Let the probability matrices be given by

$$B_{T_1} = \begin{pmatrix} 0.47 & 0.39 \\ 0.39 & 0.56 \end{pmatrix}, \quad B_{T_2} = \begin{pmatrix} 0.53 & 0.61 \\ 0.61 & 0.44 \end{pmatrix}.$$
5. Experimental data

5.1. Recovery of latent positions

An important estimation problem for the data of a random dot product graph is that of estimating the latent positions \( \mathbf{X}_i \), and so we shall investigate the performance of the MRDPG in the context of such an estimation problem. For comparison, we will consider the multiple adjacency spectral embedding (MASE) \[3\], which is an alternative method of jointly embedding adjacency matrices which follow a model that is essentially identical to the MRDPG, known as the common subspace independent edge graph model. In \[3\], the authors demonstrate that the MASE yields state-of-the-art performance on subsequent inference tasks, ahead of other competing models for studying multiple graph embeddings such as the multi-RDPG \[25\] and MREG \[38\] models, making it an ideal method to compare the UASE against.

**Definition 3.** (Common Subspace Independent Edge graphs (COSIE)).

Let \( \mathbf{U} = \{\mathbf{U}_i\}^{n}_{i=1} \in \mathbb{R}^{n \times d} \) be a matrix with orthonormal columns, and \( \mathbf{R}^{(1)}, \ldots, \mathbf{R}^{(m)} \in \mathbb{R}^{d \times d} \) be symmetric matrices such that \( \mathbf{U}_i^\top \mathbf{R}^{(r)} \mathbf{U}_j \in [0, 1] \) for all \( i, j \in \{1, \ldots, n\} \) and \( r \in \{1, \ldots, m\} \). The random adjacency matrices \( \mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)} \) are said to be jointly distributed according to the **common subspace independent-edge graph model** with bounded rank \( d \) and parameters \( \mathbf{U} \) and \( \mathbf{R}^{(1)}, \ldots, \mathbf{R}^{(m)} \) if for each \( r \in \{1, \ldots, m\} \), conditional upon \( \mathbf{U} \) and \( \mathbf{R}^{(r)} \) we have \( \mathbf{A}^{(r)}_{ij} \sim \text{Bern} \left( \mathbf{P}^{(r)}_{ij} \right) \), where \( \mathbf{P}^{(r)} = \mathbf{U} \mathbf{R}^{(r)} \mathbf{U}^\top \), in which case we write \( \langle \mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)} \rangle \sim \text{COSIE}(\mathbf{U}; \mathbf{R}^{(1)}, \ldots, \mathbf{R}^{(m)}) \).

For all intents and purposes, the COSIE and MRDPG models are equivalent. Any COSIE model gives rise to a MRDPG by simply setting the latent positions \( \mathbf{X}_i \) to be equal to the rows \( \mathbf{U}_i \), and the matrices \( \mathbf{A}_r = \mathbf{R}^{(r)} \) for each \( r \). Conversely, given a MRDPG such that the matrix \( \mathbf{X} \) of latent positions is of rank \( d \), we can define \( \mathbf{U} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1/2} \), where we have taken the positive-definite matrix square root of the matrix \( \mathbf{X}^\top \mathbf{X} \). It is clear that the columns of \( \mathbf{U} \) are orthonormal, and we obtain a COSIE model by setting \( \mathbf{R}^{(r)} = (\mathbf{X}^\top \mathbf{X})^{1/2} \mathbf{A}_r (\mathbf{X}^\top \mathbf{X})^{1/2} \). In both cases the two definitions of the matrices \( \mathbf{P}^{(r)} \) coincide.

**Definition 4.** (Multiple adjacency spectral embedding (MASE)).
Let \((A^{(1)}, \ldots, A^{(m)}) \sim \text{COSIE}(U; R^{(1)}, \ldots, R^{(m)})\). For each \(r \in \{1, \ldots, m\}\) let \(\text{rank}(R^{(r)}) = d_r\), let \(X^{(r)}_A \in \mathbb{R}^{n \times d_r}\) be the adjacency spectral embedding of \(A^{(r)}\) and define the \((n \times \sum_{r=1}^{m} d_r)\) matrix of concatenated spectral embeddings \(\tilde{M}_A = [X^{(1)}_A \cdots X^{(m)}_A]\). The multiple adjacency spectral embedding of \((A^{(1)}, \ldots, A^{(m)})\) is the matrix \(\tilde{U}_A \in \mathbb{R}^{n \times d}\) containing the \(d\) leading left singular vectors of \(\tilde{M}_A\).

We note that it is possible to use the MASE to produce estimates for the latent positions in a MRDPG in an analogous way to the method used for the UASE. Let \(\tilde{X}^i_A = \tilde{U}^i_A \tilde{\Sigma}^i_A\), where \(\tilde{\Sigma}_A \in \mathbb{R}^{d \times d}\) is the diagonal matrix of the leading \(d\) singular vectors of \(\tilde{M}_A\), and define \(\tilde{M}_P, \tilde{U}_P\) and \(\tilde{X}_P\) analogously for the matrices \(P^{(r)}\). Firstly, we note that adding columns of zeros to any of the \(X_P^{(r)}\) will not alter \(\tilde{X}_P\), and so we may assume without loss of generality that the matrices \(X_P^{(r)} \in \mathbb{R}^{n \times d}\), and thus that there exist matrices \(\hat{L}^{(r)} \in \mathbb{R}^{d \times d}\) of rank \(d_r\) such that \(X_P^{(r)} = \tilde{X} \hat{L}^{(r)}\). A similar argument to Proposition 15 then establishes the existence of a matrix \(\hat{L} \in \text{GL}(d)\) such that \(\tilde{X}_P = \hat{X} \hat{L}\). Performing a Procrustes-style alignment between \(\tilde{X}_A\) and \(\tilde{X}_P\) and then multiplying by \(\hat{L}^{-1}\) produces a set of points that in practice are a good approximation to the latent positions \(X_i\).

We first tested the performance of the two embeddings on graphs of different sizes by performing, for each value of \(n \in \{10, 25, 50, 75, 100, 250, 500, 750, 1000\}\), 1000 independent trials in which the latent positions \(X_i\) are drawn i.i.d. from a Dirichlet distribution with parameter \((1,1,1)^\top \in \mathbb{R}^3\). In each trial, we generate two graphs \(A^{(1)}, A^{(2)} \in \mathbb{R}^{n \times n}\), where \(A^{(r)}_{ij} \sim \text{Bern}(X_i^\top \Lambda_r X_j)\) for \(i < j\), where each \(\Lambda_r\) is a randomly chosen matrix. We then calculate estimates \(\hat{X}\) using the UASE and MASE as described previously, and compare the average mean squared error \(\frac{1}{n} \sum_{i=1}^{n} |\hat{X}_i - X_i|^2\) of the two embeddings across the 1000 trials.

We then investigated the effect of changing the number of graphs to be embedded on the accuracy of each embedding type. Fixing \(n = 750\), we again performed 1000 independent trials as above for \(m = 2, \ldots, 10\) embeddings, using the same procedure for generating the latent positions and adjacency matrices, and again compared the average mean squared error between the estimated and actual latent positions.

Figure 5 plots the results of the two experiments. While the MASE outperforms the UASE for values of \(n < 75\) (with the joint embedding performing significantly worse for \(n < 50\)) the UASE clearly demonstrates superior accuracy as the size of the graph grows, a trend which continues as we increase the number of graphs to be embedded.

### 5.2. Estimation of invariant subspaces

We next investigate the performance of UASE at estimating the invariant subspace \(U\) in the COSIE model. We do this by setting the matrix \(X\) of latent positions in the MRDG to be equal to \(U\), and considering the unscaled UASE, \(U_A\). Unlike the scaled embedding, which approximates the latent positions only up to linear transformation, the unscaled UASE approximates the invariant subspace \(U\) up to orthogonal transformation. Indeed, from our results for the scaled embedding the matrix \(U_P = U Q_X\) for some \(Q_X \in \text{GL}(d)\), whence the requirement that both \(U\) and \(U_P\) have
orthonormal columns forces $Q_X$ to in fact belong to $O(d)$, while the transformation applied in the Procrustes alignment between $U_A$ and $U_P$ is by definition orthogonal.

We can measure the distance between the estimate $U_A$ and the true invariant subspace $U$ using the spectral norm of the difference between the projections $\|U_A U_A^\top - U U^\top\|$ (and similarly for the estimate $\hat{U}$ produced by the MASE). This distance is zero only when there exists an orthogonal matrix $W \in O(d)$ such that $U_A = U W$ (respectively $\hat{U} = U W$).

As in the previous example, we investigated the effect of changing both the size of the graphs and the number of graphs to be embedded on the performance of the UASE and MASE. Again, we began by performing 1000 independent trials for each value of $n \in \{10, 25, 50, 75, 100, 250, 500, 750, 1000\}$, but this time the adjacency matrices $A^{(1)}$ and $A^{(2)}$ were distributed according to a 3-community multilayer stochastic blockmodel, where the matrices $B^{(r)}$ were randomly chosen, and vertices assigned to a community uniformly at random, discarding any trials for which the matrix $X$ of community assignments was not of full rank. We then calculated and compared the average of the subspace distances $\|U_A U_A^\top - U U^\top\|$ and $\|\hat{U}^\top - U U^\top\|$ across each of the 1000 trials. For the second experiment, we again fixed $n = 750$, performed 1000 independent trials as above for $m = 2, \ldots, 10$ embeddings, and compared the subspace distance between the estimated and actual invariant subspaces.

Figure 6 plots the results of the two experiments. For this task, although the performance of the two embedding types is almost indistinguishable for very small graphs, as the number of vertices grows the UASE consistently outstrips the MASE. As in the previous example, increasing the number of
5.3. Model estimation

As a final comparison of the UASE and MASE methods, we investigate the efficiency of both at the task of estimating the underlying matrices $P^{(r)}$ in the MRDPG and COSIE models, which is of particular practical interest for link prediction tasks. To establish an appropriate estimate, we first consider the case of the standard GRDPG (that is, when $m = 1$). In this case, an estimate $\hat{P}$ for the matrix $P$ can be obtained by setting $\hat{P} = X_A L_{p,q} X_A^\top$. Note that due to orthogonality of the singular vectors, the matrix $X_A \in \mathbb{R}^{n \times d}$ of the leading $d$ left singular vectors of $A$ is the projection of the full matrix of left singular vectors onto the $d$-dimensional subspace spanned by $U_A$. Since this projection corresponds to left multiplication by the matrix $U_A U_A^\top$, we have the alternative description $\hat{P} = U_A U_A^\top A U_A U_A^\top$.

Returning to the general case, we obtain for each $r \in \{1, \ldots, m\}$ an estimate $\hat{P}^{(r)} = U_A U_A^\top A^{(r)} U_A U_A^\top$ for the matrix $P^{(r)}$ using the unscaled UASE. For the MASE, we use the matrix $\hat{U} \hat{U}^\top A^{(r)} \hat{U} \hat{U}^\top$ as our estimate. For each of the trials in the previous example, we calculated these estimates, and measured the model estimation error in each case using the normalised mean squared error

$$\frac{\|\hat{P}^{(r)} - P^{(r)}\|_F}{\|P^{(r)}\|_F}.$$
Figure 7 plots the results of the two experiments, in which we see that once again the UASE consistently demonstrates greater accuracy than the MASE for all but the smallest of graphs, and for all numbers of embedded graphs.

5.4. Two-graph hypothesis testing

When $A^{(1)}, \ldots, A^{(m)}$ are identically distributed, the right embeddings $Y_{A,r}$ are identically distributed too and each subject to the same unidentifiable linear transformation (Corollaries 3 and 4). It is therefore natural to consider the effectiveness of the UASE at testing the semiparametric hypothesis that two observed graphs are drawn from the same underlying latent positions. This problem was considered for the omnibus embedding in [21], and we shall use the framework established there to test the UASE. Suppose, then, that we have points $X_1, \ldots, X_n, Y_1, \ldots, Y_n \in \mathbb{R}^d$, and that we have two graphs $G_1$ and $G_2$ whose adjacency matrices $A^{(1)}$ and $A^{(2)}$ satisfy $A^{(1)}_{ij} \sim \text{Bern}(X_i^\top I_{p,q} X_j)$ and $A^{(2)}_{ij} \sim \text{Bern}(Y_i^\top I_{p,q} Y_j)$. The UASE allows us to test the hypothesis:

$$H_0 : X_i = Y_i \forall i \in \{1, \ldots, n\}$$

by comparing the right embeddings $Y_{A,1}$ and $Y_{A,2}$. If $H_0$ holds, then the $Y_{A,i}$ are identically (although not independently) distributed, whereas if $H_0$ fails to hold then for some $k$ the $k$th row of $Y_{A,1}$ and $Y_{A,2}$ should be distributionally distinct.

The setup used in [21] to test this hypothesis, which we shall repeat here, is as follows: we begin by drawing $X_1, \ldots, X_n, Z_1, \ldots, Z_n \in \mathbb{R}^3$ identically according to a Dirichlet distribution with parameter $\alpha = (1,1,1)^\top$, select a subset $I$ of fixed size $k$ uniformly at random among all such subsets
of \{1, \ldots, n\}, and define

\[ Y_i = \begin{cases} Z_i & \text{if } i \in I \\ X_i & \text{otherwise} \end{cases} \]

We generate two graphs \( G_1 \) and \( G_2 \) with adjacency matrices \( A^{(1)} \) and \( A^{(2)} \) satisfying \( A_{ij}^{(1)} \sim \text{Bern}(X_i^T X_j) \) and \( A_{ij}^{(2)} \sim \text{Bern}(Y_i^T Y_j) \), and estimate the latent positions \( \hat{X} \) and \( \hat{Y} \) of the two graphs by using the right embeddings as described above, and (in the case of OMNI) the first and last \( n \) rows of the spectral embedding of the matrix

\[ M = \left( \frac{1}{2} (A^{(1)} + A^{(2)}) \right) \]

We note that this is only possible due to our prior knowledge of the matrix \( P \), which allows us to construct the required transformations.

In both cases we use the test statistic \( T = \sum_{i=1}^{n} \| \hat{X}_i - \hat{Y}_i \|_2^2 \); and accept or reject based on an estimate of the critical value of \( T \) under the null hypothesis obtained by using 2000 Monte Carlo iterates to estimate the distribution of \( T \).

Figure 8: Empirical power of the UASE (black) and OMNI (red) tests to detect when the two graphs being tested differ in the specified number of their latent positions. Each point is the proportion of 2000 trials for which the given technique correctly rejected the null hypothesis.

Figure 8 shows the power of the two approaches for testing the null hypothesis for different sized graphs and for different numbers of altered latent positions, by calculating the proportion (out of
2000 trials conducted for each sized graph) of trials for which we correctly reject the null hypothesis. For smaller graphs, OMNI is still the most effective method (although there is not much difference between the two where only one latent position is altered - however in this case the empirical power of both methods does not exceed 0.25). For larger graphs, particularly those with more than 500 vertices, the two methods are almost indistinguishable, and in such cases the UASE might be preferred based on size considerations, due to only requiring an $n \times mn$ rather than an $mn \times mn$ matrix.

6. Real data

6.1. Example: Dynamic link prediction on a computer network

The Los Alamos National Laboratory computer network [36] was studied in [28], in which it was demonstrated empirically that the disassortative connectivity behaviour inherent in the network leads to the GRDPG offering a marked modelling improvement over the RDPG in the task of out-of-sample link prediction between computers in the network. For a large-scale dynamic network such as this, the MRDPG offers the possibility of further refinement by allowing us to consider multiple “snapshots” of communication behaviour at different points in time simultaneously.

As an example, we extract a ten minute sample at random from the “Network Event Data” dataset, which we divide into two separate five minute samples. From the first sample we generate five graphs, each one describing the communication behaviour of the computers in the network over a period of one minute, by assigning each IP address to a node (with this assignment being kept consistent across all graphs), and recording an edge between two nodes if the corresponding edges are observed to communicate at least once within this period, and then construct the corresponding adjacency matrices $A^{(r)}$. Setting our embedding dimension $d = 10$ (which was an arbitrary choice) we then generate estimates $\hat{P}^{(r)}$ for the probability matrices as described in Section 5.3. We then use the average of these matrices to give us estimates of the probabilities of a link being generated between any given pair of computers.

In a similar manner, we generate estimates of the link probabilities for the mean adjacency matrix $\bar{A}$. We also construct adjacency matrices $A^{[1]}$ and $A^{[5]}$ from the connectivity graphs for the first minute and the full five minute period respectively (both of which follow a standard GRDPG) and generate link probability estimates accordingly.

Using these estimates, we attempt to predict which new edges will occur within the second five minute window, disregarding those involving new nodes. Figure 9 shows the receiver operating characteristic (ROC) curves for each model and for each port, where we treat the prediction task as a binary classification problem whose outcomes are either the presence or absence of an edge between nodes, which we predict by thresholding the estimated link probabilities. As one would expect, using the ASE of the graph corresponding to only a single minute of communication (the blue curve) produces the least accurate predictions, but the UASE (black), ASE(mean) (red) and the ASE for a five minute sample (green) all produce similar results, with the UASE slightly outperforming the other two methods. This can be confirmed numerically by calculating the area under each ROC curve (AUC) which is equivalent to the probability that a given classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [11]. We
calculate that the UASE has an AUC of 0.9734, which is a small improvement over ASE(mean) and the 5-minute ASE, which have AUC values of 0.9627 and 0.9635 respectively (the 1-minute ASE, by contrast, yields an AUC of 0.8767).

6.2. Example: Port-specific link prediction on a computer network

The LANL network data presents the opportunity to demonstrate another significant improvement offered by the MRDPG over the GRDPG, namely its ability to integrate data from sources which do not necessarily behave similarly. Each communication within in LANL network occurs through a specified port, which indicates the type of service being used, and it is natural to expect that different services may exhibit different communication behaviours.

We consider the first five minute sample from the previous section. During the first minute alone, there are a total of 121,737 (not necessarily unique) communications between 10,762 computers, with 4,379 different ports being used. Of these, the 8 most commonly-used ports account for over 75% of communications, and Table 1 lists these, together with the purpose to which each port is assigned.
Figure 10: Visualization of adjacency matrices showing connections between computers on the Los Alamos National Laboratory computer network during 1 minute of activity. The top-left image shows all connections during this time, while the remaining images show only connections via the specified port.
For each of these 8 ports, we generate a graph of the communications made between computers within the network through this specific port over the first minute of our sample as in the previous example. Figure 10 visualizes the adjacency matrix of each of these graphs as a 2-dimensional plot, together with the adjacency matrix of the full network graph.

| Port | Purpose                                | Proportion of traffic |
|------|----------------------------------------|-----------------------|
| 53   | DNS                                     | 27.7%                 |
| 443  | HTTPS                                  | 14.8%                 |
| 80   | HTTP                                   | 11.9%                 |
| 514  | Syslog                                 | 7.2%                  |
| 389  | Lightweight Directory Access Protocol  | 4.3%                  |
| 427  | Service Location Protocol              | 4.1%                  |
| 88   | Kerberos authentication system         | 3.4%                  |
| 445  | Microsoft-DS Active Directory          | 1.8%                  |

Table 1: Table showing the purpose and proportion of traffic utilizing the 8 most frequently used ports during 1 minute of activity on the Los Alamos National Laboratory computer network.

As before, we calculate estimates of the link probabilities for each port using the UASE, but now rather than averaging them we consider each port individually. For comparison, we also estimate link probabilities for each individual port using the corresponding GRDPG, and then use both estimates to attempt to predict which new edges will occur within the remainder of our five-minute window. Figure 11 shows the ROC curves for each model and for each port, while Table 2 gives the AUC values for each curve. We note that for Port 53 (the busiest port) the standard ASE actually outperforms the UASE, but for every other port the UASE is the superior method, offering a significant improvement over the ASE for the less active ports.

| Embedding method | Port 53 | Port 443 | Port 80 | Port 514 | Port 389 | Port 427 | Port 88 | Port 445 |
|------------------|---------|----------|---------|----------|----------|----------|---------|----------|
| UASE             | 0.8391  | 0.7860   | 0.9010  | 0.8931   | 0.8534   | 0.8580   | 0.8949  | 0.9836   |
| ASE              | 0.8568  | 0.6370   | 0.7185  | 0.7806   | 0.6532   | 0.6586   | 0.5668  | 0.5720   |

Table 2: Table displaying the area under the curve (AUC) for the ROC curves in Figure 11.

7. Chernoff information and the multilayer stochastic blockmodel

7.1. Chernoff information

A reasonable question to ask is whether there is any benefit to studying the spectral embedding of the joint matrix $A$ as opposed to the individual matrices $A^{(r)}$, and how one might quantify this. Tang and Priebe [32], in the context of comparing the performance of the spectral embeddings of the Laplacian and adjacency matrix in recovering block assignments from a stochastic blockmodel graph, proposed using the Chernoff information [14] of the limiting Gaussian distributions obtained from the Central Limit Theorem associated to each embedding as a means of doing so. In a two-cluster problem, the Chernoff information is the exponential rate at which the Bayes error (from the decision rule which assigns each data point to its most likely cluster a posteriori) decreases asymptotically. The Chernoff information is an example of a $f$-divergence [2], [9] and therefore possesses the desirable attribute of being invariant under invertible linear transformations [22].
If $F_1$ and $F_2$ are two absolutely continuous multivariate distributions supported on $\Omega \subset \mathbb{R}^d$, with density functions $f_1$ and $f_2$ respectively, the Chernoff information between $F_1$ and $F_2$ is defined by $C(F_1, F_2) = \sup_{t \in (0, 1)} C_t(F_1, F_2)$ [14] where the Chernoff divergence

$$C_t(F_1, F_2) = -\log \left( \int_{\Omega} f_1(x)^t f_2(x)^{1-t} dx \right).$$
For a $K$-cluster problem, in which we have distributions $F_1, \ldots, F_K$ with corresponding density functions $f_1, \ldots, f_K$, we consider the Chernoff information of the critical pair $\min_{i \neq j} C(F_i, F_j)$.

We are interested in the special case in which the $F_i$ are multivariate normal distributions, in which case it is known (see [26]) that the Chernoff information can be expressed as $C(F_i, F_j) = \sup_{t \in (0,1)} C_t(F_i, F_j)$, where

$$C_t(F_i, F_j) = \left( \frac{t(1-t)}{2} \left( \mathbf{x}_i - \mathbf{x}_j \right) \Sigma_t^{-1} \left( \mathbf{x}_i - \mathbf{x}_j \right) + \frac{1}{2} \log \left( \frac{\Sigma_t}{\Sigma_t^{|\Sigma_j|^{1-t}}} \right) \right),$$

where $F_i \sim \mathcal{N}(\mathbf{x}_i, \Sigma_i)$ and $\Sigma_t = t\Sigma_1 + (1-t)\Sigma_2$.

The invariance of the Chernoff information under invertible linear transformations means that it is an ideal measure to use when studying the MRDPG, as it means that no information is lost by studying the joint spectral embedding $\mathbf{X}_A$ (rather than transforming the data, which requires knowledge - that we will not typically possess - of the underlying matrices of probabilities $\mathbf{P}^{(r)}$) and also that the Chernoff information is independent of our choice of latent positions $\mathbf{X}_i$.

### 7.2. Comparison of the MRDPG and GRDPG via Chernoff information

Suppose that $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}$ are drawn according to a multilayer stochastic blockmodel (which we recall means that $(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}, \mathbf{X}) \sim \text{MRDPG}(F, \Lambda_1, \ldots, \Lambda_m)$ where $F$ draws from a set of vectors $(\mathbf{v}_1, \ldots, \mathbf{v}_K)$ with probabilities $(\pi_1, \ldots, \pi_K)$), and define the Chernoff information of the model as

$$C_A = \min_{i \neq j} \left( \mathcal{N}(\mathbf{v}_i, \Delta_A \hat{\Sigma}(\mathbf{v}_i) \Delta_A^\top), \mathcal{N}(\mathbf{v}_j, \Delta_A \hat{\Sigma}(\mathbf{v}_j) \Delta_A^\top) \right),$$

where the covariance matrices $\Delta_A \hat{\Sigma}(\mathbf{v}_i) \Delta_A^\top$ are as specified in Theorem 2 (in particular, we recall that $\hat{\Sigma}(\mathbf{v}_i) = \text{diag}(\Sigma_1(v_i), \ldots, \Sigma_m(v_i))$, where the $\Sigma_r(v_i)$ are the covariance matrices for the individual embeddings). For the remainder of this section, we will operate on the assumption that the individual covariance matrices $\Sigma(v_i)$ are non-singular.

Consider first the \textit{identically distributed} case, in which we recall from Corollary 4 that the limiting distribution in the Central Limit Theorem for the MRDPG is equal to that of the GRDPG, scaled by a factor of $\frac{1}{m}$. In this case, the MRDPG will \textit{always} outperform a single embedding, regardless of the sample size:

**Proposition 9.** Let $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}$ be identically distributed as a multilayer stochastic blockmodel. Then $C_A \geq C_{A^{(r)}}$ for any $r \in \{1, \ldots, m\}$.

**Proof.** Fix a subset $\mathcal{X}$ of $\mathbb{R}^d$ and a distribution $F$, and for any $i \in \{1, \ldots, K\}$ let

$$\Sigma_i = \frac{1}{m} \mathbf{I}_{p,q} \Delta^{-1} \mathbb{E} \left[ v_i^\top \mathbf{I}_{p,q} \xi \left( 1 - v_i^\top \mathbf{I}_{p,q} \xi \right) \Delta^{-1} \mathbf{I}_{p,q} \right],$$

where $\xi \sim F$, so that

$$C_A = \min_{i \neq j} \sup_{t \in (0,1)} C_t \left( \mathcal{N}(\mathbf{v}_i, \frac{1}{m} \Sigma_i), \mathcal{N}(\mathbf{v}_j, \frac{1}{m} \Sigma_j) \right), \quad C_{A^{(r)}} = \min_{i \neq j} \sup_{t \in (0,1)} C_t \left( \mathcal{N}(\mathbf{v}_i, \Sigma_i), \mathcal{N}(\mathbf{v}_j, \Sigma_j) \right).$$
But for any \(i \neq j\) and any \(t \in (0,1)\),
\[
C_t \left( \mathcal{N}(v_i, \frac{1}{m} \Sigma_i), \mathcal{N}(v_j, \frac{1}{m} \Sigma_j) \right) - C_t \left( \mathcal{N}(v_i, \Sigma_i), \mathcal{N}(v_j, \Sigma_j) \right) = \frac{(m-1)(1-t)}{2} (v_i - v_j)^\top \Sigma_i^{-1} (v_i - v_j),
\]
where \(\Sigma_i = t \Sigma_i + (1-t) \Sigma_j\) is by definition positive semi-definite, and thus \(C_{A,t} \geq C_{A(\cdot),t}\) from which the result follows. \(\square\)

If the adjacency matrices \(A^{(r)}\) are not identically distributed, however, the situation is not so clear-cut, as it is entirely possible to encounter situations for which \(C_{A^{(r)}} > C_A\) for some \(r \in \{1, \ldots, m\}\). For example, if we consider the two-graph multilayer stochastic blockmodel with matrices
\[
B^{(1)} = \begin{pmatrix} 0.67 & 0.46 \\ 0.46 & 0.36 \end{pmatrix}, \quad B^{(2)} = \begin{pmatrix} 0.98 & 0.49 \\ 0.49 & 0.10 \end{pmatrix},
\]
then while the ratio \(C_A/C_{A^{(1)}}\) tends to 11.98, the ratio \(C_A/C_{A^{(2)}}\) tends to 0.96.

Before proceeding further, we note that any analysis of the Chernoff information for large-scale multilayer stochastic blockmodels can be simplified by observing that the logarithmic term in the definition is independent of the number of vertices \(n\), and so becomes insignificant as \(n \to \infty\) under our simplifying assumption that the covariance matrices \(\Sigma(v_i)\) be non-singular. To this end, we consider instead the truncated terms \(\rho(F_i, F_j) = \sup_{t \in (0,1)} \rho_t(F_i, F_j)\), where
\[
\rho_t(F_i, F_j) = \frac{t(1-t)}{2} (x_i - x_j)^\top \Sigma_t^{-1} (x_i - x_j)
\]
for general Gaussian distributions \(F_i\) and \(F_j\), and for the MRDPG define
\[
\rho_A = \min_{i \neq j} \rho \left( \mathcal{N}(v_i, \Delta_A \Sigma(v_i) \Delta_A^\top), \mathcal{N}(v_j, \Delta_A \Sigma(v_j) \Delta_A^\top) \right).
\]

Considering the function \(\rho_A\) gives an accurate means of comparison between the large-scale behaviour of two multilayer stochastic blockmodels, as the ratio \(C_A/C_{A^{(r)}}\) tends to \(\rho_A/\rho_{A^{(r)}}\) as \(n\) increases for any two unfolded adjacency matrices \(A\) and \(A^{(r)}\).

Figure 12 demonstrates the behaviour of \(\rho_A\) for the multilayer stochastic blockmodel for values of \(K \in \{2, \ldots, 5\}\) and \(m \in \{2, \ldots, 12\}\). For each pair \((K,m)\), 1000 samples of matrices \(\{B^{(1)}, \ldots, B^{(m)}\}\) were generated, and the corresponding values of \(\rho_A\) and \(\rho_{A^{(r)}}\) calculated. The plot shows, for each value of \(m\), the proportion of examples for which the MRDPG outperforms every individual embedding (that is, for which the ratio \(\rho_A/\rho_{A^{(r)}} > 1\) for all \(r\)).

For \(K = 2\), the MRDPG outperforms every individual GRDPG at least 80% of the time, with the proportion roughly increasing as the number of embeddings increases. For \(K > 2\), the MRDPG shows even greater improvement, producing the best performance at least 95% of the time. For every example, it was observed that \(\rho_A/\rho_{A^{(r)}} > 1\) for at least one value of \(r\), prompting the following conjecture:

**Conjecture 10.** Let \(A^{(1)}, \ldots, A^{(m)}\) be distributed as a multilayer stochastic blockmodel. Then \(\rho_A/\rho_{A^{(r)}} \geq 1\) for at least one value of \(r \in \{1, \ldots, m\}\).
Figure 12: Proportion (out of 1000 trials) of $K$-community multilayer stochastic blockmodels for which $\rho_A/\rho_{A(r)} > 1$ for all $r$.

If true, this would suggest that considering the MRDPG can never make things worse, in the sense that it will perform at least as well as the worst out of the individual graphs (as opposed to, say, the mean embedding, which we have seen can lead to degeneracy when the matrices $P^{(r)}$ have differing signatures).

8. Conclusion

The multilayer random dot product graph is a natural extension of the generalised random dot product graph to accommodate multiple graphs with a common set of nodes but a mixture of assortativity behaviours. Its simplicity and flexibility make it an ideal models for a variety of situations, and it can be seen to perform equal to (and in many cases better than) existing models at multiple graph inference tasks such as community detection and graph-to-graph comparison. These experimental results are supported by theoretical results showing that the node representations obtained by the left- and right-sided spectral embeddings converge uniformly in the Euclidean norm to the latent positions with Gaussian error. Finally, we demonstrate the practical effectiveness of our model by applying it to the task of link prediction within a computer network, indicating its usefulness to the field of cyber-security.

9. Bibliography

[1] L. Akoglu and C. Faloutsos. Anomaly, event, and fraud detection in large network datasets. In Proceedings of the sixth ACM international conference on Web search and data mining, pages
S. M. Ali and S. D. Shelvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B*, 28:121–132, 1966.

J. Arroyo, A. Athreya, J. Cape, G. Chen, C. E. Priebe, and J. T. Vogelstein. Inference for multiple heterogeneous networks with a common invariant subspace. *arXiv preprints arXiv:1906.10026*, 2019.

A. Athreya, C. E. Priebe, M. Tang, V. Lyzinski, D. J. Marchette, and D. L. Sussman. A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A*, 78(1):1–18, 2016.

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

J. Cape, M. Tang, and C. E. Priebe. The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *The Annals of Statistics*, 2017. To appear; preprint available at http://arxiv.org/abs/1705.10735.

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

J. Cape, M. Tang, and C. E. Priebe. Signal-plus-noise matrix models: eigenvector deviations and fluctuations. *Biometrika*, 106(1):243–250, 2019.

I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

P. Erdős and A. Rényi. On the evolution of random graphs. *Proceedings of the Hungarian Academy of Sciences*, pages 17–61, 1960.

T. Fawcett. An introduction to roc analysis. *Pattern Recogn. Lett.*, 27(8):861874, 2006.

A. Fornito, A. Zalesky, and E. Bullmore. *Fundamentals of brain network analysis*. Academic Press, 2016.

E. N. Gilbert. Random graphs. *Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

Chernoff. H. 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, 1952.

P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97:1090–1098, 2002.

P. W. Holland, K. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5:109–137, 1983.

R. Horn and C. Johnson. *Matrix Analysis (Second Edition)*. Cambridge University Press, New York, NY, 2012.

S. Khor. Concurrency and network disassortativity. *Artificial life*, 16(3):225–232, 2010.
[19] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter. Multilayer networks. *Journal of complex networks*, 2(3):203–271, 2014.

[20] L. Lathauwer, B. Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21(4):1253–1278, 2000.

[21] K. Levin, A. Athreya, M. Tang, V. Lyzinski, and C. E. Priebe. A central limit theorem for an omnibus embedding of random dot product graphs. *arXiv preprint arXiv:1705.09355*, 2017.

[22] F. Liese and I. Vadja. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52:4394–4412, 2006.

[23] V. Lyzinski, M. Tang, A. Athreya, Y. Park, and C. E. Priebe. Community detection and classification in hierarchical stochastic blockmodels. *IEEE Transactions in Network Science and Engineering*, 4(1):13–26, 2017.

[24] M. E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89:208701, 2002.

[25] A. M. Nielsen and D. Witten. The multiple random dot product graph model. *arXiv preprint arXiv:1811.12172*, 2018.

[26] L. Pardo. *Statistical inference based on divergence measures*. Chapman and Hall/CRC, 2005.

[27] F. S. Passino, A. S. Bertiger, J. C. Neil, and N. A. Heard. Link prediction in dynamic networks using random dot product graphs. *arXiv preprint arXiv:1912.10419*, 2019.

[28] P. Rubin-Delanchy, J. Cape, C. E. Priebe, and M. Tang. A statistical interpretation of spectral embedding: the generalised random dot product graph. *arXiv preprint arXiv:1709.05506v3*, 2020.

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

[30] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):205–233, 2016.

[31] D. L. Sussman, M. Tang, D. E. Fishkind, and C. E. Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, 2012.

[32] M. Tang and C. Priebe. Limit theorems for eigenvectors of the normalized laplacian for random graphs. *The Annals of Statistics*, 2019. To appear.

[33] M. Tang, A. Athreya, D. L. Sussman, V. Lyzinski, Y. Park, and C. E. Priebe. A semiparametric two-sample hypothesis testing problem for random graphs. *Journal of Computational and Graphical Statistics*, 26(2):344–354, 2017.

[34] R. Tang, M. Ketcha, A. Badea, E. Calabrese, D. Margulies, J. Vogelstein, C. Priebe, and D. Sussman. Connectome smoothing via low-rank approximations. *IEEE Transactions on Medical Imaging*, 38(6):1446–1456, 2019.
[35] J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1–2):1–230, 2015.

[36] M. Turcotte, A. Kent, and C. Hash. Unified host and network data set. *Data Science for Cyber-Security*, pages 1–22, 2018.

[37] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[38] S. Wang, J. D. Arroyo-Relin, J. T. Vogelstein, and C. E. Priebe. Joint embedding of graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. To appear.

[39] Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the davis-kahan theorem for statisticians. *Biometrika*, 102:315–323, 2015.

[40] M. Zhu and A. Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51(2):918–930, 2006.
10. Appendix: Proofs of main results

Before proving Theorems 1 and 2, we first require some control over the asymptotic behaviour of the singular values of the matrices $P, A$ and $A - P$, which we establish in the following results:

**Proposition 11.** Let $\xi, \xi_1, \ldots, \xi_n \sim F$ and suppose that $\Delta = \mathbb{E}[\xi \xi^\top] \in \mathbb{R}^{d \times d}$ is of rank $d$. Let $X_i = \rho_n^{1/2} \xi_i$ for $i = 1, \ldots, n$ for a given sparsity factor $\rho_n$. Then there exist constants $c_1$ and $c_2$ such that the non-zero singular values $\sigma_i(P)$ for $i \in \{1, \ldots, d\}$ satisfy $c_1 n \rho_n \leq \sigma_i(P) \leq c_2 m^{1/2} n \rho_n$ almost surely.

**Proof.** We begin with the lower bound. Note that $P = X\Lambda Z^\top$, where $Z = I_m \otimes X \in \mathbb{R}^{mn \times md}$, and so

$$\sigma_i(P) = \sqrt{\lambda_i(XAZ^\top Z\Lambda^\top X^\top)} = \sqrt{\lambda_i(X^\top XAZ^\top Z\Lambda^\top)}.$$

as the non-zero eigenvalues of a product of matrices are invariant under cyclic permutations of its factors. The matrix $X^\top XAZ^\top Z\Lambda^\top \in \mathbb{R}^{d \times d}$, and so, using that standard fact that the eigenvalues of any matrix are bounded below in absolute value by its smallest singular value, we see that

$$\sigma_i(P) \geq \sqrt{\sigma_d(X^\top XAZ^\top Z\Lambda^\top)}.$$

Now, for any $i, j \in \{1, \ldots, d\}$, the term

$$(X^\top X - n \rho_n \Delta)_{ij} = \sum_{k=1}^n \rho_n (\xi_{ki} \xi_{kj} - \Delta_{ij})$$

is a sum of $n$ bounded zero-mean independent random variables, and thus by Hoeffding’s Inequality we have

$$\left| (X^\top X - n \rho_n \Delta)_{ij} \right| = O\left(n^{1/2} \rho_n \log^{1/2}(n)\right)$$

almost surely. Summing over all such indices $i, j$ and using the fact that the spectral norm is bounded above by the Frobenius norm then shows that

$$\|X^\top X - n \rho_n \Delta\| = O\left(n^{1/2} \rho_n \log^{1/2}(n)\right)$$

almost surely. An identical argument shows that

$$\|Z^\top Z - n \rho_n \hat{\Delta}\| = O\left((mn)^{1/2} \rho_n \log^{1/2}(n)\right)$$

almost surely, where $\hat{\Delta} = I_m \otimes \Delta \in \mathbb{R}^{md \times md}$ and consequently that

$$\|Z^\top Z\| \leq \|n \rho_n \hat{\Delta}\| + \|Z^\top Z - n \rho_n \hat{\Delta}\| = O\left(m^{1/2} n \rho_n\right)$$

almost surely by a standard application of the triangle inequality.
Next, note that
\[
X^\top XAZ^\top ZA^\top - n^2 \rho_n^2 \Delta \Delta A^\top = (X^\top X - n\rho_n \Delta)AZ^\top ZA^\top + n\rho_n \Delta (Z^\top Z - n\rho_n \Delta)A^\top,
\]
and so
\[
\|X^\top XAZ^\top ZA^\top - n^2 \rho_n^2 \Delta \Delta A^\top\| \leq \|A\|^2 (\|X^\top X - n\rho_n \Delta\| \|Z^\top Z\| + n\rho_n \|\Delta\| \|Z^\top Z - n\rho_n \Delta\|)
\]
\[
= O \left( m^{1/2} n^{3/2} \rho_n^2 \log^{1/2}(n) \right)
\]
almost surely.

A corollary of Weyl’s inequalities (see, for example, [17], Corollary 7.3.5) states that
\[
|\sigma_i(A) - \sigma_i(B)| \leq \|A - B\|
\]
for any matrices \(A\) and \(B\), and applying the reverse triangle inequality shows that
\[
\sigma_i(B) \geq \sigma_i(A) - \|A - B\|.
\]

In particular, we find that
\[
\sigma_d(X^\top XAZ^\top ZA^\top) \geq n^2 \rho_n^2 \sigma_d(\Delta \Delta A^\top) - \|X^\top XAZ^\top ZA^\top - n^2 \rho_n^2 \Delta \Delta A^\top\| \geq c(n\rho_n)^2
\]
almost surely for some constant \(c\), from which we obtain our desired lower bound by taking square roots.

The upper bound is obtained by observing that \(\sigma_1(P)^2 = \|P\|^2 \leq \|P\|_F^2 = O(mn^2 \rho_n)\), as the entries \(P_{ij}\) are \(O(\rho_n)\).

\[\Box\]

**Proposition 12.** \(\|A - P\| = O \left( (mn \rho_n \log(mn))^{1/2} \right) \) almost surely.

**Proof.** We will make use of a matrix analogue of the Bernstein inequality (see [35], Theorem 1.6.2):

**Theorem 13 (Matrix Bernstein).** Let \(M_1, \ldots, M_n\) be independent random matrices with common dimension \(d_1 \times d_2\), satisfying \(E[M_k] = 0\) and \(\|M_k\| \leq L\) for each \(1 \leq k \leq n\), for some fixed value \(L\).

Let \(M = \sum_{k=1}^n M_k\) and let \(v(M) = \max\{\|E[MM^\top]\|, \|E[M^\top M]\|\}\) denote the matrix variance statistic of \(M\). Then for all \(t \geq 0\), we have
\[
P(\|M\| \geq t) \leq (d_1 + d_2) \exp \left( \frac{-t^2/2}{v(M) + Lt/3} \right).
\]
We apply this Theorem as follows: for each $1 \leq i \leq j \leq n$ and each $1 \leq r \leq m$, let $M^{(r)}_{ij}$ be the $n \times mn$ matrix with $(i, (r-1)n + j)$th and $(j, (r-1)n + i)$th entries equal to $A^{(r)}_{ij} - P^{(r)}_{ij}$, and all other entries equal to 0 (in other words, if we divide $M^{(r)}_{ij}$ into $m$ distinct $n \times n$ blocks, then the $r^{th}$ block is the only non-zero one, and within this block only the $(i,j)$th and $(j,i)$th entries are non-zero).

Observe that $\|M^{(r)}_{ij}\| = |A_{ij} - P_{ij}| < 1$, and by definition $E[M^{(r)}_{ij}] = 0$, so

$$M = \sum_{r=1}^{m} \sum_{i,j=1}^{n} M^{(r)}_{ij} = A - P$$

satisfies the criteria for Bernstein’s Theorem. To bound the matrix variance statistic $v(M)$, observe firstly that

$$[MM^T]_{ij} = \sum_{r=1}^{m} \sum_{k=1}^{n} (A^{(r)}_{ik} - P^{(r)}_{ik})(A^{(r)}_{jk} - P^{(r)}_{jk}),$$

and thus

$$E[MM^T]_{ij} = \begin{cases} \sum_{r=1}^{m} \sum_{k=1}^{n} P^{(r)}_{ik}(1 - P^{(r)}_{ik}) & i = j \\ 0 & i \neq j \end{cases}$$

By definition, there is some constant $c$ such that $P^{(r)}_{ik}(1 - P^{(r)}_{ik}) \leq c \rho_n$ for all $i$ and $k$, and so since $E[MM^T]$ is diagonal, we see that $\|E[MM^T]\| \leq c mn \rho_n$. Similarly,

$$[M^TM]_{ij} = \sum_{k=1}^{n} (A^{(r(i))}_{ki} - P^{(r(i))}_{ki})(A^{(r(j))}_{kj} - P^{(r(j))}_{kj}),$$

where $r(i)$ denotes the block containing the $i$th column of $M$. Thus

$$E[M^TM]_{ij} = \begin{cases} \sum_{k=1}^{n} P^{(r(i))}_{ik}(1 - P^{(r(i))}_{ik}) & i = j \\ 0 & i \neq j \end{cases}$$

and so $\|E[M^TM]\| \leq cn \rho_n$, so in particular $v(M) \leq c mn \rho_n$.

Substituting these into Bernstein’s Theorem and rearranging, we find that, for any $t \geq 0$,

$$P(\|A - P\| \geq t) \leq (m + 1)n \exp\left(\frac{-3t^2}{6c mn \rho_n + 2t}\right).$$

The numerator of the exponential term dominates for $n$ sufficiently large if $t = O\left((mn \rho_n \log(mn))^{1/2}\right)$, from which the result follows. \[\square\]
Corollary 14. There exist constants $c_1$ and $c_2$ such that the leading $d$ singular values of $A$ satisfy $c_1nρ_n ≤ σ_i(A) ≤ c_2nm^{1/2}nρ_n$ almost surely.

Proof. The generalisation of Weyl’s inequalities used in Proposition 11 shows that
\[ \sigma_d(A) ≥ σ_d(P) - \|A - P\| \]
from which we obtain our desired lower bound by combining Propositions 11 and 12.

The argument for the upper bound is identical to that in Proposition 11. □

The next results establish the existence of, and some properties relating to, the matrices $L_X$ and $L_{Y,r}$ which map the latent positions $X$ to the embeddings $X_P$ and $Y_{P,r}$, respectively. In particular we show that the inverses of these matrices have bounded spectral norm which in turn shows that the matrices $L_{(n)}$ and $L_{(r)}^{(n)}$ used to transform the spectral embeddings have bounded norm.

Proposition 15. Let $n$ be large enough that $\text{rank}(X) = d$. Then there exist matrices $L_X ∈ \text{GL}(d)$ and $L_{Y,r} ∈ \mathbb{R}^{d×d}$ such that
\[ X_P = XL_X, \quad Y_{P,r} = XL_{Y,r}. \]
The matrices $L_X$ and $L_{Y,r}$ satisfy $L_XL_{Y,r}^\top$, and so in particular $\text{rank}(L_{Y,r}) = \text{rank}(Λ_r)$.

Proof. Define symmetric matrices $Π ∈ \text{GL}(d)$ and $Π_r ∈ \mathbb{R}^{d×d}$ such that
\[ Π^2 = ΛZ^\top ZΛ^\top, \quad Π_r^2 = Λ_rX^\top XΛ_r^\top, \]
and observe that
\[ \left( X_P Σ_p^{1/2} \right) \left( X_P Σ_p^{1/2} \right)^\top = UPΣ_P^2U_P^\top = PP^\top = (XΠ)(XΠ)^\top, \]
and similarly (noting that $U_PΣ_PV_{P,r} = P^{(r)}$)
\[ \left( Y_{P,r} Σ_p^{1/2} \right) \left( Y_{P,r} Σ_p^{1/2} \right)^\top = P^{(r)}P^{(r)\top} = (XΠ_r)(XΠ_r)^\top. \]
Thus there exist orthogonal matrices $Q_P, Q_{P,r} ∈ O(d)$ such that
\[ X_PΣ_p^{1/2} = XPQ_P, \quad Y_{P,r}Σ_p^{1/2} = XP_{P,r}Q_P, \]
and so
\[ L_X = ΠQ_PΣ_p^{1/2} ∈ \text{GL}(d), \quad L_{Y,r} = Π_rQ_{P,r}Σ_p^{-1/2} ∈ \mathbb{R}^{d×d} \]
are our desired matrices.

For the final statement, observe that
\[ XL_XL_{Y,r}^\top X = XPY_{P,r}^\top = P^{(r)} = XL_AX^\top, \]
and so the result follows after multiplying by $(X^\top X)^{-1}X^\top$ and $X(X^\top X)^{-1}$ on the left and right respectively. □
Corollary 16. The matrices \( L_X^{-1} \) and \( L_{Y,r}^{-1} \) (where they exist) satisfy \( \|L_X^{-1}\|, \|L_{Y,r}^{-1}\| = O(m^{1/4}) \) almost surely.

Proof. An identical argument to the proof of Proposition 11 shows us that there is some constant \( c \) for which \( \sigma_d(\Pi), \sigma_d(\Pi_r) \geq cn^{1/2} \) almost surely, and thus \( \|\Pi^{-1}\|, \|\Pi_r^{-1}\| = O((n\rho_n)^{-1/2}) \) almost surely.

Similarly, \( \|\Sigma_P\| = O(m^{1/2}n\rho_n) \) almost surely by Proposition 11, and so the result follows from submultiplicativity and unitary invariance of the spectral norm. □

Proposition 17. Let \( n \) be large enough that \( \text{rank}(X) = d \). Then the following identities hold:

\[
L_{Y,r} \Sigma_P^{-1} L_X^{-1} = \Lambda_r (AZ^TZA^T)^{-1}, \quad L_X \Sigma_P^{-1} L_{Y,r} = (X^T X)^{-1} \Lambda_r^{-1}.
\]

Proof. Recall from Proposition 15 that \( L_X L_{Y,r}^T = \Lambda_r \). Similarly, since

\[
XL_X \Sigma_P X^T X = X_P \Sigma_P X_P^T = PP^T = XAZ^TZA^T X^T,
\]

we find that

\[
L_X \Sigma_P L_X^T = AZ^TZA^T.
\]

Then

\[
L_{Y,r} \Sigma_P^{-1} L_X^{-1} = L_{Y,r} L_X^T \Sigma_P^{-1} L_X^{-1} = \Lambda_r (AZ^TZA^T)^{-1}
\]

and

\[
L_X \Sigma_P^{-1} L_{Y,r}^{-1} = (X^T X)^{-1} X^T X_P \Sigma_P^{-1} L_{Y,r}^{-1} = (X^T X)^{-1} X^T X_P \Sigma_P^{-1} X_P^T X_P \Sigma_P^{-1} L_{Y,r}^{-1} = (X^T X)^{-1} X_P^T X_P \Sigma_P^{-1} L_{Y,r}^{-1} = (X^T X)^{-1} L_{Y,r}^{-1} \Lambda_r^{-1},
\]

where we have used the identities

\[
L_X = (X^T X)^{-1} X^T X_P, \quad X_P^T X_P = \Sigma_P.
\]

The next few results provide bounds on the asymptotic growth of a number of residual terms in the proofs of our main theorems. While the proofs are similar in nature to a number of results in [23], there are some minor differences to account for the fact that the matrices \( A \) and \( P \) are not symmetric, and so we reproduce them in full. We begin an analogue of a bound appearing in Lemma 17 of [23]:

37
Proposition 18. \( \|U_P^T(A - P)V_P\|_F = O \left( (m \log(n))^{1/2} \right) \) almost surely.

Proof. Note that \( U_P^T(A - P)V_P \in \mathbb{R}^{d \times d} \), with entries

\[
(U_P^T(A - P)V_P)_{ij} = \sum_{r=1}^{m} \sum_{k=1}^{n} \sum_{l=1}^{n} u_k v_{(r-1)n+l} \left( A_{kl}^{(r)} - P_{kl}^{(r)} \right)
= \sum_{r=1}^{m} \sum_{1 \leq k < l \leq n} 2u_k v_{(r-1)n+l} \left( A_{kl}^{(r)} - P_{kl}^{(r)} \right) - \sum_{r=1}^{m} \sum_{1 \leq k < l \leq n} u_k v_{(r-1)n+k} P_{kk}^{(r)}
\]

where for ease of notation we let \( u \) and \( v \) denote the \( i \)th and \( j \)th columns of \( U_P \) and \( V_P \) respectively.

To bound the latter term, note that

\[
\left| \sum_{r=1}^{m} \sum_{1 \leq k < l \leq n} u_k v_{(r-1)n+k} P_{kk}^{(r)} \right| \leq \left( \sum_{r=1}^{m} \sum_{1 \leq k < l \leq n} |u_k| \right)^{1/2} \left( \sum_{r=1}^{m} \sum_{1 \leq k < l \leq n} |v_{(r-1)n+k}|^2 \right)^{1/2} = m^{1/2}.
\]

For the former, note that the terms \( 2u_k v_{(r-1)n+l} \left( A_{kl}^{(r)} - P_{kl}^{(r)} \right) \) are independent, zero-mean random variables, bounded in absolute value by \( 2|u_k v_{(r-1)n+l}| \), and so by Hoeffding’s inequality we see that

\[
P \left( \left| \sum_{r=1}^{m} \sum_{1 \leq k < l \leq n} 2u_k v_{(r-1)n+l} \left( A_{kl}^{(r)} - P_{kl}^{(r)} \right) \right| \geq t \right) \leq 2 \exp \left( \frac{-2t^2}{4 \sum_{r=1}^{m} \sum_{1 \leq k < l \leq n} |u_k v_{(r-1)n+l}|^2} \right)
\]

\[
\leq 2 \exp \left( \frac{-2t^2}{4 \sum_{r=1}^{m} \sum_{1 \leq k < l \leq n} |v_{(r-1)n+l}|^2} \right)
\]

\[
= 2 \exp \left( \frac{-t^2}{2} \right)
\]

and so \( \sum_{r=1}^{m} \sum_{1 \leq k < l \leq n} 2u_k v_{(r-1)n+l} \left( A_{kl}^{(r)} - P_{kl}^{(r)} \right) \) is of order \( O \left( \log^{1/2}(n) \right) \) almost surely. Summing over all \( i, j \in \{1, \ldots, d\} \) gives the result. \( \square \)
Before establishing the next set of bounds (which relate to the left and right singular vectors of the matrices $A$ and $P$), we state the following variation of the Davis-Kahan theorem (see [39], Theorem 4):

**Theorem 19 (Variant of Davis-Kahan).** Let $M_1, M_2 \in \mathbb{R}^{m \times n}$ have singular value decompositions

$$M_i = U_i \Sigma_i V_i^\top + U_{i, \bot} \Sigma_{i, \bot} V_{i, \bot}^\top,$$

where $U_i \in \mathbb{O}(m \times d)$ has orthonormal columns corresponding to the $d$ greatest singular values of $M_i$, for some $1 \leq d \leq n$. Then, if $|\sigma_d(M_1)^2 - \sigma_{d+1}(M_1)^2| > 0$, we have

$$\max \{ \| \sin \Theta(U_2, U_1) \|, \| \sin \Theta(V_2, V_1) \| \} \leq \frac{2\sqrt{d}(2\sigma_1(M_1) + \|M_2 - M_1\|)}{\sigma_d(M_1)^2 - \sigma_{d+1}(M_1)^2} \|M_2 - M_1\|,$$

where we take $\sigma_{n+1}(M_1) = -\infty$.

Using this result, we can prove the following:

**Proposition 20.** The following bounds hold almost surely:

1. $\|U_A U_A^\top - U_P U_P^\top\|, \|V_A V_A^\top - V_P V_P^\top\| = O \left( \frac{m \log^{1/2}(mn)}{(n \rho_n)^{1/2}} \right)$;

2. $\|U_A - U_P U_P^\top U_A\|_F, \|V_A - V_P V_P^\top V_A\|_F = O \left( \frac{m \log^{1/2}(mn)}{(n \rho_n)^{1/2}} \right)$;

3. $\|U_P^\top U_A \Sigma_A - \Sigma_P V_P^\top V_A\|_F, \|\Sigma_P U_P^\top U_A - V_P^\top V_A \Sigma_A\|_F = O \left( \frac{m^{3/2} \log(mn)}{n \rho_n} \right)$;

4. $\|U_P^\top U_A - V_P^\top V_A\|_F = O \left( \frac{m^{3/2} \log(mn)}{n \rho_n} \right)$

**Proof.**

1. Let $\sigma_1, \ldots, \sigma_d$ denote the singular values of $U_P^\top U_A$, and let $\theta_i = \cos^{-1}(\sigma_i)$ be the principal angles. It is a standard result that the non-zero eigenvalues of the matrix $U_A U_A^\top - U_P U_P^\top$ are precisely the $\sin(\theta_i)$ (each occurring twice) and so by Davis-Kahan we have

$$\|U_A U_A^\top - U_P U_P^\top\| = \max_{i \in \{1, \ldots, d\}} |\sin(\theta_i)| \leq \frac{2\sqrt{d}(2\sigma_1(P) + \|A - P\|)}{\sigma_d(P)^2} \|A - P\|$$

for $n$ sufficiently large.

The spectral norm bound from **Proposition 12** then shows that

$$\|U_A U_A^\top - U_P U_P^\top\| = O \left( \frac{(\sigma_1(P) + (mn \rho_n \log(mn))^{1/2}(mn \rho_n \log(mn))^{1/2})}{\sigma_d(P)^2} \right)$$

$$= O \left( \frac{m \log^{1/2}(mn)}{(n \rho_n)^{1/2}} \right)$$

after applying **Proposition 11** to bound the terms $\sigma_i(P)$. An identical argument gives the result for $\|V_A V_A^\top - V_P V_P^\top\|$. 

39
ii. Using the bound from part i., we find that

\[ \| U_A - U_P U_P^T U_A \|_F = \|(U_A U_A^T - U_P U_P^T) U_A \|_F \leq \| U_A U_A^T - U_P U_P^T \| \| U_A \|_F = O \left( \frac{m \log^{1/2}(mn)}{(n \rho_n)^{1/2}} \right). \]

An identical argument bounds the term \( \| V_A - V_P V_P^T V_A \|_F \).

iii. Observe that

\[ U_A^T \Sigma_A - \Sigma_P V_P^T V_A = U_P^T (A - P) V_A \]

\[ = U_P^T (A - P) (V_A - V_P V_P^T V_A) + U_P^T (A - P) V_P V_P^T V_A, \]

and so

\[ \| U_P^T U_A \Sigma_A - \Sigma_P V_P^T V_A \|_F \leq \| U_P^T \| \| A - P \| \| V_A - V_P V_P^T V_A \|_F + \| U_P^T (A - P) V_P \|_F \| V_P^T V_A \|_F \]

\[ = O \left( (mn \rho_n \log(mn))^{1/2} \cdot (n \rho_n)^{-1/2} \cdot m \log^{1/2}(mn) \right) + O \left( m^{1/2} \log^{1/2}(n) \right) \]

\[ = O \left( m^{3/2} \log(mn) \right), \]

where we have used Propositions 12, 18 and the result from part ii.

An identical argument bounds the term \( \| \Sigma_P U_P^T U_A - V_P^T V_A \Sigma_A \|_F \).

iv. Note that

\[ U_P^T U_A - V_P^T V_A = ( U_P^T U_A \Sigma_A - \Sigma_P V_P^T V_A ) + ( \Sigma_P U_P^T U_A - V_P^T V_A \Sigma_A ) \Sigma_A^{-1} - \Sigma_P ( U_P^T U_A - V_P^T V_A ) \Sigma_A^{-1}. \]

For any \( i, j \in \{1, \ldots, d\} \) we find (after rearranging and bounding the absolute value of the right-hand terms by the Frobenius norm):

\[ \left| ( U_P^T U_A - V_P^T V_A )_{ij} \right| \left( 1 + \frac{\sigma_i(P)}{\sigma_j(A)} \right) \leq \| U_P^T U_A \Sigma_A - \Sigma_P V_P^T V_A \|_F + \| \Sigma_P U_P^T U_A - V_P^T V_A \Sigma_A \|_F \| \Sigma_A^{-1} \|_F \]

\[ = O \left( \frac{m^{3/2} \log(mn)}{n \rho_n} \right), \]

where we have used the result from part iii. and Corollary 14. The result follows from the fact that

\[ \left( 1 + \frac{\sigma_i(P)}{\sigma_j(A)} \right) \geq 1. \]
The following result (an analogue of [23, Proposition 16]) relates to orthogonal matrix $W$ used to perform a simultaneous Procrustes alignment of $X_A$ with $X_P$ and $Y_A$ with $Y_P$.

**Proposition 21.** Let $U_P^T U_A + V_P^T V_A$ admit the singular value decomposition

$$U_P^T U_A + V_P^T V_A = W_1 \Sigma W_2^T,$$

and let $W = W_1 W_2^T$. Then

$$\|U_P^T U_A - W\|_F, \|V_P^T V_A - W\|_F = O\left(\frac{m^2 \log(mn)}{n \rho_n}\right)$$

almost surely.

**Proof.** A standard argument shows that $W$ minimises the term $\|U_P^T U_A - Q\|_F^2 + \|V_P^T V_A - Q\|_F^2$ among all $Q \in O(d)$. Let $U_P^T U_A = W_{U,1} \Sigma U_{U,2}$ be the singular value decomposition of $U_P^T U_A$, and define $W_U \in O(d)$ by $W_U = W_{U,1} W_{U,2}$. Then

$$\|U_P^T U_A - W_U\|_F = \|\Sigma - I\|_F = \left(\sum_{i=1}^d (1 - \sigma_i)^2\right)^{1/2} \leq \sum_{i=1}^d (1 - \sigma_i) \leq \sum_{i=1}^d (1 - \sigma_i^2)$$

$$= \sum_{i=1}^d \sin^2(\theta_i) \leq d \|U_A U_A^T - U_P U_P^T\| F$$

$$= O\left(\frac{m^2 \log(mn)}{n \rho_n}\right).$$

Also,

$$\|V_P^T V_A - W_U\|_F \leq \|V_P^T V_A - U_P V_A\|_F + \|U_P^T U_A - W\|_F$$

$$= O\left(\frac{m^2 \log(mn)}{n \rho_n}\right)$$

by Proposition 20.

Combining these shows that

$$\|U_P^T U_A - W\|_F^2 + \|V_P^T V_A - W\|_F^2 \leq \|U_P^T U_A - W_U\|_F^2 + \|V_P^T V_A - W_U\|_F^2$$

$$= O\left(\frac{m^4 \log^2(mn)}{(n \rho_n)^2}\right),$$

which gives the desired bound. \qed
The following bounds are a straightforward adaptation of [23], Lemma 17:

**Proposition 22.** The following bounds hold almost surely:

i. \[ \|W\Sigma_A - \Sigma_p W\|_F = O(\left(\frac{m^{5/2} \log(mn)}{n^2}\right))^5 \];

ii. \[ \|W\Sigma_A^{1/2} - \Sigma_p^{1/2} W\|_F = O\left(\frac{m^{5/2} \log(mn)}{(n^2p^2)^2}\right) \];

iii. \[ \|W\Sigma_A^{-1/2} - \Sigma_p^{-1/2} W\|_F = O\left(\frac{m^{5/2} \log(mn)}{(n^2p^2)^{3/2}}\right) \].

**Proof.**

i. Observe that

\[ W\Sigma_A - \Sigma_p W = (W - U_p^T U_A)\Sigma_A + U_p^T U_A \Sigma_A - \Sigma_p W = (W - U_p^T U_A)\Sigma_A + (U_p^T U_A \Sigma_A - \Sigma_p V_p^T V_A) + \Sigma_p (V_p^T V_A - W). \]

The terms \( (W - U_p^T U_A)\Sigma_A \) and \( \|\Sigma_p (V_p^T V_A - W)\|_F \) are both \( O\left(\frac{m^{5/2} \log(mn)}{n^2}\right) \) (as shown by Proposition 21 and Corollary 14), while \( \|U_p^T U_A \Sigma_A - \Sigma_p V_p^T V_A\|_F \) is \( O\left(\frac{m^{3/2} \log(mn)}{n^2}\right) \), and so \( \|W\Sigma_A - \Sigma_p W\|_F = O\left(\frac{m^{5/2} \log(mn)}{n^2}\right) \).

ii. We will bound the absolute value of the terms \( \left(W\Sigma_A^{1/2} - \Sigma_p^{1/2} W\right)_{ij} \). Note that

\[ \left|\left(W\Sigma_A^{1/2} - \Sigma_p^{1/2} W\right)_{ij}\right| = \left|W_{ij} \left(\sigma_j(A)^{1/2} - \sigma_i(P)^{1/2}\right)\right| = \left|\frac{W_{ij}(\sigma_j(A) - \sigma_i(P))}{\sigma_j(A)^{1/2} + \sigma_i(P)^{1/2}}\right| \]

\[ = \frac{|(W\Sigma_A - \Sigma_p W)_{ij}|}{\sigma_j(A)^{1/2} + \sigma_i(P)^{1/2}} \leq \frac{\|W\Sigma_A - \Sigma_p W\|_F}{\sigma_d(P)^{1/2}} \]

and consequently we find that \( \|W\Sigma_A^{1/2} - \Sigma_p^{1/2} W\|_F = O\left(\frac{m^{5/2} \log(mn)}{(n^2p^2)^{3/2}}\right) \) by summing over all \( i, j \in \{1, \ldots, n\} \) and applying part i.

iii. We will bound the absolute value of the terms \( \left(W\Sigma_A^{-1/2} - \Sigma_p^{-1/2} W\right)_{ij} \). Note that

\[ \left|\left(W\Sigma_A^{-1/2} - \Sigma_p^{-1/2} W\right)_{ij}\right| = \left|\frac{W_{ij}(\sigma_i(P)^{1/2} - \sigma_j(A)^{1/2})}{\sigma_i(P)^{1/2} \sigma_j(A)^{1/2}}\right| \]

\[ = \left|\frac{(W\Sigma_A - \Sigma_p W)_{ij}}{\sigma_i(P)^{1/2} \sigma_j(A)^{1/2}}\right| = \frac{\|W\Sigma_A - \Sigma_p W\|_F}{\sigma_i(P)^{1/2} \sigma_j(A)^{1/2}} \]

\[ = O\left(\frac{m^{5/2} \log(mn)}{(n^2p^2)^{3/2}}\right) \]

by part ii. The result follows by summing over all \( i, j \in \{1, \ldots, n\} \). \( \square \)
We prove the bounds only for the terms $R_{1,i}$.

Proving the bounds for $R_{1,1}$, we have

$$R_{1,1} = U_P(U_P^T U_A \Sigma_A^{1/2} - \Sigma_P^{1/2} W)$$

and

$$R_{1,2} = (I - U_P U_P^T)(A - P)(V_A - V_P W)\Sigma_A^{-1/2}$$

$$R_{1,3} = -U_P U_P^T(A - P) V_P W \Sigma_A^{-1/2}$$

$$R_{1,4} = (A - P) V_P(W \Sigma_A^{-1/2} - \Sigma_P^{-1/2} W)$$

Then the following bounds hold almost surely:

i. $\|R_{1,1}\|_{2\to\infty}, \|R_{2,1}\|_{2\to\infty} = O \left( \frac{m^{1/4} \log(nm)}{n \rho_n} \right)$;

ii. $\|R_{1,2}\|_{2\to\infty}, \|R_{2,2}\|_{2\to\infty} = O \left( \frac{m^{1/4} \log^{3/2}(nm)}{n^{3/4} \rho_n} \right)$;

iii. $\|R_{1,3}\|_{2\to\infty}, \|R_{2,3}\|_{2\to\infty} = O \left( \frac{m^{3/4} \log^{1/2}(n)}{n^{3/4} \rho_n} \right)$;

iv. $\|R_{1,4}\|_{2\to\infty}, \|R_{2,4}\|_{2\to\infty} = O \left( \frac{m^{3/4} \log^{3/2}(nm)}{n^{3/4} \rho_n} \right)$.

In particular, if $\rho_n = o \left( \frac{\log^{3/2}(n)}{n^{1/2}} \right)$ then $\|R_{i,j}\|_{2\to\infty}$ for all $i, j$.

Proof. We prove the bounds only for the terms $R_{1,i}$, the proofs for the terms $R_{2,i}$ are identical.

i. Recall that $U_P \Sigma_P^{1/2} = X L_X$ for some $L_X \in \text{GL}(d)$, which can be shown to satisfy $\|L_X\| = O \left( m^{1/4} \right)$ by an identical argument to Corollary 16. Using the relation $\|AB\|_{2\to\infty} \leq \|A\|_{2\to\infty} \|B\|$ (see, for example, [6], Proposition 6.5) we find that $\|U_P\|_{2\to\infty} \leq \|X\|_{2\to\infty} \|L_X\| \|\Sigma_P^{-1/2}\|$, and thus $\|U_P\|_{2\to\infty} = O \left( m^{1/4} n^{-1/2} \right)$ as the rows of $X$ are by definition bounded in Euclidean norm.

Thus

$$\|R_{1,1}\|_{2\to\infty} \leq \|U_P\|_{2\to\infty} \|U_P^T U_A \Sigma_A^{1/2} - \Sigma_P^{1/2} W\|$$

$$\leq \|U_P\|_{2\to\infty} \left( \|(U_P^T U_A - W) \Sigma_A^{1/2}\|_F + \|W \Sigma_A^{1/2} - \Sigma_P^{1/2} W\|_F \right)$$
The first summand is \( O \left( (n \rho_n)^{-1/2} m^{9/4} \log(mn) \right) \) by Proposition 21 and Corollary 14, while Proposition 22 shows that the second is \( O \left( (n \rho_n)^{-1/2} m^{5/2} \log(mn) \right) \), and so

\[
\| R_{1,1} \|_{2 \to \infty} = O \left( \frac{m^{11/4} \log(mn)}{n \rho_n^{1/2}} \right).
\]

ii. We first observe that

\[
\| U_P U_P^T (A - P)(V_A - V_P W) \Sigma_A^{-1/2} \|_{2 \to \infty} \leq \| U_P \|_{2 \to \infty} \| U_P^T \| \| A - P \| \| V_A - V_P W \| \| \Sigma_A^{-1/2} \|
\]

\[
= O \left( m^{1/4} n^{-1/2} \cdot (mn \rho_n \log(mn))^{1/2} \cdot \frac{m^2 \log(mn)}{n^{1/2} \rho_n} \cdot (n \rho_n)^{-1/2} \right)
\]

\[
= O \left( \frac{m^{11/4} \log^{3/2}(mn)}{n \rho_n} \right),
\]

where we have bounded \( \| V_A - V_P W \| \) by noting that

\[
\| V_A - V_P W \| \leq \| V_A - V_P V_P^T V_A \| + \| V_P (V_P^T V_A - W) \|
\]

\[
= O \left( \frac{m^2 \log(mn)}{n^{1/2} \rho_n} \right),
\]

by Propositions 20 and 21.

This leaves us to bound the term \( \| (A - P)(V_A - V_P W) \Sigma_A^{-1/2} \|_{2 \to \infty} \). Now,

\[
(A - P)(V_A - V_P W) \Sigma_A^{-1/2} = (A - P)(I - V_P V_P^T) V_A \Sigma_A^{-1/2} + (A - P)V_P (V_P^T V_A - W) \Sigma_A^{-1/2},
\]

and

\[
\| (A - P)V_P (V_P^T V_A - W) \Sigma_A^{-1/2} \|_{2 \to \infty} \leq \| A - P \| \| V_P \| \| V_P^T V_A - W \| \| \Sigma_A^{-1/2} \|
\]

\[
= O \left( (mn \rho_n \log(mn))^{1/2} \cdot \frac{m^2 \log(mn)}{n \rho_n} \cdot (n \rho_n)^{-1/2} \right)
\]

\[
= O \left( \frac{m^{5/2} \log^{3/2}(mn)}{n \rho_n} \right)
\]

by Propositions 12, 21 and Corollary 14.

To bound the remaining term, observe that we can rewrite

\[
(A - P)(I - V_P V_P^T) V_A \Sigma_A^{-1/2} = (A - P)(I - V_P V_P^T) V_A V_A^T V_A \Sigma_A^{-1/2}
\]

and so

\[
\| (A - P)(I - V_P V_P^T) V_A \Sigma_A^{-1/2} \|_{2 \to \infty} \leq \| R \|_{2 \to \infty} \| V_A \Sigma_A^{-1/2} \|,
\]
where
\[ R = (A - P)(I - V_P V_P^\top)V_A V_A^\top. \]

The term \( \|V_A \Sigma_A^{-1/2}\| \) is \( O((n \rho_n)^{-1/2}) \) by Corollary 14, so it suffices to bound \( \|R\|_{2\to\infty} \). To do this, we claim that the Frobenius norms of the rows of the matrix \( R \) are exchangeable, and thus have the same expectation, which implies that \( \mathbb{E}(\|R_i\|_F^2) = n \mathbb{E}(\|R_i\|_F) \) for any \( i \in \{1, \ldots, n\} \). Applying Markov’s inequality, we therefore see that
\[
\mathbb{P}(\|R_i\| > t) \leq \frac{\mathbb{E}(\|R_i\|_F^2)}{t^2} = \frac{\mathbb{E}(\|R\|_F^2)}{nt^2}.
\]

Now,
\[
\|R\|_F \leq \|A - P\| \|V_A - V_P V_P^\top V_A\|_F \|V_A^\top\|_F
\]
\[
= O \left( (mn \rho_n \log(mn))^{1/2}, \frac{m \log^{1/2}(mn)}{(n \rho_n)^{1/2}} \right)
\]
\[
= O \left( m^{3/2} \log(mn) \right)
\]
by Propositions 12 and 20. It follows that
\[
\mathbb{P}(\|R_i\| > n^{-1/4} m^{3/2} \log(mn)) \leq cn^{-1/2}
\]
and thus
\[
\|R\|_{2\to\infty} = O \left( \frac{m^{3/2} \log(mn)}{(n \rho_n)^{1/2}} \right)
\]
almost surely.

We must therefore show that the Frobenius norms of the rows of \( R \) are exchangeable. Let \( Q \in O(n) \) be a permutation matrix, and let \( \hat{Q} = I_m \otimes Q \in O(mn) \). For any matrix \( G \in \mathbb{R}^{n \times mn} \), right multiplication by the matrix \( \hat{Q}^\top \) simply permutes the columns of \( G \), and thus does not alter the Frobenius norms of its rows. In particular, the Frobenius norms of the rows of \( QG \hat{Q}^\top \) are the same as the Frobenius norms of the rows of \( QG \). For any matrix \( G \in \mathbb{R}^{n \times mn} \), let \( \mathcal{L}_d(G) \) denote the projection onto the subspace spanned by the left singular vectors corresponding to the leading \( d \) singular values of \( G \), and let \( \mathcal{R}_d(G) \) denote the projection onto the subspace spanned by the right singular vectors corresponding to these singular values. Similarly, let \( \mathcal{L}_\perp(G) \) and \( \mathcal{R}_\perp(G) \) denote the projections onto the orthogonal complements of these subspaces.

Note that
\[
\mathcal{R}_d(P) = U_P U_P^\top, \quad \mathcal{L}_d(P) = V_P V_P^\top
\]
and similarly
\[
\mathcal{R}_d(A) = U_A U_A^\top, \quad \mathcal{L}_d(A) = V_A V_A^\top.
\]
while for any permutation matrix $Q \in \mathbb{O}(n)$ we have
$$
\mathcal{R}_d(QP\hat{Q}^T) = QU_P U_P^T Q^T, \quad \mathcal{L}_d(QP\hat{Q}^T) = \hat{Q}V_P V_P^T \hat{Q}^T
$$
and similarly
$$
\mathcal{R}_d(QA\hat{Q}^T) = QU_A U_A^T Q^T, \quad \mathcal{L}_d(QA\hat{Q}^T) = \hat{Q}V_A V_A^T \hat{Q}^T.
$$

For any pair of matrices $G, H \in \mathbb{R}^{n \times mn}$, define an operator
$$
\mathcal{P}_\mathcal{L}(G, H) = (G - H) \mathcal{L}^\perp (H) \mathcal{L}(G),
$$
and note that $\mathcal{P}_\mathcal{L}(A, P) = R$, while
$$
\mathcal{P}_\mathcal{L}(QA\hat{Q}^T, QP\hat{Q}^T) = Q(A - P)\hat{Q}^T \mathcal{Q}(I - V_P V_P^T)\hat{Q}^T \hat{Q}V_A V_A^T \hat{Q}^T
$$
which implies that $R$ has the same distribution as $QR\hat{Q}^T$, and consequently the Frobenius norms of the rows of $R$ have the same distribution as those of $QR$, which proves our claim.

Combining these results, we see that
$$
\|R_{1,2}\|_{2 \to \infty} = O \left( \frac{n^{11/4} \log^{3/2}(mn)}{n^{3/4} \rho_n} \right),
$$
as required.

iii. Similarly to part i., we see that
$$
\|R_{1,3}\|_{2 \to \infty} \leq \|U_P\|_{2 \to \infty} \|U_P^T (A - P) V_P W \Sigma_A^{-1/2}\| F
\leq \|U_P\|_{2 \to \infty} \|U_P^T (A - P) V_P\| F \|W \Sigma_A^{-1/2}\| F
= O \left( m^{1/4} n^{-1/2} \cdot (m \log(n))^{1/2} \cdot (n \rho_n)^{-1/2} \right)
= O \left( \frac{m^{3/4} \log^{1/2}(n)}{n^{1/2} \rho_n} \right)
$$
by Proposition 18 and Corollary 14.

iv. Observe that
$$
\|R_{1,4}\|_{2 \to \infty} \leq \|R_{1,4}\|_F
\leq \|A - P\|_F \|V_P\|_F \|W \Sigma_A^{-1/2} - \Sigma_P^{-1/2} W\|_F
= O \left( \frac{m^3 \log^{3/2}(mn)}{n \rho_n} \right)
$$
by Propositions 12 and 22.
10.1. Proof of consistency in the two-to-infinity norm for the left and right embeddings

Theorem 1 (Consistency in the two-to-infinity norm). Let $F$ be a distribution on a subset $X$ of $\mathbb{R}^d$, and let $(A^{(1)}, \ldots, A^{(m)}, X) \sim \text{MRDPG}(F^n, A_1, \ldots, A_m)$. Then there exists a sequence of matrices $L(n) \in \text{GL}(d)$ such that

$$\|X_A L(n) - X\|_{2\rightarrow\infty} = O\left(\frac{m^{13/4} \log^{3/2}(mn)}{n^{1/2}\rho_n}\right)$$

almost surely. Moreover, for each $r \in \{1, \ldots, m\}$ with $\text{rank}(A_r) = d$, there exists a sequence $L^{(r)}(n) \in \text{GL}(d)$ such that

$$\|Y_{A_r} L^{(r)}(n) - X\|_{2\rightarrow\infty} = O\left(\frac{m^{13/4} \log^{3/2}(mn)}{n^{1/2}\rho_n}\right)$$

almost surely.

Proof. We first consider the left embedding $X_A$. Observe that

$$X_A - X_P W = U_A \Sigma_A^{1/2} - U_P \Sigma_P^{1/2} W$$

$$= U_A \Sigma_A^{1/2} - U_P U_P^T U_A \Sigma_A^{1/2} + U_P (U_P^T U_A \Sigma_A^{1/2} - \Sigma_P^{1/2} W)$$

$$= U_A \Sigma_A^{1/2} - U_P U_P^T U_A \Sigma_A^{1/2} + R_{1,1}.$$ 

Noting that

$$U_A \Sigma_A^{1/2} = A V_A \Sigma_A^{-1/2}, \quad U_P U_P^T P = P,$$

we see that

$$X_A - X_P W = AV_A \Sigma_A^{-1/2} - U_P U_P^T AV_A \Sigma_A^{-1/2} + R_{1,1}$$

$$= AV_A \Sigma_A^{-1/2} - PV_A \Sigma_A^{-1/2} - (U_P U_P^T AV_A \Sigma_A^{-1/2} - PV_A \Sigma_A^{-1/2}) + R_{1,1}$$

$$= (A - P) V_A \Sigma_A^{1/2} + U_P U_P^T (A - P) V_A \Sigma_A^{-1/2} + R_{1,1}$$

$$= (I - U_P U_P^T)(A - P) V_A \Sigma_A^{-1/2} + R_{1,1}$$

$$= (I - U_P U_P^T)(A - P)(V_P W + (V_A - V_P W)) \Sigma_A^{-1/2} + R_{1,1}$$

$$= (A - P) V_P W \Sigma_A^{-1/2} + R_{1,1} + R_{1,1} + R_{1,1} + R_{1,1}$$

$$= (A - P) V_P \Sigma_P^{-1/2} W + R_{1,1} + R_{1,1} + R_{1,1} + R_{1,1}.$$
Applying Proposition 23, we find that

$$\|X_A - X_P W\|_2 \rightarrow \infty = \|(A - P)V_P \Sigma_P^{-1/2}\|_2 \rightarrow \infty + O\left(\frac{m \log^{3/2}(mn)}{n^{3/4} \rho_n}\right).$$

Consequently,

$$\|X_A - X_P W\|_2 \rightarrow \infty \leq \sigma_d(P)^{-1/2}\|(A - P)V_P\|_2 \rightarrow \infty + O\left(\frac{m \log^{3/2}(mn)}{n^{3/4} \rho_n}\right).$$

Letting $v$ denote the $j$th column of $V_P$, we note that for any $i \in \{1, \ldots, n\}$ we have

$$(A - P)V_P)_{ij} = \sum_{k=1}^{mn} (A_{ik} - P_{ik})v_k$$

$$= \sum_{r=1}^{m} \sum_{1 \neq k \neq i} (A_{ik}^{(r)} - P_{ik}^{(r)})v_{(r-1)n+k} - \sum_{r=1}^{m} P_{ii}^{(r)}v_{(r-1)n+i}.$$ 

The latter term is $O(m\rho_n)$, while the former is a sum of independent zero-mean random variables satisfying

$$P \left( \left| \sum_{r=1}^{m} \sum_{k \neq i} (A_{ik}^{(r)} - P_{ik}^{(r)})v_{(r-1)n+k} \right| \geq t \right) \leq 2 \exp \left( -\frac{2t^2}{4 \sum_{r=1}^{m} \sum_{k \neq i} |v_{(r-1)n+k}|^2} \right) \leq 2 \exp \left( -\frac{t^2}{2} \right)$$

by Hoeffding’s inequality. Thus $((A - P)V_P)_{ij} = O\left(m + \log^{1/2}(n)\right)$ almost surely, and hence $\|(A - P)V_P\|_2 = O\left(m + \log^{1/2}(n)\right)$ almost surely by summing over all $j \in \{1, \ldots, d\}$. Taking the union bound over all $mn$ rows then shows that

$$\sigma_d(P)^{-1/2}\|(A - P)V_P\|_2 \rightarrow \infty = O\left(\frac{(m + \log^{1/2}(n))}{(n\rho_n)^{1/2}}\right),$$

and consequently that

$$\|X_A L_{(n)} - X\|_2 \rightarrow \infty = O\left(\frac{n^{13/4} \log^{3/2}(mn)}{n^{1/2} \rho_n}\right)$$

by multiplying on the right by $W^T L_X^{-1}$ and applying Corollary 16.

A similar argument is used for the right embedding. Suppose that $\text{rank}(A_r) = d$, and observe that

$$Y_A - Y_P W = V_A \Sigma_A^{1/2} - V_P \Sigma_P^{-1/2} W$$

$$= V_A \Sigma_A^{1/2} - V_P V_P^T V_A \Sigma_A^{1/2} + V_P (V_P^T V_A \Sigma_A^{1/2} - \Sigma_P^{1/2} W)$$

$$= V_A \Sigma_A^{1/2} - V_P V_P^T V_A \Sigma_A^{1/2} + R_{2,1}.$$
Noting that
\[ A^\top U_A \Sigma_A^{-1/2} = V_A \Sigma_A^{1/2}, \quad V_P V_P^\top = P^\top, \]
we see that
\[
Y_A - Y_P \mathbf{W} = V_A \Sigma_A^{1/2} - V_P V_P^\top V_A \Sigma_A^{1/2} + R_{2,1}
\]
\[ = A^\top U_A \Sigma_A^{-1/2} - V_P V_P^\top A^\top U_A \Sigma_A^{-1/2} + R_{2,1} \]
\[ = (A - P)^\top U_A \Sigma_A^{-1/2} - (V_P V_P^\top A^\top U_A \Sigma_A^{-1/2} - V_P V_P^\top P^\top U_A \Sigma_A^{-1/2}) + R_{2,1} \]
\[ = (A - P)^\top U_A \Sigma_A^{-1/2} - V_P V_P^\top (A - P)^\top U_A \Sigma_A^{-1/2} + R_{2,1} \]
\[ = (I - V_P V_P^\top) (A - P)^\top U_A \Sigma_A^{-1/2} + R_{2,1} \]
\[ = (I - V_P V_P^\top) (A - P)^\top (U_P \mathbf{W} + (U_A - U_P \mathbf{W})) \Sigma_A^{-1/2} + R_{2,1} \]
\[ = (A - P)^\top U_P W \Sigma_A^{-1/2} + R_{2,3} + R_{2,2} + R_{2,1} \]
\[ = (A - P)^\top U_P (\Sigma_P^{-1/2} \mathbf{W} + (W \Sigma_A^{-1/2} - \Sigma_P^{-1/2} \mathbf{W})) + R_{2,3} + R_{2,2} + R_{2,1} \]
\[ = (A - P)^\top U_P \Sigma_P^{-1/2} \mathbf{W} + R_{2,4} + R_{2,3} + R_{2,2} + R_{2,1}. \]

Applying Proposition 23 once more, we find that
\[ \|Y_A - Y_P \mathbf{W}\|_{2 \to \infty} = \|(A - P)^\top U_P \Sigma_P^{-1/2}\|_{2 \to \infty} + O \left( \frac{m^3 \log^{3/2}(mn)}{n^{3/4} \rho_n} \right). \]

Consequently,
\[ \|Y_{A,r} - Y_{P,r} \mathbf{W}\|_{2 \to \infty} \leq \sigma_d(P)^{-1/2} \|(A^{(r)} - P^{(r)}) U_P\|_{2 \to \infty} + O \left( \frac{m^3 \log^{3/2}(mn)}{n^{3/4} \rho_n} \right). \]

Letting \( u \) denote the \( j \)th column of \( U_P \), we note that for any \( i \in \{1, \ldots, n\} \) we have
\[ ((A^{(r)} - P^{(r)}) U_P)_{ij} = \sum_{k=1}^n (A_{ik}^{(r)} - P_{ik}^{(r)}) u_k \]
\[ = \sum_{k \neq i} (A_{ik}^{(r)} - P_{ik}^{(r)}) u_k - P_{ii}^{(r)} u_i. \]

The latter term is \( O(\rho_n) \), and applying Hoeffding as before shows that \( ((A^{(r)} - P^{(r)}) U_P)_{ij} = O \left( \log^{1/2}(n) \right) \) almost surely, and thus \( \|(A^{(r)} - P^{(r)}) U_P)\|_{2 \to \infty} = O \left( \log^{1/2}(n) \right) \) almost surely by summing over all \( j \in \{1, \ldots, d\} \). Taking the union bound over all \( n \) rows then shows that
\[ \sigma_d(P)^{-1/2} \|(A^{(r)} - P^{(r)}) U_P\|_{2 \to \infty} = O \left( \frac{\log^{1/2}(n)}{(n \rho_n)^{1/2}} \right), \]

49
and consequently that
\[
\|Y_{\Lambda,r}L_{(n)}^{(r)} - X\|_{2\rightarrow\infty} = O\left(\frac{m^{1/4} \log^{3/2}(mn)}{n^{1/2} \rho_n}\right)
\]
by multiplying on the right by $W^T L^{-1}_{X,Y}$ and applying Corollary 16. \hfill \Box

10.2. Proof of the Central Limit Theorem for the left and right joint adjacency spectral embeddings

**Theorem 2** (Central Limit Theorems). With the same notation as Theorem 1, let $\xi_1, \ldots, \xi_n, \xi \sim F, \Delta = \mathbb{E}[\xi^T] \in \mathbb{R}^{d \times d}, \hat{\Delta} = I_n \otimes \Delta \in \mathbb{R}^{md \times md}$, and suppose that the sparsity factor $\rho_n$ satisfies $\rho_n = o\left(\log^{3/2}(n)\right)$, $n \rightarrow \infty$. Given $x \in \mathcal{X}$ and $r \in \{1, \ldots, m\}$, let
\[
\Sigma_r(x) = \begin{cases} 
\mathbb{E}[X^T \Lambda_r(1 - X^T \Lambda_r \xi \xi^T)] & \text{if } \rho_n = 1 \\
\mathbb{E}[X^T \Lambda_r \xi \xi^T] & \text{if } \rho_n \rightarrow 0
\end{cases}
\]
and let $\tilde{\Sigma}(x) = \text{diag}(\Sigma_1(x), \ldots, \Sigma_m(x))$. Then, for all $z \in \mathbb{R}^d$ and for any fixed $i$,
\[
P\left(n^{1/2} (X_{\Lambda} L_{n} - X)^T i \leq z \mid \xi_i = x \right) \rightarrow \Phi \left(z, \Delta_{\Lambda} \tilde{\Sigma}(x) \Delta_{\Lambda}^T\right),
\]
almost surely and, if $\text{rank}(\Lambda_r) = d$,
\[
P\left(n^{1/2} Y_{\Lambda,r} L_{(n)}^{(r)} - X)^T i \leq z \mid \xi_i = x \right) \rightarrow \Phi \left(z, \Delta_r \tilde{\Sigma}_r(x) \Delta_r^T\right),
\]
almost surely, where $\Delta_{\Lambda} = (\Lambda \hat{\Delta} \Lambda^T)^{-1} \Lambda$ and $\Delta_r = (\Lambda_r \Delta_{\Lambda}^T)^{-1} \Lambda_r$.

**Proof.** We first consider the left embedding $X_{\Lambda}$. From the proof of Theorem 1, we see that
\[
n^{1/2} (X_{\Lambda} W^T L_{X}^{-1} - X) = n^{1/2} (A - P) V_p \Sigma_p^{-1/2} L_{X}^{-1} + n^{1/2} R,
\]
where $\|n^{1/2} R\|_{2\rightarrow\infty} \rightarrow 0$ by Proposition 23 and our assumption about the growth of the sparsity factor.

Now,
\[
(A - P) V_p = \sum_{r=1}^m (A^{(r)} - P^{(r)}) V_{p,r},
\]
and the matrices $L_{Y,r}$ were chosen so that
\[
XL_{Y,r} = Y_{p,r} = V_{p,r} \Sigma_p^{1/2},
\]
and so
\[
V_{p,r} \Sigma_p^{-1/2} = XL_{Y,r} \Sigma_p^{-1}.
\]
Thus
\[
n^{1/2} (A - P) V_p \Sigma_p^{-1/2} L_{X}^{-1} = n^{1/2} \sum_{r=1}^m (A^{(r)} - P^{(r)}) XL_{Y,r} \Sigma_p^{-1} L_{X}^{-1}.
\]
Consequently,

\[ n^{1/2} \left( (A - P)V_P \Sigma_P^{-1/2} L_X^{-1} \right)_{i}^\top = n^{1/2} \sum_{r=1}^{m} (L_{Y,r} \Sigma_P^{-1} L_X^{-1})^\top \left( (A^{(r)} - P^{(r)})X \right)_{i}^\top \]

\[ = \sum_{r=1}^{m} (n L_{Y,r} \Sigma_P^{-1} L_X^{-1})^\top \left[ n^{-1/2} \sum_{j=1}^{n} (A_{ij}^{(r)} - P_{ij}^{(r)})X_j \right] \]

\[ = \sum_{r=1}^{m} (n \rho_n L_{Y,r} \Sigma_P^{-1} L_X^{-1})^\top \left[ (n \rho_n)^{-1/2} \sum_{j \neq i} (A_{ij}^{(r)} - P_{ij}^{(r)})\xi_j \right] \]

\[ - \sum_{r=1}^{m} (n \rho_n L_{Y,r} \Sigma_P^{-1} L_X^{-1})^\top \left[ (n \rho_n)^{-1/2} P_{ii}^{(r)} \xi_i \right]. \]

The latter term satisfies

\[ \left\| \sum_{r=1}^{m} (n \rho_n L_{Y,r} \Sigma_P^{-1} L_X^{-1})^\top \left[ (n \rho_n)^{-1/2} P_{ii}^{(r)} \xi_i \right] \right\|_{2\rightarrow\infty} \leq \sum_{r=1}^{m} \left\| (n \rho_n L_{Y,r} \Sigma_P^{-1} L_X^{-1})^\top \left[ (n \rho_n)^{-1/2} P_{ii}^{(r)} \xi_i \right] \right\| = O \left( m(n \rho_n)^{-1/2} \right) \]

almost surely.

Now, conditional on \( \xi_i = x \), we have \( P_{ii}^{(r)} = \rho_n x^\top \Lambda_r \xi_j \), and so

\[ (n \rho_n)^{-1/2} \sum_{j \neq i} (A_{ij}^{(r)} - P_{ij}^{(r)})\xi_j \]

is a scaled sum of \( n - 1 \) independent, zero-mean random variables, each with covariance matrix given by

\[ \Sigma_r(x) = \mathbb{E} [x^\top \Lambda_r \xi (1 - \rho_n x^\top \Lambda_r \xi) \xi^\top] \]

which implies (by the multivariate central limit theorem) that

\[ n^{-1/2} \sum_{j \neq i} (A_{ij}^{(r)} - P_{ij}^{(r)})X_j \to \mathcal{N}(0, \Sigma_r(x)). \]

Finally, we consider the terms \( (n \rho_n L_{Y,r} \Sigma_P^{-1} L_X^{-1})^\top \). By Proposition 17 we know that \( L_{Y,r} \Sigma_P^{-1} L_X^{-1} = \Lambda_r (\Lambda Z Z^\top)^{-1}, \) and so

\[ (n \rho_n L_{Y,r} \Sigma_P^{-1} L_X^{-1})^\top = n(\rho_n^{-1} \Delta Z^\top Z \Lambda^\top)^{-1} \Lambda_r \to (\Delta \hat{\Delta} \Lambda^\top)^{-1} \Lambda_r \]

almost surely by the law of large numbers.

Combining all this, we find that

\[ n^{-1/2} \left( X \Delta W^\top L_X^{-1} - X \right)_{i}^\top \to \mathcal{N} \left( 0, \Delta \hat{\Sigma}(x) \Delta \hat{\Lambda} \right) \]
Similarly, for the right embedding we observe that

\[ n^{1/2}(Y_{A,r}W^T L_{Y,r}^{-1} - X) = n^{1/2}(A^{(r)} - P^{(r)}) U_P \Sigma_p^{-1/2} L_{Y,r}^{-1} + n^{1/2}R, \]

where again \( \|n^{1/2}R\|_2 \to 0. \)

The matrix \( L_X \) satisfies

\[ U_P \Sigma_p^{-1/2} = XL_X \Sigma_p^{-1} \]

and so

\[ n^{1/2}(A^{(r)} - P^{(r)}) U_P \Sigma_p^{-1/2} L_{Y,r}^{-1} = n^{1/2}(A^{(r)} - P^{(r)}) XL_X \Sigma_p^{-1} L_{Y,r}^{-1}. \]

Consequently,

\[
\begin{align*}
    n^{1/2} \left[ (A^{(r)} - P^{(r)}) U_P \Sigma_p^{-1/2} L_{Y,r}^{-1} \right]_i^T &= n^{1/2} (L_X \Sigma_p^{-1} L_{Y,r}^{-1})^T \left[ (A^{(r)} - P^{(r)}) X \right]_i^T \\
    &= (nL_X \Sigma_p^{-1} L_{Y,r}^{-1})^T \left[ n^{-1/2} \sum_{j=1}^{n} (A^{(r)}_{ij} - P^{(r)}_{ij}) X_j \right] \\
    &= (n \rho_n L_X \Sigma_p^{-1} L_{Y,r}^{-1})^T \left[ (n \rho_n)^{-1/2} \sum_{j \neq i} (A^{(r)}_{ij} - P^{(r)}_{ij}) \xi_j \right] \\
    &- (n \rho_n L_X \Sigma_p^{-1} L_{Y,r}^{-1})^T \left[ (n \rho_n)^{-1/2} P^{(r)}_{ii} \xi_i \right].
\end{align*}
\]

The latter term satisfies

\[ \left\| (n \rho_n L_X \Sigma_p^{-1} L_{Y,r}^{-1})^T \left[ (n \rho_n)^{-1/2} P^{(r)}_{ii} \xi_i \right] \right\|_{2 \to \infty} \leq \left\| (n \rho_n L_X \Sigma_p^{-1} L_{Y,r}^{-1})^T \left[ (n \rho_n)^{-1/2} P^{(r)}_{ii} X_i \right] \right\| = O \left( (n \rho_n)^{-1/2} \right) \]

almost surely.

From the proof of the left embedding, we know that, conditional on \( \xi_i = x, \)

\[ (n \rho_n)^{-1/2} \sum_{j \neq i} (A^{(r)}_{ij} - P^{(r)}_{ij}) \xi_j \to N(0, \Sigma_r(x)), \]

so we need only evaluate the term \( (n \rho_n L_X \Sigma_p^{-1} L_{Y,r}^{-1})^T \). From Proposition 17 we know that \( L_X \Sigma_p^{-1} L_{Y,r} = (X^\top X)^{-1} \Lambda_r^{-1} \), and thus

\[ (n \rho_n L_X \Sigma_p^{-1} L_{Y,r}^{-1})^T = n \Lambda_r^{-1} (\rho_n^{-1} X^\top X)^{-1} \to \Lambda_r^{-1} \Delta^{-1} = (A_r \Delta A_r^\top)^{-1} A_r \]

almost surely.
Combining all this, we find that

$$n^{1/2} \left( Y_{A_r} W_{Y_r}^T L_{Y_r}^{-1} - X \right)^T \to N \left( 0, \Delta_r \Sigma_r (x) \Delta_r^T \right)$$

almost surely, where $\Delta_r = (A_r \Delta A_r^T)^{-1} A_r$, from which we deduce the Central Limit Theorem by integrating over all possible values of $x \in \mathcal{X}$. \hfill \Box