Using Ideas from Network Theory to Detect Concept Drift

Ezra Nwobodo

Level 6 Project - 20cp

Supervised by Daniel Lawson

School of Mathematics

University of Bristol

29/04/2024

Acknowledgement of Sources

For all ideas taken from other sources (books, articles, internet), the source of the ideas is mentioned in the main text and fully referenced at the end of the report.

All material which is quoted essentially word-for-word from other sources is given in quotation marks and referenced.

Pictures and diagrams copied from the internet or other sources are labelled with a reference to the web page or book, article etc.

Signed Ezra Nwobodo

Date 04/29/2024

Contents

1. Introduction

This report is on a novel method of detecting a phenomenon called "concept drift" called unfolded adjacency spectral embedding (UASE), which was introduced by A. Jones and P. Rubin-Delanchy [1] as a method of embedding complex graphs. Using UASE for this purpose is heavily inspired by two papers by I. Gallagher et al.[2][3]. The first uses UASE in the context of a dynamic latent position model, a sequence of unweighted graphs, and the second uses it for a single weighted graph (the specifics are discussed in Section 4). The aim here is to model the relationship between features or covariates changing over time as a weighted dynamic latent position model, a sequence of weighted graphs, and combine the results of those two papers. Obtaining useful insights and properties from this model involves defining a new class of model called a weighted multilayer random dot product graph, which is some weighted extension of the multilayer random dot product graph defined in the A. Jones and P. Rubin-Delanchy paper [1].

We first explain what concept drift is, why it's important, and what problems this method aims to solve (Section 1). We then describe the method itself (Section 2) and apply it to a toy model to illustrate its properties (Section 3) before going into detail on the theory behind model and proving some important properties (Section 4).

1.1 What is concept drift and why is it important?

In many data science and machine learning contexts, data is collected over time in order to be analysed. A model will often learn from historic data for the purpose of prediction (for example, fraud detection [4], targeted advertising algorithms [5]). Concept drift refers to the phenomenon of data evolving in such a way that its distribution changes which can cause previous models to become deprecated and lose their predictive power [6].

A. Suárez-Cetrulo et al. [7] illustrate concept drift in a very simple way in the context of binary classification in Figure 1. If the data points are on one side of the decision boundary, then they are classified as being in class A, with the same for points on the other side being in class B. We can see that even though the points themselves are the same, some points have changed class. This means that the true decision boundary has changed and some form of concept drift has occurred.

Failure to detect this drift and adapt to the change in time can be potentially costly (in Figure 1, the model trained on historic data would incorrectly classify some data points). An example of this cost is detecting concept drift in the stock market [8] where you want to ensure that your trading strategy is still effective in an extremely volatile environment.

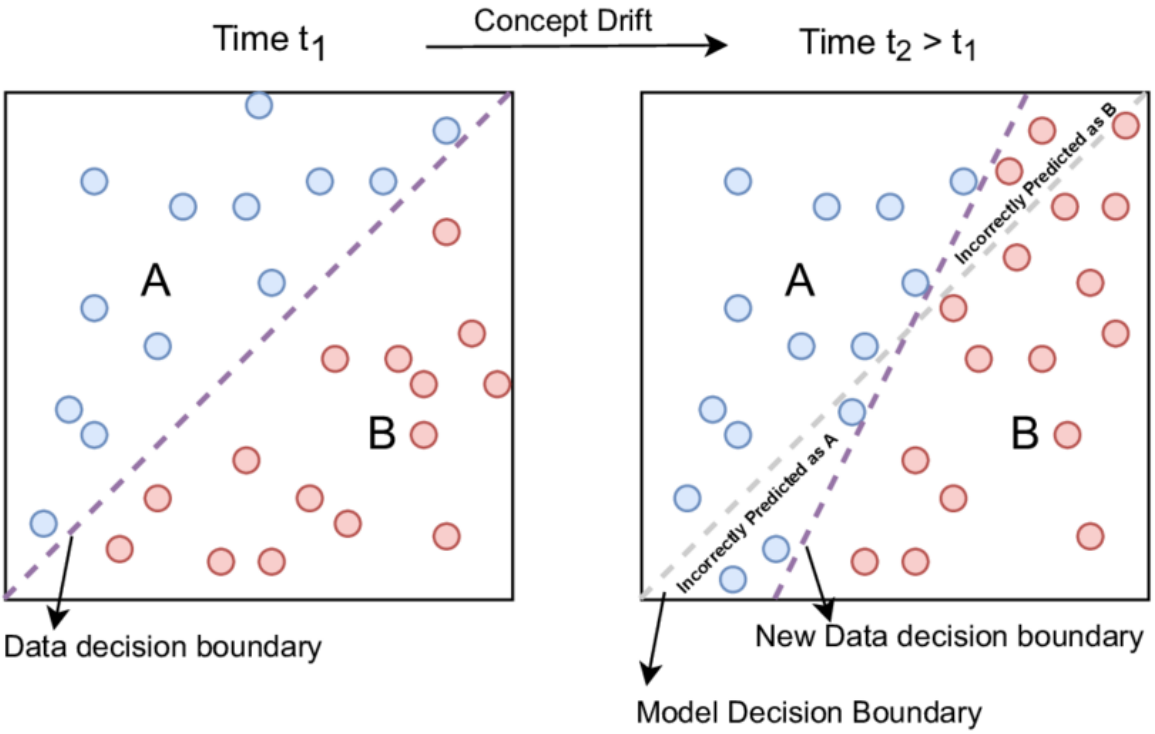

Figure 1: Illustrative example of a concept drift [7].

The focus of this report will be on detecting concept drift in the context of unsupervised learning. So we are only interested in analyzing data with no associated response variable. This is also called covariate drift [9].

We observe virtual (unsupervised) concept drift as defined by G. Webb et al. [9] for data/observations $\mathbf{X} \in \mathbb{R}^{n \times p}$ if, between times t_0 and t_1, the probability of observing that data changes:

$$\mathbb{P}_{t_0}(\mathbf{X}) \neq \mathbb{P}_{t_1}(\mathbf{X}).$$

1.1.1 Difficulties

Detecting unsupervised concept drift comes with some difficulties compared with its supervised counterpart.

- Since there is no response variable Y there is a lack of a clear signal. In the supervised context, it is common to detect drift through monitoring prediction error (comparing a fitted value to the true value) or monitoring some form of score function as shown by K. Zhang et al. [10].
- Since there's no obvious feedback on the performance of a model, it is difficult for it to adapt to concept drift. This is discussed by Hoens et al. [11] where they note that one can observe the distribution of observations changing using some form of error metric, while the relationship between the features themselves stay the same.

On that last point, Hoens et al. state that "in virtual concept drift, while the distribution

of instances may change (corresponding to a change in the class priors or the distribution of the classes), the underlying concept (i.e., the posterior distribution) does not. This may cause problems for the learner, as such changes in the probability distribution may change the error of the learned model, even if the concept did not change".

1.1.2 Complications from high-dimensional data

We'll also be focusing on contexts where our data is high-dimensional, which comes with its own challenges.

- The difficulties associated with high dimensional data is often referred to as the 'curse of dimensionality' [12][13]. As the number of features of a data set increases, the data becomes more sparse which makes it harder for a model to generalise accurately and detect meaningful patterns [14]. It can also make it more difficult to detect shifts in distributions as subtle changes in some dimensions might be significant but hard to detect over the noise from the more irrelevant features.
- Any results you do get are more difficult to understand or interpret in the context of all the other features as visualisation is much more difficult [15]. You could calculate the one dimensional drift of every feature and compare them, but that may not tell the whole story.

1.2 Existing methods

There already exists methods for detecting concept drift that do well when, for example, you want to determine whether drift has occurred at all or to determine which covariates have drifted the most. G. Webb et al. [16] attempt to categorise many different types of concept drift while discussing methods associated with them. I. Goldenberg et al. [17] survey the use of various distance measures to quantify just how much the distribution of the features has changed for numeric data, recommending the use of Hellinger distance to "measure dissimilarity between distributions for univariate or low-dimensional data".

But at present, there is a lack of interpretable methods for visualising the drift between all covariates at the same time, especially with high dimensional data. In the same paper, I. Goldenberg et al. remark: "There is a paucity of measures that work for high-dimensional numerical data.", noting that a PCA-based approach shows some promise and encouraging more work in this direction.

2. Framing the Problem

Before being able to tackle this problem effectively, we must frame it with a suitable model. Since our aim is to track the relationship between the covariates over time, we can represent the covariates as nodes of a weighted network with the arcs connecting them

representing this relationship.

Covariance is used to represent this relationship throughout but statistical discrepancies such as Wasserstein distance and many others can also be effective here. The exact properties required will be discussed later.

So we have a sequence of weighted networks represented by adjacency matrix $\mathbf{A}^{(t)} \in \mathbb{R}^{n\times n}$ with n nodes (covariates) at time point t.

$\mathbf{A}^{(t)}_{ij} = Cov(\mathbf{c}_i, \mathbf{c}_j)$ represents the relationship between covariates $\mathbf{c}_i$ and $\mathbf{c}_j$ at time t with $\mathbf{c}_1, \ldots, \mathbf{c}_n \in \mathbb{R}^m$, where m is the number of observations of data.

Before providing the method for performing UASE on the sequence of adjacency matrices, we'll briefly discuss the singular value decomposition(SVD).

2.1 Singular Value Decomposition

A method central to everything in this report is the singular value decomposition [18]. For any $\mathbf{A} \in \mathbb{R}^{m\times n}$, you can rewrite it in the form

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

where $\mathbf{U} \in \mathbb{R}^{m\times m}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$ represent the matrices of left and right singular vectors while $\mathbf{\Sigma} \in \mathbb{R}^{m\times n}$ represents the matrix of singular values.

What exactly these singular values and vectors are will not be discussed here [19], but they have a couple of important properties that are of importance. First is that the number of non-zero singular values is equal to the rank of $\mathbf{A}$. So a matrix is degenerate, its matrix of singular values from the SVD will have some zeros on its diagonal (those rows can be discarded).

Secondly, the singular vectors form two sets of orthonormal bases. So each singular vector in a given set is orthogonal and, if $\mathbf{u}_i$ and $\mathbf{v}_i$ represent the ith singular vectors of each set, we have

$$\sum_{i=1}^{m} \|\mathbf{u}_i\|^2 = 1 \text{ and } \sum_{i=1}^{n} \|\mathbf{v}_i\|^2 = 1,$$

where $\|\cdot\|$ is the Euclidean norm.

2.2 Unfolded Adjacency Spectral Embedding

First, define the *unfolding* $\mathbf{A} = [\mathbf{A}^{(1)}|\cdots|\mathbf{A}^{(T)}] \in \mathbb{R}^{n\times nT}$, which is a column concatenation of the matrices. Then compute the truncated singular value decomposition on this unfolding as:

$$\mathbf{A} \approx \mathbf{U}_\mathbf{A}\mathbf{\Sigma}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top,$$

where $\mathbf{\Sigma_A} \in \mathbb{R}^{d\times d}$ is diagonal and contains the d largest singular values of $\mathbf{A}$, and $\mathbf{U_A} \in \mathbb{O}^{n\times d}$ and $\mathbf{V_A} \in \mathbb{O}^{nT\times d}$ contain the left and right singular vectors respectively.

This representation of $\mathbf{A}$ can be broken down as follows:

$$\mathbf{A} \approx \mathbf{U_A}\mathbf{\Sigma_A}\mathbf{V_A}^\top = \mathbf{U_A}\mathbf{\Sigma_A}^{1/2}\mathbf{\Sigma_A}^{1/2}\mathbf{V_A}^\top = \hat{\mathbf{X}}\hat{\mathbf{Y}}^\top.$$

Take $\hat{\mathbf{Y}} = \mathbf{V_A}\mathbf{\Sigma_A}^{1/2} \in \mathbb{R}^{nT\times d}$ which represents our embedding and partition $\hat{\mathbf{Y}}$ into T blocks $\hat{\mathbf{Y}}^{(t)}$ (row-wise). Then the rows $\hat{\mathbf{Y}}_i$ of each block are the d-dimensional points to be plotted and is an estimate of the latent position of covariate i at time t. The notion of a latent space will be discussed later.

3. Toy Model

To make some of these ideas concrete, we'll demonstrate them on a toy model with ideal conditions and compare UASE to other methods such as independent embedding.

3.1 Dynamic Stochastic Block Model

Continuing on the theme of stealing ideas from network theory, our toy model will be an extension of the dynamic stochastic block model [20][21]. It involves a set of K communities (4 in our case), where a node from community i will be "connected" to a node from community j with probability p.

In our case, 'connected' will be refer to 'degree of correlation'. If two nodes (covariates) are connected, they will be more correlated with each other. Setting up a model with these properties is surprisingly involved.

3.2 Setting up the model

We have our matrices of connection probabilities between covariates at times 1 and 2:

$$\mathbf{B}^{(1)} = \begin{pmatrix} 0.24 & 0.06 & 0.54 & 0.06 \\ 0.06 & 0.60 & 0.12 & 0.06 \\ 0.54 & 0.12 & 0.06 & 0.10 \\ 0.06 & 0.06 & 0.10 & 0.18 \end{pmatrix}, \ \mathbf{B}^{(2)} = \begin{pmatrix} 0.48 & 0.48 & 0.12 & 0.06 \\ 0.48 & 0.48 & 0.12 & 0.06 \\ 0.12 & 0.12 & 0.27 & 0.10 \\ 0.06 & 0.06 & 0.10 & 0.18 \end{pmatrix}.$$

Now, we want to generate a data frame where each covariate $\mathbf{c}_i$ belongs to exactly one of the 4 communities. For example, a covariate in community 4 will be connected with a covariate in community 2 with probability 0.3.

At $t = 2$ we see that communities 1 and 2 have the exact same connection probabilities with every other community, so they have merged between $t = 1$ and $t = 2$ and they

should be indistinguishable in our embedding.

Similarly, community 4 has the same connection probabilities at both time points.

3.3 Relation to detecting concept drift

From this, we want to demonstrate two properties of the UASE method:

(i) Cross-sectional stability: The embeddings for communities 1 and 2 at time 2 are close,

(ii) Longitudinal stability: The embeddings for community 4 at times 1 and 2 are close.

These properties are desirable in the context of detecting concept drift as we ideally want the embedding of a covariate to remain in the same place if no concept drift has occurred (i.e. the distribution of the connection probabilities remain constant). We similarly want the embedding of two covariates to be in a similar place if they have the same distribution at any given time point.

3.4 Generating our simulated data

In our case, two covariates being 'connected' refers to correlation. So we need to generate synthetic data where two covariates are highly correlated when they are connected and independent otherwise.

This is done by first generating the observed matrices of connections based on the connection probability matrices. We then independently sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from a $\mathcal{N}(0, \sigma^2)$ distribution, and let the ith component of observation k be the linear combination of samples corresponding with the covariates connected with covariate i.

For example, if $\mathbf{c}_1$ is connected with $\mathbf{c}_2$ and $\mathbf{c}_5$, then the first component of the first observation (denoted $\mathbf{D}_{1,1}$) will be $\mathbf{x}_2 + \mathbf{x}_5$. It follows that if two covariates aren't connected, they will have no samples in common and thus be independent.

This process is detailed in Algorithm 1.

Algorithm 1 Generate simulated data

Input: Connection matrices $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$
Output: Data frames $\mathbf{D}^{(1)}$, $\mathbf{D}^{(2)} \in \mathbb{R}^{m \times n}$

```
Divide covariates into k = 4 communities
Use connection probabilities to simulate adjacency matrix C^(1) where C^(1)_ij = 1 if and
only if c_i and c_j are connected
for observation k ∈ [m] do
    Simulate x ∈ R^n with x_1, ..., x_n ~iid N(0, σ^2)
    for i ∈ [n] do
        D^(t)_{k,i} = Σ_j x_j for j where c_i and c_j are connected
    end for
end for
```

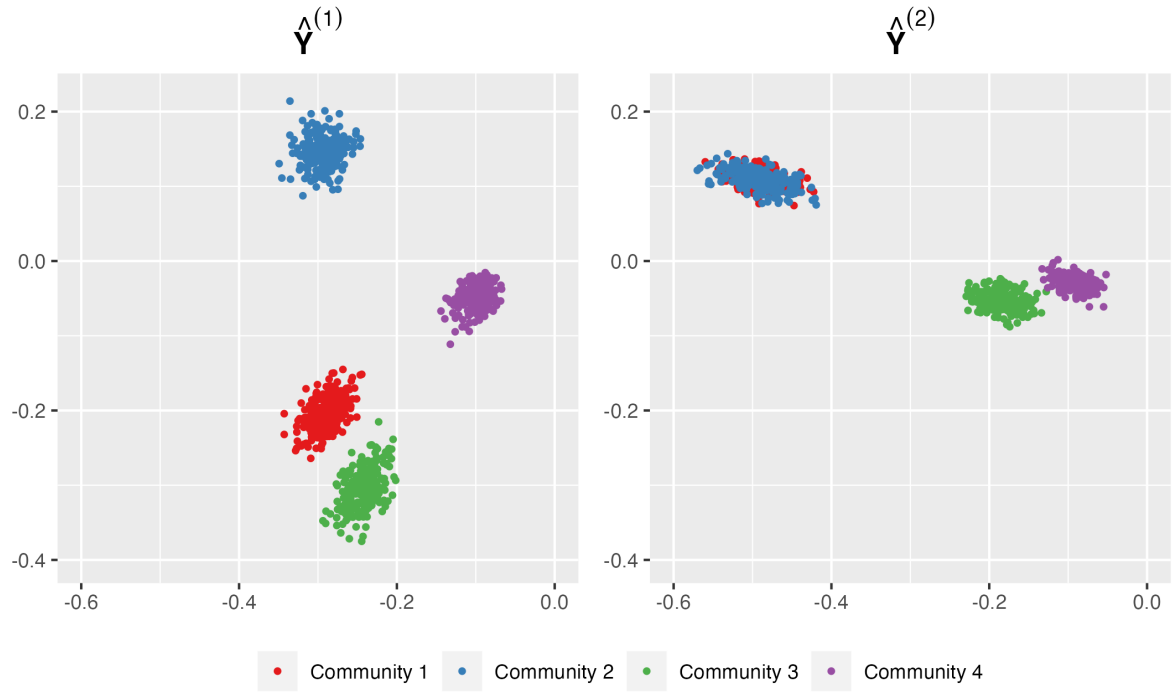

Figure 2: Plot showing the stability of the UASE method at different time points, with $n = 300$ and $m = 800$.

3.5 Computing the embedding

Once we have our data, we can compute the embeddings associated with it using UASE.

Since the relationship between our covariates is being represented with covariance, each adjacency matrix $\mathbf{A}^{(t)}$ will be the covariance matrix of the data at time t. Under this model, a covariate is never connected to itself, so we want each diagonal element to be zero (so the matrix is hollow).

So we let $\mathbf{A}^{(1)} = \text{Hollow}(\text{Cov}(\mathbf{D}^{(1)}))$ and $\mathbf{A}^{(2)} = \text{Hollow}(\text{Cov}(\mathbf{D}^{(2)}))$ be our adjacency matrices. Then column concatenate $\mathbf{A} = [\mathbf{A}^{(1)}|\mathbf{A}^{(2)}] \in \mathbb{R}^{n\times 2n}$ which is the unfolding. Compute the truncated (d-dimensional) SVD on this unfolding to obtain the embedding:

$$\mathbf{A} \approx \mathbf{U}_\mathbf{A}\mathbf{\Sigma}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top = \hat{\mathbf{X}}\hat{\mathbf{Y}}^\top.$$

Finally, plot the first 2 entries for each row of $\hat{\mathbf{Y}}^{(1)}$ and $\hat{\mathbf{Y}}^{(2)}$.

3.6 Results

In Figure 2, we can see the properties we the properties we set out to show, with $n = 300$ and $m = 800$. Community 4 seems to have only moved very slightly while communities 1 and 2 have overlapped entirely.

We contrast this with the method of computing the SVD of each individual covariance matrix which we can see in Figure 3. While it displays cross-sectional stability in that com-

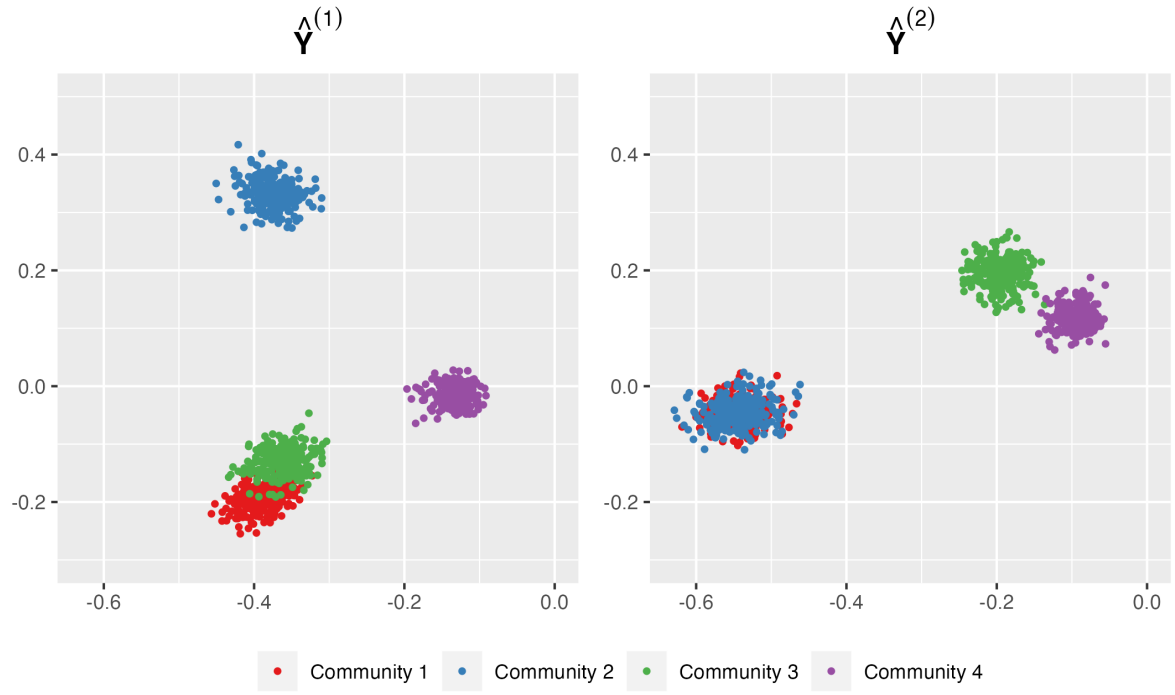

Figure 3: Plot showing the contrasting non stability of embedding each adjacency matrix independently, with $n = 300$ and $m = 800$.

munities 1 and 2 are overlapping, it doesn't display longitudinal stability as community 4 has moved considerably. If we are attempting to detect concept drift, we wouldn't want to flag the distribution of community 4 as having changed substantially when it hasn't.

3.7 What's next?

We have discussed the method itself, but in order to prove these stability properties we'll need a richer model. For example, under certain assumptions, we can prove:

- The cross sectional, and longitudinal stability of UASE (Corollary 8);
- That each embedding $\hat{\mathbf{Y}}^{(t)}$ converges to some underlying 'true' or 'noise-free' $\tilde{\mathbf{Y}}^{(t)}$ as the size of n increases, up to rotation (Theorem 5);
- That the point clouds converge to a multivariate Gaussian distribution with n (Theorem 6).

Figure 4 provides a good illustration of these properties. It uses the same toy model, but with $n = 1200$ and $m = 2500$ representing a much larger network. The Gaussian clusters that we observed before have become much tighter around one singular point, which we refer to as the noise-free embedding. This idea is expanded on in the next section.

We show that our adjacency matrices follow an extension of a model called the multilayer random dot product graph (MRDPG) [1].

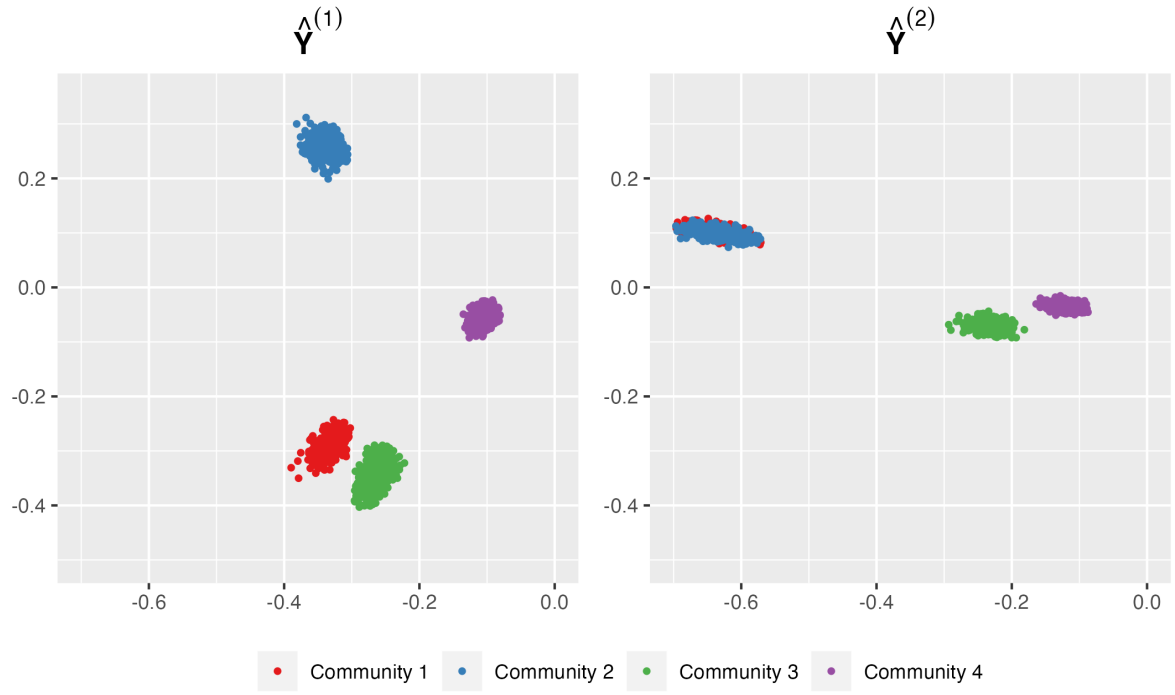

Figure 4: Plot showing the convergence of point clouds to a multivariate Gaussian distribution centred at the 'noise free' embeddings, with $n = 1200$ and $m = 2500$.

4. Theory

The stability of the UASE method applied to networks following a dynamic latent position model was proven by I. Gallagher et al. [2]. The aim here is to prove a weighted version of the same thing where each network follows a weighted generalised random dot product graph model [3].

They proved some asymptotic properties of the model by showing that it follows a MRDPG. However, for our use case, we must show that our model follows some weighted version of this (it is called a weighted multilayer random dot product graph (WMRDPG) here for convenience, but it's not a full extension of the MRDPG as bipartite nor directed graphs are considered).

In addition, despite this model being used in the context of detecting concept drift, it's applications are a lot broader and the majority of the remainder on this report will be more focused on the theory surrounding this.

4.1 Weighted dynamic latent position model

In order to show that our model follows a WMRDPG, we must make some assumptions on the form of each adjacency matrix and the asymptotic properties of the sequence of adjacency matrices. These assumptions form what's called a weighted dynamic latent

position model, a combination of the dynamic latent position model [22], and the weighted generalised random dot product graph model.

4.1.1 Latent space

First we must understand the concept of a latent space [23]. The toy model is very helpful for this since, despite our data being n dimensional (n covariates) we say that its behaviour can be explained by some lower d dimensional latent variable. This is also how we decide how many dimensions to reduce in the truncated SVD.

For example, consider data that measures students' results in a large number of different subjects and sports events. Many of these results will be correlated with each other since a student scoring high in science will usually score high in maths too.

The latent variables in this scenario would be communities like intelligence, athletic ability, household income etc. and we attempt to model how these communities move in the latent space over time. In the toy model, we could say that the latent space is the set $\mathcal{Z} \times \mathcal{Z}$, where $\mathcal{Z} = \{1, 2, 3, 4\}$.

We assume that there exists a map $\phi_t : \mathcal{Z} \to \mathbb{R}^{D_t}$ for each t where D_t is the dimension of the embedding $\hat{\mathbf{Y}}^{(t)}$ (which can change over time). D_t should be much smaller than n.

We are comfortable in making this assumption as it is equivalent to the manifold hypothesis [24] which states that "many high-dimensional data sets that occur in the real world actually lie along low-dimensional latent manifolds inside that high-dimensional space"[1]. An example of this hypothesis being studied in a similar context to ours can be found in a paper by N. Whiteley et al. [25]. They discuss how a low dimensional manifold structure embedded in a higher dimensional space emerges from the Latent Metric Model when using SVD, along with some other methods and give a statistical justification for this.

Finally, we impose some asymptotic properties on our model to allow us to prove stability properties later.

Definition 1 (Weighted dynamic latent position model)**.** *Let* $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(T)} \in \mathbb{R}^{n \times n}$ *be a sequence of symmetric adjacency matrices where*

$$\mathbf{A}_{ij}^{(t)} \overset{ind}{\sim} H(\mathbf{Z}_i^{(t)}, \mathbf{Z}_j^{(t)}),$$

for $1 \leq i < j \leq n$*, and* $t \in [T]$ *where* $\mathbf{Z}_i^{(t)} \in \mathbb{R}^k$ *represents the unknown position of node* i *at time* t*.* $\{H(\mathbf{z}_1, \mathbf{z}_2) : \mathbf{z}_1, \mathbf{z}_2 \in \mathcal{Z}\}$ *is a family of real valued distributions with the following properties, for all* $\mathbf{z}_1, \mathbf{z}_2 \in \mathcal{Z}$*:*

- *(Symmetry)* $H(\mathbf{z}_1, \mathbf{z}_2) = H(\mathbf{z}_2, \mathbf{z}_1)$.
- *(Low-rank expectation) There exists a map* $\phi_t : \mathcal{Z} \to \mathbb{R}^{D_t}$ *such that, if* $\mathbf{A}^{(t)} \sim$

[1] https://en.wikipedia.org/wiki/Manifold_hypothesis

$H(\mathbf{z}_1, \mathbf{z}_2)$

$$\mathbb{E}(\mathbf{A}^{(t)}) = \phi(\mathbf{z}_1)^\top \mathbf{I}_{p_t,q_t} \phi(\mathbf{z}_2),$$

where $\mathbf{I}_{p,q} = diag(1, \ldots, 1, -1 \ldots -1)$ *is a diagonal matrix with* p_t *ones followed by* q_t *negative ones, with* $p_t + q_t = D_t$.

We say that for each node there is an associated *latent position sequence* $(\mathbf{Z}_i)_{t\in[T]}$ which is distributed according to a joint distribution $\mathcal{F}$ on $\mathcal{Z}^T$ where $\mathcal{Z}$ is some bounded sample space and each sequence $(\mathbf{Z}_i)_{t\in[T]}$ is independent.

For the purpose of making more precise statements on some asymptotic results related to UASE, we make some further assumptions:

- (Bounded expectation) There exist universal constants $a, b \in \mathbb{R}$ such that $\mathbb{E}(\mathbf{A}) \in [a, b]$.
- (Full rank marginal second moment matrices) For $\mathbf{Z} \sim \mathcal{F}$ with $\mathbf{Z} = [\mathbf{Z}^{(1)}| \cdots |\mathbf{Z}^{(T)}]$, each component of the random vector $\xi = [\xi_1| \cdots |\xi_T]$ with $\xi_t = \phi_t(\mathbf{Z}^{(t)})$ has second moment matrix

  $$\Delta_t = \mathbb{E}[\xi_t \xi_t^\top] \in \mathbb{R}^{D_t \times D_t}$$

  with rank D_t. Let $\mathcal{F}^*$ be the joint distribution such that $\xi \sim \mathcal{F}^*$ with marginal distributions $\mathcal{F}_1^*, \ldots, \mathcal{F}_T^*$ such that $\xi_t \sim \mathcal{F}_t^*$.[2]
- (Exponential tails) There exist universal constants $\alpha > 0$ and $\beta_\rho > 0$ for each $\rho \in \mathbb{R}$, such that

  $$\mathbb{P}[|\mathbf{A}| < \beta_\rho log^\alpha(t)] > 1 - t^\rho.$$

4.2 Matrices follow a WMRDPG

We are now able to show that a sequence of graphs satisfying the above properties follow a WMRDPG.

The main feature of this model is the existence of random matrices $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $\mathbf{Y}^{(t)} \in \mathbb{R}^{n\times d_t}$ and matrices $\mathbf{\Lambda}^{(t)} \in \mathbb{R}^{d\times d_t}$ which determine the distribution of the expected value of the graphs as $\mathbb{E}[\mathbf{A}_{ij}^{(t)}] = \mathbf{X}_i \mathbf{\Lambda}^{(t)} \mathbf{Y}_j^{(t)\top}$.

The most important thing to note is that $\mathbf{X}$ is the same for each matrix. It acts as an 'anchor', providing stability for the method and allowing us to prove the relevant stability properties.

Theorem 2. *If symmetric matrices* $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(T)}$ *follow a weighted dynamic latent position model, then there exists* $\mathbf{\Lambda}^{(t)} \in \mathbb{R}^{d\times d_t}$ *of rank* d *and random matrices* $\mathbf{X} \in \mathbb{R}^{n\times d}$ *and*

[2]To avoid possible confusion, note that this section on minimal rank is only relevant in proving that this model can be characterised as a WMRDPG. Once that model is established, its asymptotic properties will be discussed using similar language but with some notation redefined.

$\mathbf{Y}^{(t)} \in \mathbb{R}^{n \times d_t}$ *(generated by some joint distribution* $\mathcal{G}$*) such that* $\mathbb{E}[\mathbf{A}_{ij}^{(t)}] = \mathbf{X}_i \mathbf{\Lambda}^{(t)} \mathbf{Y}_j^{(t)\top}$ *for each* $t \in [T]$*, where* $\mathbf{X}_i$ *and* $\mathbf{Y}_j^{(t)}$ *denote the* i*th and* j*th rows of* $\mathbf{X}$ *and* $\mathbf{Y}^{(t)}$ *respectively.*

Then, $(\mathbf{A}, \mathbf{X}, \mathbf{Y}) \sim \text{WMRDPG}(\mathcal{G}, \mathbf{\Lambda})$*, where* $\mathbf{A} = [\mathbf{A}^{(1)}|\cdots|\mathbf{A}^{(T)}] \in \mathbb{R}^{n \times nT}$*.*

Proof. See the appendix. □

For the remainder of this report, we'll use the convention that $(\mathbf{A}, \mathbf{X}, \mathbf{Y}) \sim \text{WMRDPG}(\mathcal{F}, \mathbf{\Lambda})$.

4.2.1 Important properties and assumptions

Define the Gram matrix $\mathbf{P} = [\mathbf{P}^{(1)}|\cdots|\mathbf{P}^{(T)}]$ with $\mathbf{P}^{(t)} = \mathbf{X}_i \mathbf{\Lambda}^{(t)} \mathbf{Y}_j^{(t)\top} = \mathbb{E}[\mathbf{A}_{ij}^{(t)}]$. Embedding this matrix in the same way as $\mathbf{A}$ results in 'noise free' embeddings which we'll denote as $\tilde{\mathbf{Y}}^{(t)}$. Also note that

Using SVD, we embed these matrices using UASE as before, with

$$\mathbf{A} \approx \mathbf{U_A \Sigma_A V_A^\top} \text{ and } \mathbf{P} = \mathbf{U_P \Sigma_P V_P^\top}$$

where$\mathbf{X_A} = \mathbf{U_A \Sigma_A^{1/2}} \in \mathbb{R}^{nT \times d}$ is the left UASE, and for $t \in [T]$ the tth right UASE is the matrix $\mathbf{Y_A}^{(t)} \in \mathbb{R}^{n_t \times d}$ obtained by partitioning $\mathbf{Y_A} = \mathbf{V_A \Sigma_A^{1/2}}$ into T blocks of sizes $n_1 \times d, \ldots, n_T \times d$.

4.3 Regularity conditions

The approach to asymptotics here is equivalent to that in A. Jones & P. Rubin-Delanchy [1]. Since we are analysing how n grows, we assume that the number of graphs T is either fixed, or grows very slowly such that $\lim_{T,n\to\infty} \frac{T}{n} = 0$.

Also assume that the number of latent positions $Y_i^{(t)}$ grows similarly to the number of latent positions X_i, so that for each $t \in [T]$, there exists a positive constant c_t such that $\lim_{n_t,n\to\infty} \frac{n_t}{n} = c_t$.

Now to provide some notation for asymptotic growth that will be used throughout. Let f and g be real-valued functions of $n, n_1, \ldots, n_T$. If there is a constant $c > 0$ and integers $N, N_1, \ldots, N_T$ such that for all $n \geq N$ and $n_t \geq N_t$ the subsequent conditions are satisfied, characterise each type of growth as follows:

- $f = \Omega(g)$ if $f(n, n_1, \ldots, n_T) \geq cg(n, n_1, \ldots, n_T)$;
- $f = \mathrm{O}(g)$ if $f(n, n_1, \ldots, n_T) \leq cg(n, n_1, \ldots, n_T)$;
- $f = \omega(g)$ if $f(n, n_1, \ldots, n_T) \geq cg(n, n_1, \ldots, n_T)$ and $\lim_{n,n_1,\ldots,n_T\to\infty} \left|\frac{f(n,n_1,\ldots,n_T)}{g(n,n_1,\ldots,n_T)}\right| = \infty$.

As in [1], "each of the collections $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ and $(\mathbf{Y}_1^{(t)}, \ldots, \mathbf{Y}_{n_t}^{(t)})$ is marginally i.i.d., with marginal distributions $\mathcal{F}_X$ on $\mathcal{X}$ and $\mathcal{F}_{Y,t}$ on $\mathcal{Y}_T$ for each t.

Given such a joint distribution $\mathcal{F}$, let $\xi \sim \mathcal{F}_X$ and $v_t \sim \mathcal{F}_{Y,t}$ for each t. Our next requirement is that these marginal distributions be non-degenerate, in the sense that:

- The second moment matrices $\Delta_X = \mathbb{E}\left[\xi\xi^\top\right] \in \mathbb{R}^{d\times d}$ and $\Delta_{Y,t} = \mathbb{E}\left[v_t v_t^\top\right] \in \mathbb{R}^{d_t\times d_t}$ for each t are all invertible."

In order to prove the following proposition, we use a slightly simplified version of the matrix Hoeffding lemma (as in [26], Theorem 1.3).

Lemma 3 (Matrix Hoeffding lemma)**.** *Consider a finite sequence $\{\mathbf{X}_k\}$ of independent, random, self- adjoint matrices with dimension d, and let $\{\mathbf{A}_k\}$ be a sequence of fixed self-adjoint matrices. Assume that each random matrix satisfies Assume that $\mathbb{E}[\mathbf{X}]$ is finite and $\mathbf{X}^2 \preceq C\mathbf{I}$ for some positive constant C almost surely. Then, for all $\tau \geq 0$,*

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^{(i)} - \mathbb{E}[\mathbf{X}]\right\| \geq \tau\right\} \leq d\exp\left\{-\frac{\tau^2}{8C^2}\right\}.$$

Here, $\|\mathbf{A}\|$ represents the spectral norm, or the largest singular value of a matrix. Let $\lambda_{\max}$ represent the largest eigenvalue and $\sigma_{\max}$ the largest singular value:

$$\|\mathbf{A}\| = \sqrt{\lambda_{\max}(\mathbf{A}^*\mathbf{A})} = \sigma_{\max}(\mathbf{A}).$$

Proposition 4. *The spectral norm $||\mathbf{X}^\top\mathbf{X} - n\Delta_X||$ is of order $\mathrm{O}(n^{1/2}\log(n))$ and for each t, $||\mathbf{Y}^{(t)\top}\mathbf{Y}^{(t)} - n_t\Delta_{Y,t}|| = \mathrm{O}(n_t^{1/2}\log(n_t))$ almost surely.*

Proof. $\mathbf{X}^\top\mathbf{X} = \sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^\top$. Then, using the matrix Hoeffding lemma and that $\mathbf{X}^2 \preceq C\mathbf{I}$ (due to $\mathbf{A}$ having a bounded expectation) for some constant C,

$$\begin{aligned}\mathbb{P}\{\|\mathbf{X}^\top\mathbf{X} - n\Delta_X\| \geq \tau\} &= \mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^\top - \mathbb{E}\left[\xi\xi^\top\right]\right\| \geq \frac{\tau}{n}\right\} \\ &\leq d\exp\left\{-\frac{(\frac{\tau}{n})^2}{8C^2}\right\}.\end{aligned}$$

The numerator dominates when $\tau = n^{1/2}\log(n)$. Proof for $\mathbf{Y}$ terms is similar. □

We also define what's called a global sparsity function ρ with $\rho : \mathbb{Z}_+ \to [0,1]$ which is either constant or tends to 0 as n grows. It is used to uniformly scale the latent positions $\mathbf{X}$ and $\mathbf{Y}^{(t)}$. When ρ tends to 0, this corresponds with the *sparse* regime where the majority of covariates are independent or not statistically close to one another in the context of detecting concept drift.

This scaling is applied to the model by first letting $(\xi_1, \ldots, \xi_n, v_1^{(1)}, \ldots, v_{n_T}^{(T)}) \sim \mathcal{F}$, and then defining the latent positions as $\mathbf{X}_i = \rho^{1/2}\xi_i$ and $\mathbf{Y}_i^{(t)} = \rho^{1/2}v_i^{(t)}$. Call this scaled distribution $\mathcal{F}_\rho$.

In order to avoid the graphs being too sparse to provide any useful information, we impose a limit on the rate of decay of ρ:

$$\rho = \omega\left(\frac{\log^c(n)}{n^{1/2}}\right)$$

for some constant c.

4.4 Central Limit Theorem

The machinery required to prove distributional results such as the central limit theorem has finally been set up. The proofs for each of the following results are in the appendix and their statements are equivalent to those in I. Gallagher et al. (Propositions 2 and 3, [2])

The first result is that, up to rotation, the embeddings $\hat{\mathbf{Y}}_i$ converge in Euclidean norm to noise free embeddings $\tilde{\mathbf{Y}}_i$.

Theorem 5 (Convergence to noise-free embeddings)**.** *There exists a a sequence of orthogonal matrices $\tilde{\mathbf{W}} \in \mathbb{O}(d)$ such that*

$$\max_{i\in[n]} \|\hat{\mathbf{Y}}_i^{(t)}\tilde{\mathbf{W}} - \tilde{\mathbf{Y}}_i^{(t)}\| = \mathrm{O}\left(\frac{\log^{1/2}(n)}{\rho^{1/2}n^{1/2}}\right),$$

with high probability for each t.

The need for rotation by $\tilde{\mathbf{W}}$ is a result of the low rank expectation assumption of the weighted dynamic latent position model (Definition 1). Any orthogonal matrix $\mathbf{Q}$ that has the property $\mathbf{Q}\mathbf{I}_{p_t,q_t}\mathbf{Q}^\top = \mathbf{I}_{p_t,q_t}$, allows us to define a new map $\phi'(\mathbf{z}) = \mathbf{Q}^\top\phi(\mathbf{z})$ which is also be a perfectly valid choice of map to satisfy the low rank expectation:

$$\mathbb{E}(\mathbf{A}^{(t)}) = \phi'(\mathbf{z}_1)^\top\mathbf{I}_{p_t,q_t}\phi'(\mathbf{z}_2) = \phi(\mathbf{z}_1)^\top\mathbf{Q}\mathbf{I}_{p_t,q_t}\mathbf{Q}^\top\phi(\mathbf{z}_2) = \phi(\mathbf{z}_1)^\top\mathbf{I}_{p_t,q_t}\phi(\mathbf{z}_2).$$

Matrices that satisfy this property form the group of indefinite orthogonal transformations ([3], Section 4.2). $\tilde{\mathbf{W}}$ are the matrices in this group that align $\hat{\mathbf{Y}}_i^{(t)}$ with the noise-free embeddings $\tilde{\mathbf{Y}}_i^{(t)}$.

While there is no way to compute these matrices, they are defined to be the solution to the one-mode orthogonal Procrustes problem [27] whose details will be in the appendix. It is an optimisation problem which aims to find the orthogonal matrix which reduces the distance between the UASEs associated with $\mathbf{A}$ and $\mathbf{P}$ in the Frobenius norm (see Definition 10).

$\tilde{\mathbf{W}}$ being the same for each embedding at each time point t implies some form of stability in the noise-free embeddings $\tilde{\mathbf{Y}}^{(t)}$ which is captured by UASE.

Equipped with the rate of convergence to the noise free embeddings, we can show that

the points $\hat{\mathbf{Y}}_i^{(t)}$ converge to a multivariate Gaussian distribution as $n \to \infty$ after applying a second transformation.

The proof for this culminates in the use of the multivariate central limit theorem (a generalisation of the standard central limit theorem) to show that the points converge to the distribution we require.

Theorem 6. *Let $\xi \sim \mathcal{F}$, and for $\mathbf{z} \in \mathcal{Z}$ define*

$$\mathbf{\Sigma}_X^{(t)}(\mathbf{z}) = \begin{cases} \mathbb{E}\left[\xi^\top \mathbf{\Lambda}_t(\mathbf{z})(1 - \xi^\top \mathbf{\Lambda}_t(\mathbf{z})) \cdot \xi\xi^\top\right] & \text{if } \rho = 1 \\ \mathbb{E}\left[\xi^\top \mathbf{\Lambda}_t(\mathbf{z}) \cdot \xi\xi^\top\right] & \text{if } \rho \to 0 \end{cases}$$

Then for all $\mathbf{y} \in \mathbb{R}^{d_t}$ and for any fixed $i \in [n]$ and $t \in [T]$,

$$\mathbb{P}\left(n^{1/2}(\hat{\mathbf{Y}}_i^{(t)}\tilde{\mathbf{W}} - \tilde{\mathbf{Y}}_i^{(t)}) \leq \mathbf{y} \mid \mathbf{Z}_i^{(t)} = \mathbf{z}\right) \to \mathbf{\Phi}\left(\mathbf{y}, \mathbf{\Lambda}_t^{-1}\Delta_X^{-1}\mathbf{\Sigma}_X^{(t)}(\mathbf{z})\mathbf{\Lambda}_t^{-1}\Delta_X^{-\top}\right)$$

The matrices $\mathbf{\Lambda}_t$ and Δ_X are analogous to $\tilde{\mathbf{W}}$ from the previous result in that they can only be explicitly constructed given knowledge of the underlying function f and distribution $\mathcal{F}$ (which is information that we do not have).

The main property that makes UASE so attractive, not only in the context of concept drift but also in general, is that it satisfies the two forms of stability mentioned earlier (cross-sectional and longitudinal stability). Here we use the definition provided in the paper by I. Gallagher et al ([2], Definition 4) so that we can prove that UASE still has these properties in this new context.

"We say that two space-time positions $(\mathbf{z}, t)$ and $(\mathbf{z}', t')$ are exchangeable if $f(\mathbf{z}, \zeta_t) = f(\mathbf{z}', \zeta_t)$ with probability one, where $\zeta = (\zeta_1 | \cdots | \zeta_T) \sim \mathcal{F}$, and that the positions are exchangeable up to degree if $f(\mathbf{z}, \zeta_t) = f(\mathbf{z}', \zeta_t)$ for some > 0. Equivalently, $(\mathbf{z}, t)$ and $(\mathbf{z}', t')$ are exchangeable if conditional on $\mathbf{Z}_i^{(t)} = \mathbf{z}$ and $\mathbf{Z}_j^{(t)} = \mathbf{z}'$ the ith row of $\mathbf{P}^{(t)}$ and jth row of $\mathbf{P}^{(t)}$ are equal with probability one."

In our context, $f(\mathbf{x}, \mathbf{y}) = \mathbf{x}\mathbf{\Lambda}_t\mathbf{y}$). They give a definition that can be applied to dynamic networks more broadly.

Definition 7. *Given a generic method for dynamic network embedding, with output denoted $(\hat{\mathbf{Z}}_i^{(t)})_{i\in[n];t\in[T]}$, define the following stability properties:*

(1) Cross-sectional stability: Given exchangeable $(\mathbf{z}, t)$ and $(\mathbf{z}', t)$, $\hat{\mathbf{Z}}_i^{(t)}$ and $\hat{\mathbf{Z}}_j^{(t)}$ are asymptotically equal, with identical error distribution, conditional on $\mathbf{Z}_i^{(t)} = \mathbf{z}$ and $\mathbf{Z}_j^{(t)} = \mathbf{z}'$.

(2) Longitudinal stability: Given exchangeable $(\mathbf{z}, t)$ and $(\mathbf{z}, t')$, $\hat{\mathbf{Z}}_i^{(t)}$ and $\hat{\mathbf{Z}}_i^{(t)}$ are asymptotically equal, with identical error distribution, conditional on $\mathbf{Z}_i^{(t)} = \mathbf{Z}_i^{(t)} = \mathbf{z}$.

Now to state the result that UASE exhibits both types of stability:

Corollary 8. *Conditional on* $\mathbf{Z}_i^{(t)} = \mathbf{z}$ *and* $\mathbf{Z}_j^{(t)} = \mathbf{z}'$*, the following properties hold:*

(1) If $(\mathbf{z}, t)$ *and* $(\mathbf{z}', t')$ *are exchangeable then* $\hat{\mathbf{Y}}_i^{(t)}$ *and* $\hat{\mathbf{Y}}_j^{(t)}$ *are asymptotically equal, with identical error distribution.*

(2) If $(\mathbf{z}, t)$ *and* $(\mathbf{z}', t')$ *are exchangeable up to degree then* $\hat{\mathbf{Y}}_i^{(t)}$ *and* $\hat{\mathbf{Y}}_j^{(t)}$ *are asymptotically equal and, under a sparse regime* ($\rho \to 0$)*, their error distributions are equal up to scale, satisfying* $\Sigma_t(\mathbf{z}) = \Sigma_t(\mathbf{z}')$.

Under our toy model, we can see that a pair $(\mathbf{z}, t)$ and $(\mathbf{z}', t')$ is exchangeable if and only if the corresponding rows of $\mathbf{B}^t$ and $\mathbf{B}^{t'}$ are the same.

We can use the theory we've developed here to make sense of the properties we observed earlier. For example, Theorem 6 predicts that the points will converge to a finite number of Gaussian clusters and Theorem 8 says that those clusters will be identical if each pair is exchangeable (which is the case for the fourth rows $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$).

5. Conclusion

In this report, we've introduced the use of UASE to detect concept drift alongside a rich statistical theory to make sense of it and provide more tools for analysis. Using these tools, we've proven a central limit theorem to show that the method has desirable stability properties. It's unique in that it's able to visualise concept drift in a high number of dimensions while showing how the individual features move in relation to one another. We showed these properties through the use of a toy model.

There's still a lot of work to be done for in improving the effectiveness of this model. The most notable being the choice of statistical discrepancy between each feature. This report uses covariance as the canonical discrepancy, but there is actually a lot of freedom to choose here. It is an open question as to what distance measure to use in which context, and work could be carried out in a similar way to how I. Goldenberg et al. compares different distance measures for detecting concept drift [17]. In addition, I. Gallagher et al. shows that there are a range of entry-wise transformations that can be made to the unfolding $\mathbf{A}$ that can change the quality of an embedding such as taking the logarithm of each entry or setting all values below a certain threshold to 0 [3], which would be an interesting line of inquiry.

Most importantly, there are many application areas that this model can and should be tested on, which would be the best showcase of its validity.

References

[1] A. Jones & P. Rubin-Delanchy, *The multilayer random dot product graph*, 2021. arXiv: 2007.10455 [stat.ML].

[2] I. Gallagher, A. Jones & P. Rubin-Delanchy, *Spectral embedding for dynamic networks with stability guarantees*, 2022. arXiv: 2106.01282 [stat.ML].

[3] I. Gallagher, A. Jones, A. Bertiger, C. Priebe & P. Rubin-Delanchy, *Spectral embedding of weighted graphs*, 2023. arXiv: 1910.05534 [stat.ML].

[4] R. B. Velasco, I. Carpanese, R. Interian, O. C. Paulo Neto & C. C. Ribeiro, 'A decision support system for fraud detection in public procurement,' *International Transactions in Operational Research*, 28, no. 1, pp. 27–47, 2021.

[5] J. Plummer, S. D. Rappaport, T. Hall & R. Barocci, *The online advertising playbook: Proven strategies and tested tactics from the advertising research foundation*. John Wiley & Sons, 2007.

[6] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama & G. Zhang, 'Learning under concept drift: A review,' *IEEE transactions on knowledge and data engineering*, 31, no. 12, pp. 2346–2363, 2018.

[7] A. Suárez-Cetrulo, D. Quintana & A. Cervantes, 'A survey on machine learning for recurring concept drifting data streams,' *Expert Systems with Applications*, 213, p. 118 934, Oct. 2022. DOI: 10.1016/j.eswa.2022.118934.

[8] Y. Hu, K. Liu, X. Zhang *et al.*, 'Concept drift mining of portfolio selection factors in stock market,' *Electronic Commerce Research and Applications*, 14, no. 6, pp. 444–455, 2015.

[9] G. I. Webb, L. K. Lee, B. Goethals & F. Petitjean, 'Analyzing concept drift and shift from sample data,' *Data Mining and Knowledge Discovery*, 32, pp. 1179–1199, 2018.

[10] K. Zhang, A. T. Bui & D. W. Apley, 'Concept drift monitoring and diagnostics of supervised learning models via score vectors,' *Technometrics*, 65, no. 2, pp. 137–149, 2023.

[11] T. R. Hoens, R. Polikar & N. V. Chawla, 'Learning from streaming data with concept drift and imbalance: An overview,' *Progress in Artificial Intelligence*, 1, pp. 89–101, 2012.

[12] G. V. Trunk, 'A problem of dimensionality: A simple example,' *IEEE Transactions on pattern analysis and machine intelligence*, no. 3, pp. 306–307, 1979.

[13] B. Chandrasekaran & A. K. Jain, 'Quantization complexity and independent measurements,' *IEEE Transactions on Computers*, 100, no. 1, pp. 102–106, 1974.

[14] G. J. McLachlan, *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons, 2005.

[15] L. Van Der Maaten, E. Postma, J. Van den Herik *et al.*, 'Dimensionality reduction: A comparative,' *J Mach Learn Res*, 10, no. 66-71, 2009.

[16] G. I. Webb, R. Hyde, H. Cao, H. L. Nguyen & F. Petitjean, 'Characterizing concept drift,' *Data Mining and Knowledge Discovery*, 30, no. 4, pp. 964–994, 2016.

[17] I. Goldenberg & G. I. Webb, 'Survey of distance measures for quantifying concept drift and shift in numeric data,' *Knowledge and Information Systems*, 60, no. 2, pp. 591–615, 2019.

[18] S. Banerjee & A. Roy, *Linear algebra and matrix analysis for statistics*. Crc Press Boca Raton, 2014, vol. 181.

[19] I. Gohberg & M. G. Krein, *Introduction to the theory of linear nonselfadjoint operators*. American Mathematical Soc., 1978, vol. 18.

[20] C. Matias & V. Miele, 'Statistical clustering of temporal networks through a dynamic stochastic block model,' *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79, no. 4, pp. 1119–1141, 2017.

[21] M. Pensky & T. Zhang, 'Spectral clustering in the dynamic stochastic block model,' 2019.

[22] B. Kim, K. H. Lee, L. Xue & X. Niu, 'A review of dynamic network models with latent variables,' *Statistics Surveys*, 12, no. none, pp. 105–135, 2018. DOI: 10.1214/18-SS121. [Online]. Available: https://doi.org/10.1214/18-SS121.

[23] Y. Liu, E. Jun, Q. Li & J. Heer, 'Latent space cartography: Visual analysis of vector space embeddings,' in *Computer graphics forum*, Wiley Online Library, vol. 38, 2019, pp. 67–78.

[24] A. N. Gorban & I. Y. Tyukin, 'Blessing of dimensionality: Mathematical foundations of the statistical physics of data,' *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376, no. 2118, p. 20 170 237, 2018.

[25] N. Whiteley, A. Gray & P. Rubin-Delanchy, *Statistical exploration of the manifold hypothesis*, 2024. arXiv: 2208.11665 [stat.ME].

[26] J. A. Tropp, 'User-friendly tail bounds for sums of random matrices,' *Foundations of Computational Mathematics*, 12, no. 4, pp. 389–434, Aug. 2011, ISSN: 1615-3383. DOI: 10.1007/s10208-011-9099-z. [Online]. Available: http://dx.doi.org/10.1007/s10208-011-9099-z.

[27] J. Gower, 'Multivariate analysis: Ordination, multidimensional scaling and allied topics,' 1984.

[28] J. Cape, M. Tang & C. E. Priebe, *The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics*, 2018. arXiv: 1705.10735 [math.ST].

[29] R. A. Horn & C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[30] A.-L. Cauchy, 'Sur les formules qui résultent de l'emploie du signe et sur> ou<, et sur les moyennes entre plusieurs quantités,' *Cours d'Analyse, 1er Partie, Analyse Algebrique*, pp. 373–377, 1821.

[31] T. Popoviciu, 'Sur les équations algébriques ayant toutes leurs racines réelles,' *Mathematica*, 9, no. 129-145, p. 20, 1935.

[32] W. Hoeffding, 'Probability inequalities for sums of bounded random variables,' *Journal of the American Statistical Association*, 58, no. 301, pp. 13–30, 1963. DOI: 10.1080/01621459.1963.10500830. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1963.10500830. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500830.

[33] M. Huber, 'Halving the bounds for the markov, chebyshev, and chernoff inequalities using smoothing,' *The American Mathematical Monthly*, 126, no. 10, pp. 915–927, 2019.

[34] J. A. Tropp, *An introduction to matrix concentration inequalities*, 2015. arXiv: 1501.01571 [math.PR].

[35] V. Lyzinski, M. Tang, A. Athreya, Y. Park & C. E. Priebe, 'Community detection and classification in hierarchical stochastic blockmodels,' *IEEE Transactions on Network Science and Engineering*, 4, no. 1, pp. 13–26, 2016.

6. Appendix

6.1 Important inequalities and notation

In order to understand the proofs of the main theorems in this paper, we will establish inequalities and notation that are used throughout.

Definition 9 (Two-to-infinity norm)**.** *For* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *the two-to-infinity norm [28]* $\|\cdot\|_{2\to\infty}$ *denotes the maximum row-wise Euclidean norm of a matrix:*

$$\|\mathbf{A}\|_{2\to\infty} = \max_{i\in[n]} \|\mathbf{A}_i\|.$$

Definition 10 (Frobenius norm)**.** *For* $\mathbf{A} \in \mathbb{R}^{m\times n}$*, the Frobenius norm [29]* $\|\cdot\|_F$ *is defined:*

$$\|\mathbf{A}\|_F = \sqrt{\sum_i^m \sum_j^n |a_{ij}|^2} = \sqrt{\text{trace}(\mathbf{A}^*\mathbf{A})} = \sqrt{\sum_{i=1}^{min\{m,n\}} \sigma_i^2(\mathbf{A})},$$

where $\sigma_i(\mathbf{A})$ *denotes the ith singular value of* $\mathbf{A}$*.*

An important property of the Frobenius norm is that, for all $\mathbf{A}$,$\|\mathbf{A}\| \leq \|\mathbf{A}\|_F$.

Theorem 11 (Cauchy-Schwarz inequality (Euclidean)). *For vectors $u, v \in \mathbb{R}^n$, the Cauchy-Schwarz inequality [30] states that*

$$\left(\sum_{i=1}^{n} u_i v_i\right)^2 = \left(\sum_{i=1}^{n} u_i^2\right)\left(\sum_{i=1}^{n} v_i^2\right).$$

Theorem 12 (Popoviciu's inequality). *For any random variable with variance σ^2, and bounded above and below by B and A, Popoviciu's inequality [31] states that*

$$\sigma^2 \leq \frac{1}{4}(B-A)^2.$$

Theorem 13 (Hoeffding's inequality). *Let $X_1, \ldots, X_n$ be independent random variables with $a_i \leq X_i \leq b_i$ almost surely. For the sum of these random variables $S_n = X_1 + \cdots + X_n$, Hoeffding's inequality [32] states that, for all $t < 0$,*

$$\mathbb{P}(|S_n - \mathbb{E}S_n| \geq t) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Theorem 14 (Markov's inequality). *If X is a non-negative random variable and $a > 0$, Markov's inequality [33] states that the probability that X is at least a is at most the expectation of X divided by a:*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}X}{a}.$$

6.2 Proof of Theorem 1

As stated before, the main difference between the WMRDPG and the weighted dynamic latent position model is the existence of random matrices $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{Y}^{(t)} \in \mathbb{R}^{n \times d_t}$ generated by some joint distribution $\mathcal{G}$, and matrices $\mathbf{\Lambda}^{(t)} \in \mathbb{R}^{d \times d_t}$.

The weighted dynamic latent position model is characterised by the existence of maps $\phi_t : \mathcal{Z} \to \mathbb{R}^{D_t}$ for each t with a further assumption that each map has a full rank second moment matrix

$$\Delta_t = \mathbb{E}[\xi_t \xi_t^\top] \in \mathbb{R}^{D_t \times D_t},$$

where each $\mathcal{F}^*$ is the joint distribution such that $\xi \sim \mathcal{F}^*$ with marginal distributions $\mathcal{F}_1^*, \ldots, \mathcal{F}_T^*$ so $\xi_t \sim \mathcal{F}_t^*$.

To prove the equivalence, we construct suitable maps between the vectors generated by $\mathcal{F}^*$ and the random matrices generated by $\mathcal{G}$ so that the second moment matrices of the marginal distributions generated by $\mathcal{G}$ are also full rank.

This is done by first constructing matrices whose rows form the basis of the support of $\mathcal{F}^*$ and then performing a series of singular value decompositions to obtain full rank

matrices. Since the number of non-zero singular values is equal to the rank of the matrix being decomposed, we can construct our maps in peace.

Proof. Define second moment matrix $\Delta = \mathbb{E}[\xi\xi^\top] \in \mathbb{R}^{(D_1+\cdots+D_T)\times(D_1+\cdots+D_T)}$ with $rank(\Delta) = r$.

Let $\mathbf{M} \in \mathbb{R}^{r\times(D_1+\cdots+D_T)}$ be a matrix whose rows form a basis of $supp(\mathcal{F}^*)$, and similarly let $\mathbf{N}_t \in \mathbb{R}^{D_t\times((D_1+\cdots+D_T)}$ be a matrix whose rows form a basis of $supp(\mathcal{F}_t^*)$. Let $\mathbf{N} = \text{diag}(\mathbf{N}_1, \ldots, \mathbf{N}_T)$, and define the matrix $\mathbf{\Pi} := \mathbf{MDN} \in \mathbb{R}^{r\times(D_1+\cdots+D_T)}$ where $\mathbf{D} = \text{diag}(\mathbf{I}_{p_1,q_1}, \ldots, \mathbf{I}_{p_T,q_T})$

Construct the singular value decomposition $\mathbf{\Pi} = \mathbf{U\Sigma V}^\top$ with $\mathbf{U} \in O(r \times d), \mathbf{\Sigma} \in \mathbb{R}^{d\times d}, \mathbf{V} \in O((D_1 + \cdots + D_T) \times d)$ and $d = rank(\mathbf{\Pi})$. Then, writing $\mathbf{V} = [\mathbf{V}_1|\cdots|\mathbf{V}_T]$ where $\mathbf{V}_t \in \mathbb{R}^{D_t\times d}$ has rank d_t, we construct singular value decomposition $\mathbf{V}_t = \mathbf{U}_t\mathbf{\Sigma}_t\mathbf{W}_t^\top$ with $\mathbf{U} \in O(D_t \times d_t), \mathbf{\Sigma} \in \mathbb{R}^{d_t\times d_t}, \mathbf{W} \in O(d \times d_t)$ where $d_t = rank(\mathbf{V_t})$.

Define $\mathbf{\Lambda} = \mathbf{\Sigma W} \in \mathbb{R}^{d\times(d_1+\cdots+d_T)}$, where $\mathbf{W} = (\mathbf{W}_1\mathbf{\Sigma}_1|\cdots|\mathbf{W}_T\mathbf{\Sigma}_T)$, and let $L : \mathbb{R}^{(D_1+\cdots+D_T)} \to \mathbb{R}^d$ be the linear map sending the ith row of $\mathbf{M}$ to the ith row of $\mathbf{U}$, and similarly let $L_t : \mathbb{R}^{D_t} \to \mathbb{R}^{d_t}$ be the linear map sending the ith row of $\mathbf{N}_t$ to the ith row of $\mathbf{U}_t$. Finally, let $\varphi : \mathcal{Z}^T \to \mathbb{R}^d$ and $\varphi_t : \mathcal{Z} \to \mathbb{R}^{d_t}$ be the maps satisfying $\varphi(\mathbf{z}) = L\left(\left(\phi\left(\mathbf{z}^{(1)}\right)|\cdots|\phi\left(\mathbf{z}^{(T)}\right)\right)\right.$ and $\varphi_t\left(\mathbf{z}^{(t)}\right) = L_t\left(\phi\left(\mathbf{z}^{(t)}\right)\right)$ for any $\mathbf{z} = \left(\mathbf{z}^{(1)}|\cdots|\mathbf{z}^{(T)}\right) \in \mathcal{Z}^T$.

Then, setting $\mathcal{G}$ to be the joint distribution on $\mathbb{R}^d \times \mathbb{R}^{d_1} \times \cdots \mathbb{R}^{d_T}$ obtained by first assigning a random vector $\zeta = (\zeta_1|\cdots|\zeta_T)$ via $\mathcal{F}$ and then sending this to the tuple $(\varphi(\zeta), \varphi_1(\zeta_1), \ldots, \varphi_t(\zeta_t))$ and letting $\mathbf{X}_i = \varphi(\mathbf{Z}_i)$ and $\mathbf{Y}_i^{(t)} = \varphi_t\left(\mathbf{Z}_i^{(t)}\right)$, we find that $(\mathbf{A}, \mathbf{X}, \mathbf{Y}) \sim \text{WMRDPG}(\mathcal{G}, \mathbf{\Lambda})$, where $\mathbf{A} = [\mathbf{A}^{(1)}|\cdots|\mathbf{A}^{(T)}] \in \mathbb{R}^{n\times nT}$. □

6.3 Asymptotic properties

There is a large number of regularity conditions that must be proven in order to begin the proof of the central limit theorem. Again, the vast majority of these proofs follow the same format and structure as the paper by A. Jones & P. Rubin-Delanchy [1], with a few key differences since these graphs are weighted. When the arguments are identical, the corresponding proof will be referenced.

Through the proofs, we track how the rate of growth of various norms change. It is very important to understand how to manipulate these in terms of big O notation as outlined in Chapter 4.3 as this is used heavily.

We first note the rate of growth of the spectral norms of the unfolding and the Gram matrix and their associated UASEs.

Proposition 15. $\sigma_i(\mathbf{A}) = \sigma_i(\mathbf{P}) = \text{O}(\rho n) = \Omega(\rho n)$.

The proof follows identical arguments to those in [1], Proposition 7 and Proposition 10. It is centred around using the rate of growth of the second moment matrix as in Proposition

4.

The following proposition makes use of a matrix analogue of Bernstein's theorem ([34],Theorem 1.6.2).

Theorem 16 (Matrix Bernstein)**.** *Let $\mathbf{M}_1, \ldots, \mathbf{M}_n$ be independent random matrices with common dimensions $m_1 \times m_2$, satisfying $\mathbb{E}[\mathbf{M}_k] = 0$ and $\|\mathbf{M}_k\| \leq L$ for each $1 \leq k \leq n$, for some fixed value L.*

Let $\mathbf{M} = \sum \mathbf{M}_k$ and let $v(\mathbf{M}) = \max\left\{\left\|\mathbb{E}\left[\mathbf{M}\mathbf{M}^\top\right]\right\|, \left\|\mathbb{E}\left[\mathbf{M}^\top\mathbf{M}\right]\right\|\right\}$ denote the matrix variance statistic of $\mathbf{M}$. Then for all $t \geq 0$, we have

$$\mathbb{P}(\|\mathbf{M}\| \geq \tau) \leq (m_1 + m_2) \exp\left(\frac{-\tau^2/2}{v(\mathbf{M}) + L\tau/3}\right).$$

Proposition 17. $||\mathbf{A} - \mathbf{P}|| = O(\rho T^{1/4} n^{1/2} \log^{\alpha+1/2}(n))$.

Proof. Define matrix $\mathbf{T}_{ij}^{(t)} \in \mathbb{R}^{n \times (n_1 + \cdots + n_T)}$ for each $i \in [n]$ and $j \in [n_t]$ whose $(i, n_1 + \cdots + n_{t-1} + j)$th entry is equal to $\mathbf{A}_{ij}^{(t)} - \mathbf{P}_{ij}^{(t)}$ with all other entries 0. Define $\mathbf{M}_{ij}^{(t)}$ for each $t \in [T]$ as $\mathbf{M}_{ij}^{(t)} = \mathbf{T}_{ij}^{(t)} + \mathbf{T}_{ji}^{(t)}$ for each i, j with $i < j$.

Note that $||\mathbf{M}_{ij}^{(t)}|| = |\mathbf{A}_{ij} - \mathbf{P}_{ij}| < 2\beta\rho \log^\alpha(n)$ almost surely and $\mathbb{E}\mathbf{M}_{ij}^{(t)} = 0$, so the sum $\mathbf{M} = \sum_{i,j,t} \mathbf{M}_{ij}^{(r)}$ satisfies criteria for Bernstein's theorem.

To bound the variance statistic $v(\mathbf{M})$, let $\mathbf{M}_r = \sum_{i,j} \mathbf{M}_{ij}^{(t)}$. Note that

$$\mathbf{M}_t \mathbf{M}_t^\top = \sum_{l \neq i,j} (\mathbf{A}_{il}^{(t)} - \mathbf{P}_{il}^{(t)})(\mathbf{A}_{jl}^{(t)} - \mathbf{P}_{jl}^{(t)}).$$

Hence,

$$\mathbb{E}[\mathbf{M}_t \mathbf{M}_t^\top]_{ij} = \begin{cases} \mathrm{Var}(\mathbf{A}_{ij}^{(t)}), & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases}$$

By Popoviciu's inequality, $\mathrm{Var}(\mathbf{A}_{ij}^{(t)})$ is bounded by $\beta^2\rho^2 \log^{2\alpha}(n_t)$. Since $\mathbb{E}[\mathbf{M}_t \mathbf{M}_t^\top]$ is diagonal, $||\mathbb{E}[\mathbf{M}_t \mathbf{M}_t^\top]|| \leq \beta^2\rho^2 \log^{2\alpha}(n_t)$. Since $\mathbf{M}\mathbf{M}^\top = \sum_t \mathbf{M}_t \mathbf{M}_t^\top$ and $\mathbf{M}_t$ are independent, $||\mathbb{E}[\mathbf{M}\mathbf{M}^\top]|| = \sum_{t=1}^T \beta^2\rho^2 n_t \log^{2\alpha}(n_t) \leq \beta^2\rho^2 T^{1/2} n_{\max} \log^{2\alpha}(n_{\max})$ by the Cauchy-Schwarz inequality.

By analogous arguments for $||\mathbb{E}[\mathbf{M}^\top\mathbf{M}]||$, we see that $v(\mathbf{M}) = O(\beta^2\rho^2 T^{1/2} n \log^{2\alpha}(n)$. Then by Bernstein's theorem and rearranging,

$$\mathbb{P}(\|\mathbf{M}\| \geq \tau) \leq (n + n_1 + \cdots + n_T) \exp\left(\frac{-3\tau^2}{6\beta^2\rho^2 T^{1/2} n \log^{2\alpha}(n) + 4\beta\rho \log^\alpha(n)\tau}\right).$$

The numerator dominates for sufficiently large n when $\tau = O(\rho T^{1/4} n^{1/2} \log^{\alpha+1/2}(n))$, so $||\mathbf{A} - \mathbf{P}|| = O(\rho T^{1/4} n^{1/2} \log^{\alpha+1/2}(n))$ almost surely.

□

Proposition 18. $||\mathbf{U_P}^\top(\mathbf{A}-\mathbf{P})\mathbf{V_P}||_F = O(\rho\log^{\alpha+1/2}(n))$.

Proof. Condition on some choice of latent positions. For each $i,j\in[d]$ and $t\in[T]$, let

$$\mathbf{E}_{ij}^{(t)} = \sum_{p\le q}(u_p v_q^{(t)} + u_q v_p^{(t)})(\mathbf{A}_{pq}^{(t)} - \mathbf{P}_{pq}^{(t)}) \text{ and } \mathbf{F}_{ij}^{(t)} = \sum_{p=1}^{n} u_p v_p^{(t)} \mathbf{P}_{pp}^{(t)},$$

where u and v_t denote the ith and jth columns respectively, so that

$$(\mathbf{U_P}^\top(\mathbf{A}-\mathbf{P})\mathbf{V_P})_{ij} = \sum_{t=1}^{T}\mathbf{E}_{ij}^{(t)} - \sum_{t=1}^{T}\mathbf{F}_{ij}^{(t)}.$$

We ignore the last term in our asymptotic analysis by exploiting the orthonormality of the singular vectors and the bounded expectation of $\mathbf{P}$. The argument follows along the lines of [1] (Proposition 11) and [35] (Lemma 17).

Each of the $\mathbf{E}_{ij}^{(t)}$ is a sum of independent zero-mean random variables, with each of the individual terms bounded in absolute value by $|u_p v_q^{(t)} + u_q v_p^{(t)}|$. Applying Hoeffding's inequality, we thus observe that

$$\begin{aligned}\mathbb{P}(|\sum_{t=1}^{T}\mathbf{E}_{ij}^{(t)}| > \tau) &\le 2\exp\left(\frac{-2\tau^2}{16(\beta^2\rho^2\log^{2\alpha}(n))\sum_{t=1}^{T}\sum_{p-q}((u_p v_q^{(t)} + u_q v_p^{(t)})^2}\right)\\ \text{(expand + Cauchy-Schwarz)} &\le 2\exp\left(\frac{-2\tau^2}{64\beta^2\rho^2\log^{2\alpha}(n)}\right).\end{aligned}$$

Thus $\sum_{t=1}^{T}\mathbf{E}_{ij}^{(t)} = O(\rho\log^{\alpha+1/2}(n))$ almost surely, and the result follows after integrating over all possible choices of latent positions.

□

From this point onwards, there are many results that have analogous proofs for terms involving $\mathbf{U}$ terms and those involving $\mathbf{V}$. Many of the proofs involving $\mathbf{V}$ terms will be omitted, unless the difference is significant.

Proposition 19. *The following bounds hold almost surely:*

(i) $\left\|\mathbf{U_A}\mathbf{U_A}^\top - \mathbf{U_P}\mathbf{U_P}^\top\right\|, \left\|\mathbf{V_A}\mathbf{V_A}^\top - \mathbf{V_P}\mathbf{V_P}^\top\right\| = \mathrm{O}(T^{1/4}n^{-1/2}\log^{\alpha+1/2});$

(ii) $\left\|\mathbf{U_A} - \mathbf{U_P}\mathbf{U_P}^\top\mathbf{U_A}\right\|_F, \left\|\mathbf{V_A} - \mathbf{V_P}\mathbf{V_P}^\top\mathbf{V_A}\right\|_F = \mathrm{O}(T^{1/4}n^{-1/2}\log^{\alpha+1/2});$

(iii) $\left\|\mathbf{U_P}^\top\mathbf{U_A}\mathbf{\Sigma_A} - \mathbf{\Sigma_P}\mathbf{V_P}^\top\mathbf{V_A}\right\|_F, \left\|\mathbf{\Sigma_P}\mathbf{U_P}^\top\mathbf{U_A} - \mathbf{V_P}^\top\mathbf{V_A}\mathbf{\Sigma_A}\right\|_F = \mathrm{O}(\rho T^{1/2}\log^{2\alpha+1});$

(iv) $\left\|\mathbf{U_P}^\top\mathbf{U_A} - \mathbf{V_P}^\top\mathbf{V_A}\right\|_F = \mathrm{O}(T^{1/2}n^{-1/2}\log^{2\alpha+1})$

Proof. (i) Let $\sigma_1, \ldots, \sigma_d$ denote the singular values of $\mathbf{U}_\mathbf{P}^\top \mathbf{U}_\mathbf{A}$, and let $\theta_i = \cos^{-1}(\sigma_i)$ be the principal angles. By an argument detailed in [1] (Proposition 13) which includes the use of a variant of the Davis-Kahan theorem, we have

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

for n sufficiently large. Applying the bounds from Propositions 15 and 17 then shows that

$$\begin{aligned}\left\|\mathbf{U}_\mathbf{A}\mathbf{U}_\mathbf{A}^\top - \mathbf{U}_\mathbf{P}\mathbf{U}_\mathbf{P}^\top\right\| &= \mathrm{O}\left(\frac{(\rho n)\rho T^{1/4} n^{1/2}\log^{\alpha+1/2}(n)}{\rho^2 n^2}\right)\\ &= \mathrm{O}(T^{1/4}n^{-1/2}\log^{\alpha+1/2})\end{aligned}$$

An identical argument gives the result for $\left\|\mathbf{V}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top - \mathbf{V}_\mathbf{P}\mathbf{V}_\mathbf{P}^\top\right\|$.

(ii) Using the bound from part (i), we find that

$$\left\|\mathbf{U}_\mathbf{A} - \mathbf{U}_\mathbf{P}\mathbf{U}_\mathbf{P}^\top\mathbf{U}_\mathbf{A}\right\|_F = \left\|\left(\mathbf{U}_\mathbf{A}\mathbf{U}_\mathbf{A}^\top - \mathbf{U}_\mathbf{P}\mathbf{U}_\mathbf{P}^\top\right)\mathbf{U}_\mathbf{A}\right\|_F = \mathrm{O}(T^{1/4}n^{-1/2}\log^{\alpha+1/2}(n)).$$

An identical argument bounds the term $\left\|\mathbf{V}_\mathbf{A} - \mathbf{V}_\mathbf{P}\mathbf{V}_\mathbf{P}^\top\mathbf{V}_\mathbf{A}\right\|_F$.

(iii) Observe that

$$\mathbf{U}_\mathbf{P}^\top\mathbf{U}_\mathbf{A}\mathbf{\Sigma}_\mathbf{A} - \mathbf{\Sigma}_\mathbf{P}\mathbf{V}_\mathbf{P}^\top\mathbf{V}_\mathbf{A} = \mathbf{U}_\mathbf{P}^\top(\mathbf{A}-\mathbf{P})\mathbf{V}_\mathbf{A}$$

and that we may rewrite the right-hand term to find that

$$\begin{aligned}\mathbf{U}_\mathbf{P}^\top\mathbf{U}_\mathrm{A}\mathbf{\Sigma}_\mathbf{A} - \mathbf{\Sigma}_\mathbf{P}\mathbf{V}_\mathbf{P}^\top\mathbf{V}_\mathbf{A} =& \mathbf{U}_\mathbf{P}^\top(\mathbf{A}-\mathbf{P})\left(\mathbf{V}_\mathrm{A} - \mathbf{V}_\mathbf{P}\mathbf{V}_\mathbf{P}^\top\mathbf{V}_\mathbf{A}\right)\\ &+ \mathbf{U}_\mathbf{P}^\top(\mathbf{A}-\mathbf{P})\mathbf{V}_\mathbf{P}\mathbf{V}_\mathbf{P}^\top\mathbf{V}_\mathbf{A}.\end{aligned}$$

These terms satisfy

$$\begin{aligned}\left\|\mathbf{U}_\mathbf{P}^\top(\mathbf{A}-\mathbf{P})\left(\mathbf{V}_\mathbf{A} - \mathbf{V}_\mathbf{P}\mathbf{V}_\mathbf{P}^\top\mathbf{V}_\mathbf{A}\right)\right\|_F &\leq \left\|\mathbf{U}_\mathbf{P}^\top\right\|\|(\mathbf{A}-\mathbf{P})\|\left\|\left(\mathbf{V}_\mathbf{A} - \mathbf{V}_\mathbf{P}\mathbf{V}_\mathbf{P}^\top\mathbf{V}_\mathbf{A}\right)\right\|_F\\ &= \mathrm{O}\left(\rho T^{1/2}\log^{2\alpha+1}\right)\end{aligned}$$

and

$$\begin{aligned}\|\mathbf{U}_\mathbf{P}^\top(\mathbf{A}-\mathbf{P})\mathbf{V}_\mathbf{P}\mathbf{V}_\mathbf{P}^\top\mathbf{V}_\mathbf{A}\| &\leq \|\mathbf{U}_\mathbf{P}^\top(\mathbf{A}-\mathbf{P})\mathbf{V}_\mathbf{P}\|_F\|\mathbf{V}_\mathbf{P}^\top\mathbf{V}_\mathbf{A}\|\\ &= \mathrm{O}(\rho\log^{\alpha+1/2}(n))\end{aligned}$$

by Propositions 18, 19 and the result from part (ii), and thus

$$\|\mathbf{U}_\mathbf{P}^\top\mathbf{U}_\mathbf{A}\mathbf{\Sigma}_\mathbf{A} - \mathbf{\Sigma}_\mathbf{P}\mathbf{V}_\mathbf{P}^\top\mathbf{V}_\mathbf{A}\|_F = \mathrm{O}(\rho T^{1/2}\log^{2\alpha+1}(n)).$$

An identical argument bounds the term $\left\|\mathbf{\Sigma_P U_P^\top U_A} - \mathbf{V_P^\top V_A \Sigma_A}\right\|_F$.

(iv) Note that

$$\mathbf{U_P^\top U_A} - \mathbf{V_P^\top V_A} = \left(\left(\mathbf{U_P^\top U_A \Sigma_A} - \mathbf{\Sigma_P V_P^\top V_A}\right) + \left(\mathbf{\Sigma_P U_P^\top U_A} - \mathbf{V_P^\top V_A \Sigma_A}\right)\right)\mathbf{\Sigma_A^{-1}} - \mathbf{\Sigma_P}\left(\mathbf{U_P^\top U_A} - \mathbf{V_P^\top V_A}\right)\mathbf{\Sigma_A^{-1}}.$$

For any i, j we find (after rearranging and bounding the absolute value of the right-hand terms by the Frobenius norm):

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

where we have used the result from part (iii) and Proposition 15. Consequently, we find that

$$\left|\left(\mathbf{U_P^\top U_A} - \mathbf{V_P^\top V_A}\right)_{ij}\right| = \mathrm{O}\left(\frac{\rho T^{1/2}\log^{2\alpha+1}(n)}{\rho n}\right) = \mathrm{O}(T^{1/2}n^{-1/2}\log^{2\alpha+1}(n))$$

by noting that $\left(1 + \frac{\sigma_i(\mathbf{P})}{\sigma_j(\mathbf{A})}\right) \geq 1$.

□

Recalling the use of the one-mode orthogonal Procrsutes problem to align the UASE matrices mentioned earlier, the following result (an analogue of [35], Proposition 16) utilises this.

Proposition 20. *Let* $\mathbf{U_P^\top U_A} + \mathbf{V_P^\top V_A}$ *admit the singular value decomposition*

$$\mathbf{U_P^\top U_A} + \mathbf{V_P^\top V_A} = \mathbf{W}_1\mathbf{\Sigma}\mathbf{W}_2^\top,$$

and let $\mathbf{W} = \mathbf{W}_1\mathbf{W}_2^\top$. *Then*

$$\max\left\{\left\|\mathbf{U_P^\top U_A} - \mathbf{W}\right\|_F, \left\|\mathbf{V_P^\top V_A} - \mathbf{W}\right\|_F\right\} = \mathrm{O}(T^{1/2}n^{-1}\log^{2\alpha+1}(n)).$$

almost surely.

Proof. $\mathbf{W}$ can be shown to be the solution to the solution to the one-mode orthogonal Procrustes problem [27], it minimises the term $\left\|\mathbf{U_P^\top U_A} - \mathbf{Q}\right\|_F^2 + \left\|\mathbf{V_P^\top V_A} - \mathbf{Q}\right\|_F^2$ among all $\mathbf{Q} \in \mathbb{O}(d)$. Let $\mathbf{U_P^\top U_A} = \mathbf{W_{U,1}\Sigma_U W_{U,2}^\top}$ be the singular value decomposition of

$\mathbf{U}_\mathbf{P}^\top \mathbf{U}_\mathbf{A}$, and define $\mathbf{W}_\mathbf{U} \in \mathbb{O}(d)$ by $\mathbf{W}_\mathbf{U} = \mathbf{W}_{\mathbf{U},1}\mathbf{W}_{\mathbf{U},2}^\top$. Then

$$\begin{aligned}\left\|\mathbf{U}_\mathbf{P}^\top \mathbf{U}_\mathbf{A} - \mathbf{W}_\mathbf{U}\right\|_F = \|\mathbf{\Sigma} - \mathbf{I}\|_F = \left(\sum_{i=1}^d (1-\sigma_i)^2\right)^{1/2} \leq \sum_{i=1}^d (1-\sigma_i) \\ \leq \sum_{i=1}^d \left(1-\sigma_i^2\right) = \sum_{i=1}^d \sin^2(\theta_i) \leq d\left\|\mathbf{U}_\mathbf{A}\mathbf{U}_\mathbf{A}^\top - \mathbf{U}_\mathbf{P}\mathbf{U}_\mathbf{P}^\top\right\|^2\end{aligned}$$

and so

$$\left\|\mathbf{U}_\mathbf{P}^\top \mathbf{U}_\mathbf{A} - \mathbf{W}_\mathbf{U}\right\|_F = \mathrm{O}(T^{1/2} n^{-1/2} \log^{2\alpha+1}),$$

using that the Frobenius norm of a matrix is less than or equal to the spectral norm. Also,

$$\left\|\mathbf{V}_\mathbf{P}^\top \mathbf{V}_\mathbf{A} - \mathbf{W}_\mathbf{U}\right\|_F \leq \left\|\mathbf{V}_\mathbf{P}^\top \mathbf{V}_\mathbf{A} - \mathbf{U}_\mathbf{P}^\top \mathbf{U}_\mathbf{A}\right\|_F + \left\|\mathbf{U}_\mathbf{P}^\top \mathbf{U}_\mathbf{A} - \mathbf{W}_\mathbf{U}\right\|_F$$

and so

$$\left\|\mathbf{V}_\mathbf{P}^\top \mathbf{V}_\mathbf{A} - \mathbf{W}_\mathbf{U}\right\|_F = \mathrm{O}(T^{1/2} n^{-1/2} \log^{2\alpha+1})$$

by Proposition 20. Combining these shows that

$$\left\|\mathbf{U}_\mathbf{P}^\top \mathbf{U}_\mathbf{A} - \mathbf{W}\right\|_F^2 + \left\|\mathbf{V}_\mathbf{P}^\top \mathbf{V}_\mathbf{A} - \mathbf{W}\right\|_F^2 \leq \left\|\mathbf{U}_\mathbf{P}^\top \mathbf{U}_\mathbf{A} - \mathbf{W}_\mathbf{U}\right\|_F^2 + \left\|\mathbf{V}_\mathbf{P}^\top \mathbf{V}_\mathbf{A} - \mathbf{W}_\mathbf{U}\right\|_F^2$$

and thus

$$\max\left\{\left\|\mathbf{U}_\mathbf{P}^\top \mathbf{U}_\mathbf{A} - \mathbf{W}\right\|_F, \left\|\mathbf{V}_\mathbf{P}^\top \mathbf{V}_\mathbf{A} - \mathbf{W}\right\|_F\right\} = \mathrm{O}(T^{1/2} n^{-1} \log^{2\alpha+1}(n))$$

as required. □

Proposition 21. *The following bounds hold almost surely:*

(i) $\|\mathbf{W}\mathbf{\Sigma}_\mathbf{A} - \mathbf{\Sigma}_\mathbf{P}\mathbf{W}\|_F = \mathrm{O}(\rho T^{1/2} \log^{2\alpha+1}(n))$;

(ii) $\left\|\mathbf{W}\mathbf{\Sigma}_\mathbf{A}^{1/2} - \mathbf{\Sigma}_\mathbf{P}^{1/2}\mathbf{W}\right\|_F = \mathrm{O}(\rho^{1/2} T^{1/2} n^{-1/2} \log^{2\alpha+1}(n))$;

(iii) $\left\|\mathbf{W}\mathbf{\Sigma}_\mathbf{A}^{-1/2} - \mathbf{\Sigma}_\mathbf{P}^{-1/2}\mathbf{W}\right\|_F = \mathrm{O}(\rho^{-1/2} T^{1/2} n^{-3/2} \log^{2\alpha+1}(n))$.

Proof. (i) Observe that

$$\mathbf{W}\mathbf{\Sigma}_\mathbf{A} - \mathbf{\Sigma}_\mathbf{P}\mathbf{W} = \left(\mathbf{W} - \mathbf{U}_\mathbf{P}^\top \mathbf{U}_\mathbf{A}\right)\mathbf{\Sigma}_\mathbf{A} + \mathbf{U}_\mathbf{P}^\top \mathbf{U}_\mathbf{A}\mathbf{\Sigma}_\mathbf{A} - \mathbf{\Sigma}_\mathbf{P}\mathbf{W},$$

and that the right-hand expression may be rewritten as

$$\left(\mathbf{W} - \mathbf{U}_\mathbf{P}^\top \mathbf{U}_\mathbf{A}\right)\mathbf{\Sigma}_\mathbf{A} + \left(\mathbf{U}_\mathbf{P}^\top \mathbf{U}_\mathbf{A}\mathbf{\Sigma}_\mathbf{A} - \mathbf{\Sigma}_\mathbf{P}\mathbf{V}_\mathbf{P}^\top \mathbf{V}_\mathbf{A}\right) + \mathbf{\Sigma}_\mathbf{P}\left(\mathbf{V}_\mathbf{P}^\top \mathbf{V}_\mathbf{A} - \mathbf{W}\right).$$

The Frobenius norm of each term is $\mathrm{O}\left(\rho T^{1/2} \log^{2\alpha+1}(n)\right)$ (as shown by Propositions 15, 19 and 20), so $\|\mathbf{W}\mathbf{\Sigma}_\mathbf{A} - \mathbf{\Sigma}_\mathbf{P}\mathbf{W}\|_F = \mathrm{O}\left(\rho^{1/2} T^{1/2} n^{-1/2} \log^{2\alpha+1}(n)\right)$ as required.

(ii) Note that

$$\begin{aligned}\left(\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{A}}^{1/2}-\boldsymbol{\Sigma}_{\mathbf{P}}^{1/2}\mathbf{W}\right)_{ij} &= \mathbf{W}_{ij}\left(\sigma_j(\mathbf{A})^{1/2}-\sigma_i(\mathbf{P})^{1/2}\right)\\ &= \frac{\mathbf{W}_{ij}\left(\sigma_j(\mathbf{A})-\sigma_i(\mathbf{P})\right)}{\sigma_j(\mathbf{A})^{1/2}+\sigma_i(\mathbf{P})^{1/2}}\\ &= \frac{\left(\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{A}}-\boldsymbol{\Sigma}_{\mathbf{P}}\mathbf{W}\right)_{ij}}{\sigma_j(\mathbf{A})^{1/2}+\sigma_i(\mathbf{P})^{1/2}}\end{aligned}$$

and so we find that $\left\|\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{A}}^{1/2}-\boldsymbol{\Sigma}_{\mathbf{P}}^{1/2}\mathbf{W}\right\|_F = \mathrm{O}\left(\rho^{1/2}T^{1/2}n^{-1/2}\log^{2\alpha+1}(n)\right)$ by applying part (i) and summing over all $i,j\in\{1,\ldots,d\}$.

(iii) Note that

$$\begin{aligned}\left(\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{A}}^{-1/2}-\boldsymbol{\Sigma}_{\mathbf{P}}^{-1/2}\mathbf{W}\right)_{ij} &= \frac{\mathbf{W}_{ij}\left(\sigma_i(\mathbf{P})^{1/2}-\sigma_j(\mathbf{A})^{1/2}\right)}{\sigma_i(\mathbf{P})^{1/2}\sigma_j(\mathbf{A})^{1/2}}\\ &= \frac{\left(\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{A}}^{1/2}-\boldsymbol{\Sigma}_{\mathbf{P}}^{1/2}\mathbf{W}\right)_{ij}}{\sigma_i(\mathbf{P})^{1/2}\sigma_j(\mathbf{A})^{1/2}}\end{aligned}$$

and so we find that $\left\|\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{A}}^{-1/2}-\boldsymbol{\Sigma}_{\mathbf{P}}^{-1/2}\mathbf{W}\right\|_F = \mathrm{O}\left(\rho^{-1/2}T^{1/2}n^{-3/2}\log^{2\alpha+1}(n)\right)$ by applying part (ii) and summing over all $i,j\in\{1,\ldots,d\}$.

□

Proposition 22. *If* $\mathbf{X}$ *is of rank* d *then there exists a matrix* $\widetilde{\mathbf{L}}\in\mathrm{GL}(d)$ *such that* $\mathbf{X}_{\mathbf{P}}=\mathbf{X}\widetilde{\mathbf{L}}$. *In addition, if* $\mathbf{Y}^{(t)}$ *is of rank* d_t *then* $\mathbf{Y}_{\mathbf{P}}^{(t)}=\mathbf{Y}^{(t)}\widetilde{\mathbf{R}}_t$, *where the matrix* $\widetilde{\mathbf{R}}_t\in\mathbb{R}^{d_t\times d}$ *satisfies* $\widetilde{\mathbf{L}}\widetilde{\mathbf{R}}_t^\top=\boldsymbol{\Lambda}_t$. *In particular,* $\mathrm{rank}\left(\widetilde{\mathbf{R}}_t\right)=\mathrm{rank}\left(\boldsymbol{\Lambda}_t\right)$.

Proof. As in [1], Proposition 16. □

Corollary 23. *The matrices* $\widetilde{\mathbf{L}}$ *and* $\widetilde{\mathbf{R}}_t$ *satisfy* $\|\widetilde{\mathbf{L}}\|=\left\|\widetilde{\mathbf{L}}^{-1}\right\|=\left\|\widetilde{\mathbf{R}}_t\right\|=\mathrm{O}\left(1\right)$ *almost surely.*

Moreover, if the matrix $\boldsymbol{\Lambda}_t$ *is of rank* d_t, *then* $\left\|\widetilde{\mathbf{R}}_t^+\right\|=\mathrm{O}\left(1\right)$ *almost surely, where* $\widetilde{\mathbf{R}}_t^+=\widetilde{\mathbf{R}}_t^\top\left(\widetilde{\mathbf{R}}_t\widetilde{\mathbf{R}}_t^\top\right)^{-1}$ *is the Moore-Penrose inverse of* $\widetilde{\mathbf{R}}_t$.

Proof. Proposition 15 shows us that $\|\boldsymbol{\Sigma}_{\mathbf{P}}\|=\mathrm{O}(\rho n)$ and $\left\|\boldsymbol{\Sigma}_{\mathbf{P}}^{-1}\right\|=\mathrm{O}\left(\frac{1}{\rho n}\right)$, and an identical line of reasoning shows that $\|\Pi_{\boldsymbol{\Lambda},\mathbf{Y}}\|=\mathrm{O}\left(\rho^{1/2}n^{1/2}\right)$ and $\left\|\Pi_{\boldsymbol{\Lambda},\mathbf{Y}}^{-1}\right\|=\mathrm{O}\left(\frac{1}{\rho^{1/2}n^{1/2}}\right)$, and similarly that $\|\Pi_{\mathbf{X}}\|=\mathrm{O}\left(\rho^{1/2}n^{1/2}\right)$ and $\left\|\Pi_{\mathbf{X}}^{-1}\right\|=\mathrm{O}\left(\frac{1}{\rho^{1/2}n^{1/2}}\right)$.

The rest of the proof follows in the same way as [1], Corollary 17, exploiting various properties of the spectral norm and of Hermitian matrices (real, symmetric matrices in this context). □

Proposition 24. *If* $\mathbf{X}$ *is of rank* d *and each* $\mathbf{Y}^{(t)}$ *is of rank* d_t *then*

$$\left(\widetilde{\mathbf{R}}_t\boldsymbol{\Sigma}_{\mathbf{P}}^{-1}\widetilde{\mathbf{L}}^{-1}\right)^\top=\left(\boldsymbol{\Lambda}\mathbf{Y}^\top\mathbf{Y}\boldsymbol{\Lambda}^\top\right)^{-1}\boldsymbol{\Lambda}_t.$$

Moreover, if $d_r = d$ and the matrix $\mathbf{\Lambda}_r$ is invertible, then

$$\left(\tilde{\mathbf{L}}\mathbf{\Sigma}_{\mathbf{P}}^{-1}\widetilde{\mathbf{R}}_t^{-1}\right)^{\top} = \mathbf{\Lambda}_t^{-1}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}.$$

Proof. Equivalent to that in [1], Proposition 18. □

Proposition 25. *Let*

$$\begin{aligned}
\mathbf{R}_{1,1} &= \mathbf{U}_{\mathbf{P}}\left(\mathbf{U}_{\mathbf{P}}^{\top}\mathbf{U}_{\mathbf{A}}\mathbf{\Sigma}_{\mathbf{A}}^{1/2} - \mathbf{\Sigma}_{\mathbf{P}}^{1/2}\mathbf{W}\right)\\
\mathbf{R}_{1,2} &= \left(\mathbf{I} - \mathbf{U}_{\mathbf{P}}\mathbf{U}_{\mathbf{P}}^{\top}\right)(\mathbf{A} - \mathbf{P})\left(\mathbf{V}_{\mathbf{A}} - \mathbf{V}_{\mathbf{P}}\mathbf{W}\right)\mathbf{\Sigma}_{\mathbf{A}}^{-1/2}\\
\mathbf{R}_{1,3} &= -\mathbf{U}_{\mathbf{P}}\mathbf{U}_{\mathbf{P}}^{\top}(\mathbf{A} - \mathbf{P})\mathbf{V}_{\mathbf{P}}\mathbf{W}\mathbf{\Sigma}_{\mathbf{A}}^{-1/2}\\
\mathbf{R}_{1,4} &= (\mathbf{A} - \mathbf{P})\mathbf{V}_{\mathbf{P}}\left(\mathbf{W}\mathbf{\Sigma}_{\mathbf{A}}^{-1/2} - \mathbf{\Sigma}_{\mathbf{P}}^{-1/2}\mathbf{W}\right)
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{R}_{2,1} &= \mathbf{V}_{\mathbf{P}}\left(\mathbf{V}_{\mathbf{P}}^{\top}\mathbf{V}_{\mathbf{A}}\mathbf{\Sigma}_{\mathbf{A}}^{1/2} - \mathbf{\Sigma}_{\mathbf{P}}^{1/2}\mathbf{W}\right)\\
\mathbf{R}_{2,2} &= \left(\mathbf{I} - \mathbf{V}_{\mathbf{P}}\mathbf{V}_{\mathbf{P}}^{\top}\right)(\mathbf{A} - \mathbf{P})^{\top}\left(\mathbf{U}_{\mathbf{A}} - \mathbf{U}_{\mathbf{P}}\mathbf{W}\right)\mathbf{\Sigma}_{\mathbf{A}}^{-1/2}\\
\mathbf{R}_{2,3} &= -\mathbf{V}_{\mathbf{P}}\mathbf{V}_{\mathbf{P}}^{\top}(\mathbf{A} - \mathbf{P})^{\top}\mathbf{U}_{\mathbf{P}}\mathbf{W}\mathbf{\Sigma}_{\mathbf{A}}^{-1/2}\\
\mathbf{R}_{2,4} &= (\mathbf{A} - \mathbf{P})^{\top}\mathbf{U}_{\mathbf{P}}\left(\mathbf{W}\mathbf{\Sigma}_{\mathbf{A}}^{-1/2} - \mathbf{\Sigma}_{\mathbf{P}}^{-1/2}\mathbf{W}\right)
\end{aligned}$$

Then the following bounds hold almost surely:

(i) $\|\mathbf{R}_{1,1}\|_{2\to\infty}, \|\mathbf{R}_{2,1}\|_{2\to\infty} = \mathrm{O}\left(\frac{\rho^{1/2}T^{1/2}\log^{2\alpha+1}(n)}{n}\right)$;

(ii) $\|\mathbf{R}_{1,2}\|_{2\to\infty}, \|\mathbf{R}_{2,2}\|_{2\to\infty} = \mathrm{O}\left(\frac{\rho^{1/2}T^{1/2}\log^{2\alpha+1}(n)}{n^{3/4}}\right)$;

(iii) $\|\mathbf{R}_{1,3}\|_{2\to\infty}, \|\mathbf{R}_{2,3}\|_{2\to\infty} = \mathrm{O}\left(\frac{\log^{\alpha+1/2}(n)}{\rho^{1/2}n}\right)$;

(iv) $\|\mathbf{R}_{1,4}\|_{2\to\infty}, \|\mathbf{R}_{2,4}\|_{2\to\infty} = \mathrm{O}\left(\frac{\rho^{1/2}T^{3/4}\log^{3\alpha+3/2}(n)}{n}\right)$.

Proof. Similarly to what was mentioned previously, full proofs will only be given for the $\mathbf{R}_{1,i}$, noting any differences for the proofs for the terms $\mathbf{R}_{2,i}$.

(i) Recall that $\mathbf{U}_{\mathbf{P}}\mathbf{\Sigma}_{\mathbf{P}}^{1/2} = \mathbf{X}\widetilde{\mathbf{L}}$, where the matrix $\widetilde{\mathbf{L}} \in \mathrm{GL}(d)$ satisfies $\|\widetilde{\mathbf{L}}\| = \mathrm{O}(1)$ by Corollary 23. Using the relation $\|\mathbf{AB}\|_{2\to\infty} \leq \|\mathbf{A}\|_{2\to\infty}\|\mathbf{B}\|$ (see, for example, [28], Proposition 6.5) we find that

$$\|\mathbf{U}_{\mathbf{P}}\|_{2\to\infty} \leq \|\mathbf{X}\|_{2\to\infty}\|\widetilde{\mathbf{L}}\|\|\mathbf{\Sigma}_{\mathbf{P}}^{-1/2}\|,$$

and thus $\|\mathbf{U}_{\mathbf{P}}\|_{2\to\infty} = \mathrm{O}\left(n^{-1/2}\right)$ as the rows of $\mathbf{X}$ are by definition of order $\mathrm{O}\left(\rho^{1/2}\right)$ (n increasing doesn't result in the length of each row of $\mathbf{X}$ increasing, so the Euclidean norm of a row of $\mathbf{X}$ doesn't grow with n). Similarly, we find that $\|\mathbf{V}_{\mathbf{P}}\|_{2\to\infty} = \mathrm{O}\left(n^{-1/2}\right)$ by splitting $\mathbf{V}_{\mathbf{P}}\mathbf{\Sigma}_{\mathbf{P}}^{1/2}$ into the separate terms $\mathbf{V}_{\mathbf{P}}^{(r)}\mathbf{\Sigma}_{\mathbf{P}}^{1/2}$ and evaluating each separately.

Thus

$$\begin{aligned}\|\mathbf{R}_{1,1}\|_{2\to\infty} &\leq \|\mathbf{U_P}\|_{2\to\infty} \left\|\mathbf{U_P^\top U_A \Sigma_A^{1/2}} - \mathbf{\Sigma_P^{1/2} W}\right\| \\ &\leq \|\mathbf{U_P}\|_{2\to\infty} \left(\left\|\left(\mathbf{U_P^\top U_A} - \mathbf{W}\right)\mathbf{\Sigma_A^{1/2}}\right\|_F + \left\|\mathbf{W\Sigma_A^{1/2}} - \mathbf{\Sigma_P^{1/2} W}\right\|_F\right)\end{aligned}$$

The first term is $\mathrm{O}\left(\rho^{1/2}T^{1/2}n^{-1/2}\log^{2\alpha+1}(n)\right)$ by Propositions 20 and 15, while Proposition 21 shows that the second is also $\mathrm{O}\left(\rho^{1/2}T^{1/2}n^{-1/2}\log^{2\alpha+1}(n)\right)$, and so

$$\|\mathbf{R}_{1,1}\|_{2\to\infty} = \mathrm{O}\left(\rho^{1/2}T^{1/2}n^{-1}\log^{2\alpha+1}(n)\right)$$

(ii) Begin by splitting the term $\mathbf{R}_{1,2} = \mathbf{M}_1 + \mathbf{M}_2$, where

$$\begin{aligned}\mathbf{M}_1 &= \mathbf{U_P U_P^\top}(\mathbf{A} - \mathbf{P})\left(\mathbf{V_A} - \mathbf{V_P W}\right)\mathbf{\Sigma_A^{-1/2}} \\ \mathbf{M}_2 &= (\mathbf{A} - \mathbf{P})\left(\mathbf{V_A} - \mathbf{V_P W}\right)\mathbf{\Sigma_A^{-1/2}}\end{aligned}$$

Now,

$$\|\mathbf{M}_1\|_{2\to\infty} \leq \|\mathbf{U_P}\|_{2\to\infty}\|\mathbf{A} - \mathbf{P}\|\,\|\mathbf{V_A} - \mathbf{V_P W}\|\,\|\mathbf{\Sigma_A^{-1/2}}\|$$

where the term

$$\|\mathbf{U_P}\|_{2\to\infty}\|\mathbf{A} - \mathbf{P}\|\|\mathbf{\Sigma_A^{-1/2}}\| = \mathrm{O}\left(\rho^{1/2}T^{1/4}n^{-1/2}\log^{\alpha+1/2}(n)\right)$$

by Propositions 15 and 17, while

$$\begin{aligned}\|\mathbf{V_A} - \mathbf{V_P W}\| &\leq \left\|\mathbf{V_A} - \mathbf{V_P V_P^\top V_A}\right\| + \left\|\mathbf{V_P}\left(\mathbf{V_P^\top V_A} - \mathbf{W}\right)\right\| \\ &= \mathrm{O}\left(T^{1/4}n^{-1/2}\log^{\alpha+1/2}(n)\right)\end{aligned}$$

where the first term is $\mathrm{O}\left(T^{1/4}n^{-1/2}\log^{\alpha+1/2}(n)\right)$ and the second $\mathrm{O}\left(T^{1/2}n^{-1}\log^{2\alpha+1}(n)\right)$ by Propositions 19 and 20 and the asymptotic growth conditions imposed on ρ. Thus

$$\|\mathbf{M}_1\|_{2\to\infty} = \mathrm{O}\left(\rho^{1/2}T^{1/2}n^{-1}\log^{2\alpha+1}(n)\right)$$

Next, note that

$$\mathbf{M}_2 = (\mathbf{A} - \mathbf{P})\left(\mathbf{I} - \mathbf{V_P V_P^\top}\right)\mathbf{V_A}\mathbf{\Sigma_A^{-1/2}} + (\mathbf{A} - \mathbf{P})\mathbf{V_P}\left(\mathbf{V_P^\top V_A} - \mathbf{W}\right)\mathbf{\Sigma_A^{-1/2}},$$

where

$$\begin{aligned}\left\|(\mathbf{A} - \mathbf{P})\mathbf{V_P}\left(\mathbf{V_P^\top V_A} - \mathbf{W}\right)\mathbf{\Sigma_A^{-1/2}}\right\|_{2\to\infty} &\leq \|\mathbf{A} - \mathbf{P}\|\left\|\mathbf{V_P^\top V_A} - \mathbf{W}\right\|\left\|\mathbf{\Sigma_A^{-1/2}}\right\| \\ &= \mathrm{O}\left(\rho^{1/2}T^{3/4}n^{-1}\log^{3(\alpha+1/2)}(n)\right)\end{aligned}$$

by Propositions 17, 20, and 15.

To bound the remaining term, let $\mathbf{M} = (\mathbf{A} - \mathbf{P})\left(\mathbf{I} - \mathbf{V_P}\mathbf{V_P}^\top\right)\mathbf{V_A}\mathbf{V_A}^\top$, so that

$$(\mathbf{A} - \mathbf{P})\left(\mathbf{I} - \mathbf{V_P}\mathbf{V_P}^\top\right)\mathbf{V_A}\mathbf{\Sigma_A}^{-1/2} = \mathbf{M}\mathbf{V_A}\mathbf{\Sigma_A}^{-1/2}$$

and so

$$\left\|(\mathbf{A} - \mathbf{P})\left(\mathbf{I} - \mathbf{V_P}\mathbf{V_P}^\top\right)\mathbf{V_A}\mathbf{\Sigma_A}^{-1/2}\right\|_{2\to\infty} \leq \|\mathbf{M}\|_{2\to\infty}\left\|\mathbf{V_A}\mathbf{\Sigma_A}^{-1/2}\right\|$$

The term $\|\mathbf{V_A}\mathbf{\Sigma_A}^{-1/2}\|$ is $\mathrm{O}\left(\rho^{-1/2}n^{-1/2}\right)$ by Proposition 15, so it suffices to bound $\|\mathbf{M}\|_{2\to\infty}$. To do this, we claim that the Frobenius norms of the rows of the matrix $\mathbf{M}$ are exchangeable (see [1], Proposition 19), and thus have the same expectation, which implies that $\mathbb{E}\left(\|\mathbf{M}\|_F^2\right) = n\mathbb{E}\left(\|\mathbf{M}_i\|^2\right)$ for any $i \in \{1, \ldots, n\}$. Applying Markov's inequality, we then see that

$$\mathbb{P}\left(\|\mathbf{M}_i\| > t\right) \leq \frac{\mathbb{E}\left(\|\mathbf{M}_i\|^2\right)}{t^2} = \frac{\mathbb{E}\left(\|\mathbf{M}\|_F^2\right)}{nt^2}.$$

Now,

$$\begin{aligned}\|\mathbf{M}\|_F &\leq \|\mathbf{A} - \mathbf{P}\|\left\|\mathbf{V_A} - \mathbf{V_P}\mathbf{V_P}^\top\mathbf{V_A}\right\|_F\left\|\mathbf{V_A}^\top\right\|_F \\ &= \mathrm{O}\left(\rho T^{1/2}n^{-1/2}\log^{2\alpha+1}(n)\right)\end{aligned}$$

by Propositions 17 and 19. It follows that

$$\mathbb{P}\left(\|\mathbf{M}_i\| > \rho T^{1/2}n^{-1/4}\log^{2\alpha+1}(n)\right) = \mathrm{O}\left(n^{-1/2}\right)$$

and thus

$$\|\mathbf{M}\|_{2\to\infty} = \mathrm{O}\left(\rho T^{1/2}n^{-1/4}\log^{2\alpha+1}(n)\right)$$

and

$$\left\|(\mathbf{A} - \mathbf{P})\left(\mathbf{I} - \mathbf{V_P}\mathbf{V_P}^\top\right)\mathbf{V_A}\mathbf{\Sigma_A}^{-1/2}\right\|_{2\to\infty} = \mathrm{O}\left(\rho^{1/2}T^{1/2}n^{-3/2}\log^{2\alpha+1}(n)\right)$$

almost surely.

Combining these results, we see that

$$\|\mathbf{R}_{1,2}\|_{2\to\infty} = \mathrm{O}\left(\rho^{1/2}T^{1/2}n^{-3/2}\log^{2\alpha+1}(n)\right)$$

almost surely, as required.

The proof of the bound for the term $\mathbf{R}_{2,2}$ follows similarly, and culminates in showing that the term

$$\mathbf{N} = (\mathbf{A} - \mathbf{P})^\top\left(\mathbf{I} - \mathbf{U_P}\mathbf{U_P}^\top\right)\mathbf{U_A}\mathbf{U_A}^\top$$

satisfies

$$\|\mathbf{N}\|_{2\to\infty} = \mathrm{O}\left(\rho T^{1/2} n^{-1/4} \log^{2\alpha+1}(n)\right)$$

almost surely.

(iii) Similarly to part (i), we see that

$$\begin{aligned}
\|\mathbf{R}_{1,3}\|_{2\to\infty} &\leq \|\mathbf{U_P}\|_{2\to\infty} \left\|\mathbf{U_P^\top}(\mathbf{A}-\mathbf{P})\mathbf{V_P}\mathbf{W}\mathbf{\Sigma_A^{-1/2}}\right\| \\
&\leq \|\mathbf{U_P}\|_{2\to\infty} \left\|\mathbf{U_P^\top}(\mathbf{A}-\mathbf{P})\mathbf{V_P}\right\|_F \left\|\mathbf{W}\mathbf{\Sigma_A^{-1/2}}\right\|_F \\
&= \mathrm{O}\left(\rho^{-1/2} n^{-1} \log^{\alpha+1/2}(n)\right)
\end{aligned}$$

by Propositions 21 and 15.

(iv) Observe that

$$\begin{aligned}
\|\mathbf{R}_{1,4}\|_{2\to\infty} &\leq \|\mathbf{R}_{1,4}\|_F \\
&\leq \|\mathbf{A}-\mathbf{P}\| \, \|\mathbf{V_P}\|_F \left\|\mathbf{W}\mathbf{\Sigma_A^{-1/2}} - \mathbf{\Sigma_P^{-1/2}}\mathbf{W}\right\|_F \\
&= \mathrm{O}\left(\rho^{1/2} T^{3/4} n^{-} \log^{3(\alpha+1/2)}(n)\right)
\end{aligned}$$

by Propositions 17 and 21 .

□

6.4 Proof of Theorem 5

Proof. We first consider the left embedding $\mathbf{X_A}$. Observe that

$$\begin{aligned}
\mathbf{X_A} - \mathbf{X_P}\mathbf{W} &= \mathbf{U_A}\mathbf{\Sigma_A^{1/2}} - \mathbf{U_P}\mathbf{\Sigma_P^{1/2}}\mathbf{W} \\
&= \mathbf{U_A}\mathbf{\Sigma_A^{1/2}} - \mathbf{U_P}\mathbf{U_P^\top}\mathbf{U_A}\mathbf{\Sigma_A^{1/2}} + \mathbf{U_P}\left(\mathbf{U_P^\top}\mathbf{U_A}\mathbf{\Sigma_A^{1/2}} - \mathbf{\Sigma_P^{1/2}}\mathbf{W}\right) \\
&= \mathbf{U_A}\mathbf{\Sigma_A^{1/2}} - \mathbf{U_P}\mathbf{U_P^\top}\mathbf{U_A}\mathbf{\Sigma_A^{1/2}} + \mathbf{R}_1
\end{aligned}$$

Noting that $\mathbf{U_A}\mathbf{\Sigma_A^{1/2}} = \mathbf{A}\mathbf{V_A}\mathbf{\Sigma_A^{-1/2}}$ and $\mathbf{U_P}\mathbf{U_P^\top}\mathbf{P} = \mathbf{P}$, we see that

$$\begin{aligned}
\mathbf{X_A} - \mathbf{X_P}\mathbf{W} &= \mathbf{A}\mathbf{V_A}\mathbf{\Sigma_A^{-1/2}} - \mathbf{U_P}\mathbf{U_P^\top}\mathbf{A}\mathbf{V_A}\mathbf{\Sigma_A^{-1/2}} + \mathbf{R}_{1,1} \\
&= (\mathbf{A}-\mathbf{P})\mathbf{V_A}\mathbf{\Sigma_A^{-1/2}} - \left(\mathbf{U_P}\mathbf{U_P^\top}\mathbf{A} - \mathbf{P}\right)\mathbf{V_A}\mathbf{\Sigma_A^{-1/2}} + \mathbf{R}_{1,1} \\
&= (\mathbf{A}-\mathbf{P})\mathbf{V_A}\mathbf{\Sigma_A^{-1/2}} - \mathbf{U_P}\mathbf{U_P^\top}(\mathbf{A}-\mathbf{P})\mathbf{V_A}\mathbf{\Sigma_A^{-1/2}} + \mathbf{R}_{1,1} \\
&= \left(\mathbf{I} - \mathbf{U_P}\mathbf{U_P^\top}\right)(\mathbf{A}-\mathbf{P})\mathbf{V_A}\mathbf{\Sigma_A^{-1/2}} + \mathbf{R}_{1,1} \\
&= \left(\mathbf{I} - \mathbf{U_P}\mathbf{U_P^\top}\right)(\mathbf{A}-\mathbf{P})\left(\mathbf{V_P}\mathbf{W} + (\mathbf{V_A} - \mathbf{V_P}\mathbf{W})\right)\mathbf{\Sigma_A^{-1/2}} + \mathbf{R}_{1,1} \\
&= (\mathbf{A}-\mathbf{P})\mathbf{V_P}\mathbf{W}\mathbf{\Sigma_A^{-1/2}} + \mathbf{R}_{1,3} + \mathbf{R}_{1,2} + \mathbf{R}_{1,1} \\
&= (\mathbf{A}-\mathbf{P})\mathbf{V_P}\mathbf{\Sigma_P^{-1/2}}\mathbf{W} + \mathbf{R}_{1,4} + \mathbf{R}_{1,3} + \mathbf{R}_{1,2} + \mathbf{R}_{1,1}
\end{aligned}$$

Applying Proposition 25, we find that

$$\begin{aligned}\|\mathbf{X_A} - \mathbf{X_P W}\|_{2\to\infty} &= \left\|(\mathbf{A}-\mathbf{P})\mathbf{V_P}\mathbf{\Sigma_P}^{-1/2}\right\|_{2\to\infty} + \mathrm{O}\left(\rho^{1/2}T^{1/2}n^{-3/4}\log^{2\alpha+1}(n)\right) \\ &\leq \sigma_d(\mathbf{P})^{-1/2}\left\|(\mathbf{A}-\mathbf{P})\mathbf{V_P}\right\|_{2\to\infty} + \mathrm{O}\left(\rho^{1/2}T^{1/2}n^{-3/4}\log^{2\alpha+1}(n)\right)\end{aligned}$$

almost surely.

We condition on some set of latent positions. For any $i \in \{1,\ldots,n\}, j \in \{1,\ldots,d\}$ and $t \in \{1,\ldots,T\}$, let

$$\mathbf{E}_{ij}^{(t)} = \sum_{l\neq i}\left(\mathbf{A}_{il}^{(t)} - \mathbf{P}_{il}^{(t)}\right)v_l^{(t)}.$$

and

$$\mathbf{F}_{ij}^{(t)} = \mathbf{P}_{ii}^{(t)}v_i^{(t)},$$

where $v^{(t)}$ denotes the j th column of $\mathbf{V}_{\mathbf{P}}^{(t)}$, so that

$$((\mathbf{A}-\mathbf{P})\mathbf{V_P})_{ij} = \sum_{t=1}^{T}\mathbf{E}_{ij}^{(t)} - \sum_{t=1}^{T}\mathbf{F}_{ij}^{(t)}.$$

The latter term is of order $\mathrm{O}\left(\rho T^{1/2}\right)$(as can be seen by applying the Cauchy-Schwarz inequality and using orthonormality) and thus can be discounted from our asymptotic analysis. The former is a sum of independent, zero-mean random variables, with each individual term bounded in absolute value by $|v_l^{(t)}|$, and thus we can apply Hoeffding's inequality and the orthonormality of singular vectors to see that

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{t=1}^{T}\mathbf{E}_{ij}^{(t)}\right| > \tau\right) &\leq 2\exp\left(\frac{-2\tau^2}{4\beta^2\rho^2\log^{2\alpha}(n)\sum_{t=1}^{T}\sum_{l=1}^{n_t}|v_l^{(t)}|^2}\right) \\ &= 2\exp\left(\frac{-\tau^2}{2\beta^2\rho^2\log^{2\alpha}(n)}\right).\end{aligned}$$

Since the numerator dominates when $\tau = \rho\log^{\alpha+1/2}(n)$, we see that $((\mathbf{A}-\mathbf{P})\mathbf{V_P})_{ij} = \mathrm{O}\left(\rho\log^{\alpha+1/2}(n)\right)$ almost surely, and hence $|((\mathbf{A}-\mathbf{P})\mathbf{V_P})_i| = \mathrm{O}\left(\rho\log^{\alpha+1/2}(n)\right)$ almost surely by summing over all $j \in \{1,\ldots,d\}$. Taking the union bound over all n rows then shows that

$$\sigma_d(\mathbf{P})^{-1/2}\left\|(\mathbf{A}-\mathbf{P})\mathbf{V_P}\right\|_{2\to\infty} = \mathrm{O}\left(\frac{\log^{1/2}(n)}{\rho^{1/2}n^{1/2}}\right)$$

almost surely, and consequently that

$$\|\mathbf{X_A} - \mathbf{XL}\|_{2\to\infty} = \mathrm{O}\left(\frac{\log^{1/2}(n)}{\rho^{1/2}n^{1/2}}\right)$$

almost surely by setting $\mathbf{L} = \widetilde{\mathbf{L}}\mathbf{W}$. The second bound follows from Corollary 23 and the fact that $\|\mathbf{AB}\|_{2\to\infty} \leq \|\mathbf{A}\|_{2\to\infty}\|\mathbf{B}\|$. Integrating over all possible sets of latent positions

gives the result.

A similar argument is used for the right embedding, culminating with the result:

$$\left\|\mathbf{Y}_{\mathbf{A}}^{(t)} - \mathbf{Y}\widetilde{\mathbf{R}}_t\mathbf{W}\right\|_{2\to\infty} = \mathrm{O}\left(\frac{\log^{1/2}(n)}{\rho^{1/2}n^{1/2}}\right).$$

Recalling that $\tilde{\mathbf{Y}} = \mathbf{Y}_{\mathbf{P}}^{(t)} = \mathbf{Y}\widetilde{\mathbf{R}}_t$ from Proposition 22 gives us the form of the result. Proposition 22 also shows that $\widetilde{\mathbf{L}}\widetilde{\mathbf{R}}_t^\top = \mathbf{\Lambda}_t$, and the remaining bounds follow from Corollary 23 as for the left embedding. Integrating over all possible sets of latent positions gives the final result. □

6.5 Proof of Theorem 6

Recall that $\mathbf{X}_{\mathbf{P}} = \mathbf{X}\tilde{\mathbf{L}}$, and that $\mathbf{L} = \widetilde{\mathbf{L}}\mathbf{W}$. Then using the orthogonality of these matrices and recalling from the proof of Theorem 4 (ignoring the residual terms for now),

$$\begin{aligned}
\mathbf{X}_{\mathbf{A}} - \mathbf{X}_{\mathbf{P}}\mathbf{W} &= (\mathbf{A}-\mathbf{P})\mathbf{V}_{\mathbf{P}}\mathbf{\Sigma}_{\mathbf{P}}^{-1/2}\mathbf{W} \\
\Rightarrow \mathbf{X}_{\mathbf{A}} - \mathbf{X}\tilde{\mathbf{L}}\mathbf{W} &= (\mathbf{A}-\mathbf{P})\mathbf{V}_{\mathbf{P}}\mathbf{\Sigma}_{\mathbf{P}}^{-1/2}\mathbf{W} \\
\Rightarrow \mathbf{X}_{\mathbf{A}}\mathbf{W}^{-1} - \mathbf{X}\tilde{\mathbf{L}} &= (\mathbf{A}-\mathbf{P})\mathbf{V}_{\mathbf{P}}\mathbf{\Sigma}_{\mathbf{P}}^{-1/2} \\
\Rightarrow \mathbf{X}_{\mathbf{A}}\mathbf{L}^{-1} - \mathbf{X} &= (\mathbf{A}-\mathbf{P})\mathbf{V}_{\mathbf{P}}\mathbf{\Sigma}_{\mathbf{P}}^{-1/2}\tilde{\mathbf{L}}^{-1}.
\end{aligned}$$

Now, consider

$$n^{1/2}(\mathbf{X}_{\mathbf{A}}\mathbf{L}^{-1} - \mathbf{X}) = n^{1/2}(\mathbf{A}-\mathbf{P})\mathbf{V}_{\mathbf{P}}\mathbf{\Sigma}_{\mathbf{P}}^{-1/2}\tilde{\mathbf{L}}^{-1} + n^{1/2}\mathbf{R},$$

where the residual term $\mathbf{R} = \mathbf{R}_{1,4} + \mathbf{R}_{1,3} + \mathbf{R}_{1,2} + \mathbf{R}_{1,1}$ from the proof of the previous theorem. We then see that

$$\|n^{1/2}\mathbf{R}\|_{2\to\infty} = \mathrm{O}\left(\rho^{1/2}T^{1/2}n^{-1/4}\log^{2\alpha+1}(n)\right) \to 0.$$

The proof then follows identically to the corresponding proof in [1] (Theorem 3), culminating in the use of the multivariate central limit theorem to give the desired result.

6.6 Proof of Corollary 8

A full proof is not given here, however, following the proofs for Proposition 3 and Proposition 5 from the supplemental material of I. Gallagher et al.'s paper [2], one can rewrite the central limit result here in a way that shows the existence of matrices $\mathbf{W}$, $\tilde{\mathbf{W}}$, and $\mathbf{R}_*$ which are independent of t. The exchangeability property follows from this.