To be read before the Royal Statistical Society at an online meeting on October 8th 2025, from 16.00, with Maria De Iorio in the Chair

Statistical exploration of the Manifold Hypothesis

Nick Whiteley*[a], Annie Gray[b], and Patrick Rubin-Delanchy[c]

[a]School of Mathematics, University of Bristol, United Kingdom
[b]The Alan Turing Institute, United Kingdom
[c]School of Mathematics, University of Edinburgh, United Kingdom

August 28, 2025

Abstract

The Manifold Hypothesis is a widely accepted tenet of Machine Learning which asserts that nominally high-dimensional data are in fact concentrated near a low-dimensional manifold, embedded in high-dimensional space. This phenomenon is observed empirically in many real world situations, has led to development of a wide range of statistical methods in the last few decades, and has been suggested as a key factor in the success of modern AI technologies. We show that rich and sometimes intricate manifold structure in data can emerge from a generic and remarkably simple statistical model — the Latent Metric Space model — via elementary concepts such as latent variables, correlation and stationarity. This establishes a general statistical explanation for why the Manifold Hypothesis seems to hold in so many situations. Informed by the Latent Metric Space model we derive procedures to discover and interpret the geometry of high-dimensional data, and explore hypotheses about the data generating mechanism. These procedures operate under minimal assumptions and make use of well-known dimension reduction methods and graph-analytic algorithms.

1 Introduction

The manifold hypothesis is a widely accepted tenet of Machine Learning which posits that [Cayton, 2005]:

"*...the dimensionality of many data sets is only artificially high; though each data point consists of perhaps thousands of features, it may be described as a function of only a few underlying parameters. That is, the data points are actually samples from a low-dimensional manifold that is embedded in a high-dimensional space*".

This phenomenon has impacted a wide range of methods and algorithms. Presence of manifold structure is the premise of manifold estimation and testing [Fefferman et al., 2016, Genovese et al., 2012b,a], nonlinear dimension reduction techniques [Roweis and Saul, 2000, Tenenbaum et al., 2000, Hinton and Roweis, 2002, Belkin and Niyogi, 2003, Weinberger et al., 2004, Van der Maaten and Hinton, 2008, McInnes et al., 2018], intrinsic dimension estimation [Kégl, 2002, Levina and Bickel, 2004, Hein and Audibert, 2005, Carter et al., 2009], and regression and classification techniques specially adapted to settings in which covariates are valued on manifolds [Bickel and Li, 2007, Aswani et al., 2011, Cheng and Wu, 2013, Yang and Dunson, 2016, Lin et al., 2019, Niu et al., 2019]. Assumptions that data are concentrated near low-dimensional topological or geometric structures underpin clustering techniques and topological data analysis [Edelsbrunner and Harer, 2008, Niyogi et al., 2008, Carlsson, 2009, Balakrishnan et al., 2012, Chazal et al., 2013, Chazal and Michel, 2021]. Some nonparametric techniques, such as nearest neighbour or

* *Corresponding author:* Prof. Nick Whiteley, School of Mathematics, Woodland Rd, Bristol BS8 1UG, United Kingdom; email: nick.whiteley@bristol.ac.uk

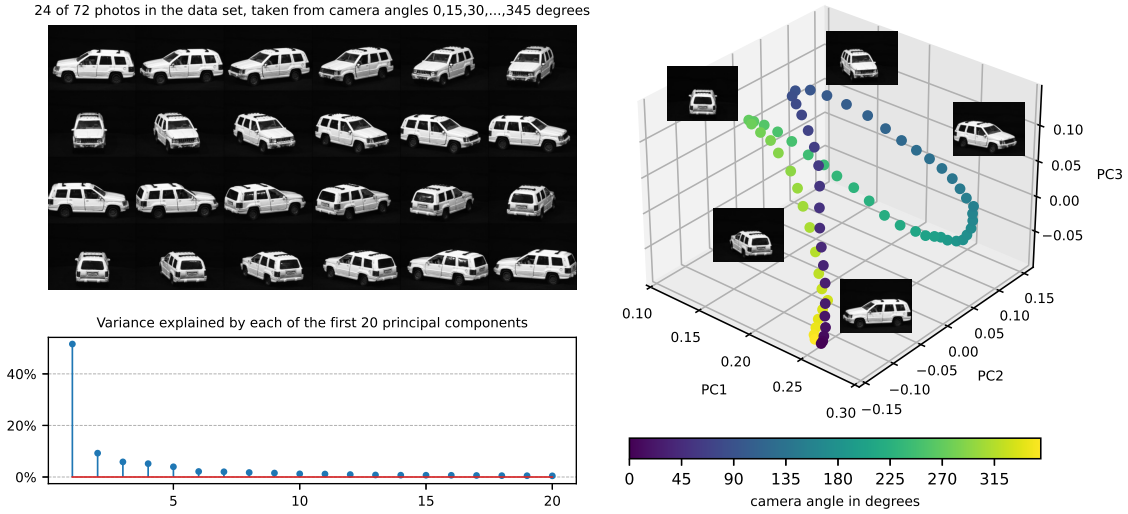

Figure 1: A collection of images reduced in dimension using PCA.

tree-based regression methods, function without manifold structure necessarily being present, but benefit significantly when it is there, since their convergence rates depend on intrinsic rather than ambient dimension of covariates [Kpotufe, 2011, Kpotufe and Dasgupta, 2012]. It has been proved that deep neural networks exhibit a similar property [Nakada and Imaizumi, 2020] and, more broadly, the presence of manifold structure has been suggested as a key factor in the success of deep learning methods [Brahma et al., 2015]. Assumptions that data lie on a low-dimensional manifold embedded in high-dimensional space are central to very recent practical and theoretical developments in generative modelling in AI, especially diffusion models [Song and Ermon, 2019, Song et al., 2020, Ho et al., 2020, De Bortoli, 2022, De Bortoli et al., 2022, Stanczuk et al., 2022, Chung et al., 2022, Pidstrigach, 2022, He et al., 2023, Elhag et al., 2023]. Manifold structure is also found in internal high-dimensional representations constructed by large neural networks such as language models, and thought to reflect machine encodings of basic human-understandable concepts such as time, space, or colour [Gurnee and Tegmark, 2024, Engels et al., 2025, Modell et al., 2025].

Why might manifold structure be present in data? In some situations, such as image analysis, an intuitive albeit heuristic explanation can be given in terms of the physical mechanism which generated the data (see e.g., Pless and Souvenir [2009] for a review of manifold estimation in this context). Figure 1 shows 24 grayscale images of a car, a subset of $n = 72$ images from [Geusebroek et al., 2005], taken from angles $0, 5, 10, \ldots, 355$ degrees around the circumference of a circle. Each image is of resolution 384×288 pixels and so can be represented as a vector of length $p = 110592$. However, at least intuitively, we can account for the variation across the collection of images using far fewer dimensions, in terms of the position of the camera in the three-dimensional space of the world around us. Figure 1 shows the result of using principal component analysis (PCA) to reduce dimension, upon which we make the following observations.

The first 20 principal components account for 91.5% of the total variance, suggesting that the data are concentrated somewhere in a low-dimensional linear subspace of $\mathbb{R}^{110592}$. The first three dimensions — the coordinates of the data with respect to the eigenvectors associated with the three largest eigenvalues — exhibit points around a loop which is somewhat irregular in shape but resembles the circle of camera positions, subject to deformation by bending and twisting. The points appear roughly equally spaced around the loop, like the camera positions which are equally spaced at intervals of 5 degrees around a circle.

Evidently reducing the dimension of these image data by PCA allows us to access some of the geometric structure of the data generating mechanism, but questions remain. We have chosen to plot the first three dimensions for ease of visualisation, is this a "good" choice? What might the other dimensions convey? What explains the precise shape of the loop and the spacing of the

points along it, relative to the underlying circle of camera positions?

In other situations, embedded topological and geometric structure may appear in different forms and have different interpretations. Figure 2 shows two approaches to visualising expression levels of $p = 5821$ genes measured across $n = 5000$ individual cells from adult planarians, a type of flatworm. In the field of single-cell transcriptomics — as set out in the 2018 Science paper [Plass et al., 2018] — such data offer the possibility of discovering the cell lineage tree of an entire animal: the aim is to find out if the data reflect the tree-structured process by which stem cells differentiate into a variety of distinct cell types. The data are preprocessed in the same way as the original paper [Plass et al., 2018], using the Python package Scanpy [Wolf et al., 2018].

The left plot in figure 2 shows the result of dimension reduction from 5821 to 2 using PCA. The right plot shows the result of first reducing from 5821 to 14 dimensions using PCA, followed by reduction to 2 dimensions using t-SNE [Van der Maaten and Hinton, 2008], a very popular nonlinear dimension reduction method which finds a lower dimensional representation of a data set by minimising a particular measure of distortion of pairwise distances. We used the default t-SNE parameter settings in `scikit-learn` [Pedregosa et al., 2011]. In both plots, the points are coloured by cell type, but neither PCA nor t-SNE have access to this information. Similarly to figure 1, it is evident from figure 2 that performing some form of dimension reduction allows us to access structure underlying the data, albeit in the form of discrete cell types rather than the geometry of camera positions. In figure 1, using only PCA to reduce dimension was enough to make this structure visible. However, in figure 2, using only PCA and reducing to 2 dimensions, distinct cell types are not clearly separated, whereas PCA down to 14 dimensions followed by t-SNE seems to be more effective. The t-SNE visualisation hints at the presence of tree structure underlying the data, with some areas having branch-like arms originating at the central point cloud, but other lineages lack clarity or seem to be disconnected.

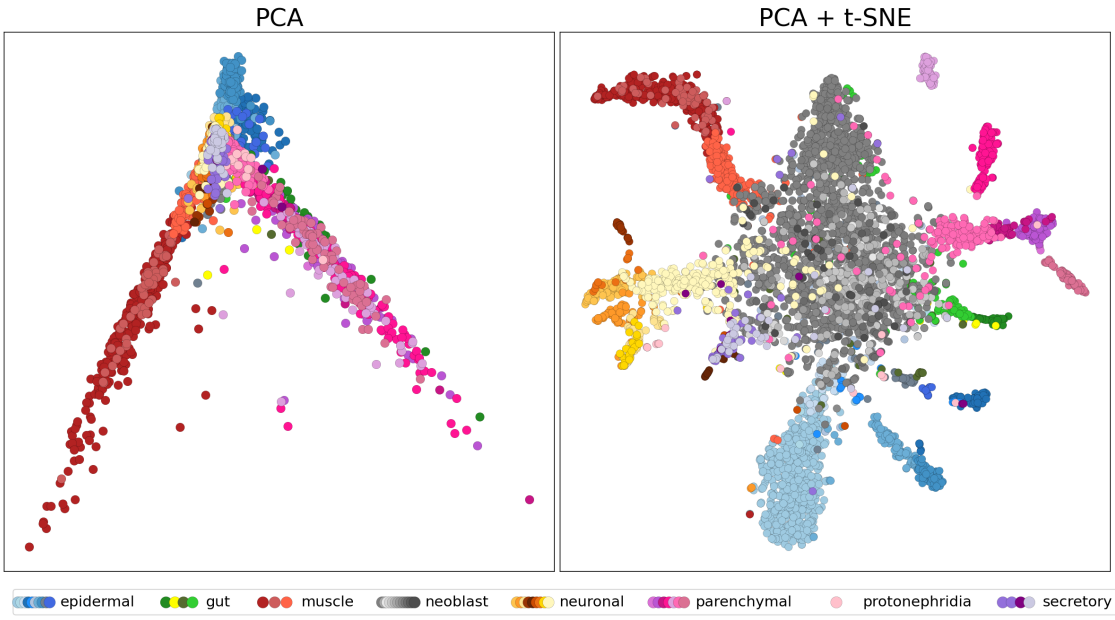

Figure 2: Planaria example. Left: first 2 dimensions of the PCA embedding. Right: representation of the data in 2 dimensions obtained by first reducing to 14 dimensions using PCA, then applying t-SNE.

These examples illustrate just some of the ways in which underlying structure can manifest itself in embedded topological and geometric patterns in data. Many other examples can be found: in genomics, where genotyping DNA sites has revealed striking geographic patterns [Novembre et al., 2008, Lao et al., 2008, Diaz-Papkovich et al., 2019]; neuroscience, where simultaneous recordings from Grid cells have been shown to exhibit toroidal structure seemingly independent of behavioural tasks [Gardner et al., 2022]; as well as manifold structure in data from wireless sensor networks [Patwari and Hero, 2004], visual speech recognition [Bregler and Omohundro,

1995], drug discovery [Reutlinger and Schneider, 2012], RNA sequencing [Moon et al., 2018], and human motion synthesis [Lee and Elgammal, 2006].

Main contributions. In this work we put forward a perspective that embedded topological and geometric structure in data can be explained as a general statistical phenomenon, without reference to physical properties or other domain-specific details of the data generating mechanism.

Our first main contribution is to propose a simple and generic statistical model which produces hidden, low-dimensional manifold structure in high-dimensional data, thus providing *a statistical justification for the manifold hypothesis*.

Our second main contribution is to describe how this hidden manifold relates to a true latent domain defined by the model, explaining, for example, why the points in the right panel of figure 1 are not in a perfect circle, as the camera positions are, but still form a loop. More precisely, we give mild conditions under which the relationship between the manifold and the latent domain is a homeomorphism (a topological equivalence), and stronger conditions under which it becomes an isometry (a metric equivalence).

Our third main contribution is to show that our model and theory enable a combination of *simple or well-known techniques* to be given a new, model-based interpretation and put to use in exploring hypotheses and uncovering information about the latent domain and broader data generating mechanism. Given data vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \in \mathbb{R}^p$, we rationalise the following workflow:

1. Dimension selection, using $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ to choose number of dimensions $\hat{r}$.
2. Linear dimension reduction of $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ by PCA, resulting in an $\hat{r}$-dimensional embedding, $\zeta_1, \ldots, \zeta_n$.
3. Spherical projection of the embedding, setting $\zeta_i^{\text{sp}} := \zeta_i / \|\zeta_i\|$, $i = 1, \ldots, n$
4. Nearest neighbour graph construction from $\zeta_1^{\text{sp}}, \ldots, \zeta_n^{\text{sp}}$.
5. Analysis and visualisation of the nearest neighbour graph, e.g., shortest paths, minimum spanning tree, topology.

For step 1., we introduce a new Wasserstein distance-based dimension selection method.

The remainder of this article is structured as follows. In section 2 we introduce the Latent Metric Space model, and the associated manifold $\mathcal{M}$, which arises as a consequence of correlation over a latent domain $\mathcal{Z}$. In section 3 we describe how this manifold structure hides in the data and how the manifold relates to $\mathcal{Z}$. We establish a representation formula (proposition 1) uncovering the perhaps surprising fact that, under the Latent Metric Space model, data are noisy, random projections of points in $\mathcal{M}$. Standard statistical concepts, such as stationarity, give rise to striking geometric relationships between $\mathcal{M}$ and $\mathcal{Z}$, such as isometry. In section 4 we develop theory and methodology supporting the workflow above, elucidating the benefits of applying PCA (theorem 1), proposing a new dimension selection method, and more. In section 5 we demonstrate the workflow on real data, revisiting the image and transcriptomics data from section 1, as well as a temperature time series example. The key new feature of these analyses is that we can explore manifold hypotheses grounded in a statistical model. In section 6 we draw together conclusions and discuss connections to the literature, including geometric representation of high-dimensional data, PCA in high dimensions, Gaussian process latent variable models, nonlinear dimension reduction, and exploratory data analysis.

2 The Latent Metric Space model

The Latent Metric Space (LMS) model is constructed from three independent sources of randomness.

Latent Variables. $Z_1, \ldots, Z_n$ are independent and identically distributed random elements of a metric space $(\mathcal{Z}, d_{\mathcal{Z}})$, that is, $\mathcal{Z}$ is a set, and $d_{\mathcal{Z}}(\cdot, \cdot)$ is a distance function on $\mathcal{Z}$. It is assumed that the metric space $(\mathcal{Z}, d_{\mathcal{Z}})$ is compact, and $Z_1, \ldots, Z_n$ are distributed according to a Borel probability measure μ supported on $\mathcal{Z}$.

Random Functions. $X_1, \ldots, X_p$ are random $\mathbb{R}$-valued functions, each with domain $\mathcal{Z}$. That is, for each $z \in \mathcal{Z}$ and $j = 1, \ldots, p$, $X_j(z)$ is an $\mathbb{R}$-valued random variable. It is not assumed that $X_1, \ldots, X_p$ are identically distributed, but it is assumed that $\mathbb{E}[|X_j(z)|^2] < \infty$, for all $j = 1, \ldots, p$ and $z \in \mathcal{Z}$.

Noise. $\mathbf{E} \in \mathbb{R}^{n \times p}$ is a matrix of random variables whose elements are each zero-mean and unit-variance. The columns of $\mathbf{E}$ are assumed independent and elements in distinct rows of $\mathbf{E}$ are assumed pairwise uncorrelated.

The data matrix $\mathbf{Y} \in \mathbb{R}^{n \times p}$ is defined by:

$$\mathbf{Y}_{ij} \coloneqq X_j(Z_i) + \sigma \mathbf{E}_{ij} \tag{1}$$

for some $\sigma \geq 0$. It will sometimes be convenient to think of data vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \in \mathbb{R}^p$ such that $[\mathbf{Y}_1 | \cdots | \mathbf{Y}_n]^\top \equiv \mathbf{Y}$, so $\mathbf{Y}_{ij}$ is the jth element of $\mathbf{Y}_i$. Similarly we shall write noise vectors $[\mathbf{E}_1 | \cdots | \mathbf{E}_n]^\top \equiv \mathbf{E}$.

We call:

$$f(z, z') \coloneqq \frac{1}{p} \sum_{j=1}^{p} \mathbb{E}[X_j(z) X_j(z')] \tag{2}$$

the *mean correlation kernel* associated with the LMS model. The following assumption is taken to hold throughout the paper without further mention.

A1. *For each $j = 1, \ldots, p$, $\mathbb{E}[X_j(z) X_j(z')]$ is a continuous function of $(z, z') \in \mathcal{Z} \times \mathcal{Z}$.*

Assumption **A1** implies $f(z, z')$ is continuous in z, z', and by a generalisation of Mercer's theorem [Mercer, 1909, Steinwart and Christmann, 2008, Thm 4.49] given in section A, when **A1** holds there exists a countable collection of non-negative real numbers $(\lambda_k^f)_{k \geq 1}$, $\lambda_1^f \geq \lambda_2^f \geq \cdots$, and a sequence of functions $(u_k^f)_{k \geq 1}$ which are orthonormal in $L_2(\mu)$ such that

$$f(z, z') = \sum_{k=1}^{\infty} \lambda_k^f u_k^f(z) u_k^f(z') = \langle \phi(z), \phi(z') \rangle_{\ell_2}, \tag{3}$$

where the series converges absolutely and uniformly in z, z'. The inner product in (3) is $\langle x, x' \rangle_{\ell_2} \coloneqq \sum_{k=1}^{\infty} x_k x_k'$, between infinitely long vectors $x = [x_1\, x_2 \cdots]^\top$ belonging to $\ell_2 \coloneqq \{x \in \mathbb{R}^{\mathbb{N}} : \|x\|_{\ell_2} < \infty\}$, where $\|x\|_{\ell_2} \coloneqq \left(\sum_{k=1}^{\infty} |x_k|^2\right)^{1/2} = \langle x, x \rangle_{\ell_2}^{1/2}$. The function $\phi : \mathcal{Z} \to \ell_2$ is the "feature map":

$$\phi(z) \coloneqq \left[(\lambda_1^f)^{1/2} u_1^f(z) \;\; (\lambda_2^f)^{1/2} u_2^f(z) \;\; \cdots \right]^\top, \tag{4}$$

with range we denote

$$\mathcal{M} \coloneqq \{\phi(z); z \in \mathcal{Z}\}. \tag{5}$$

The set $\mathcal{M}$ can be checked to be a subset of ℓ_2 using $\|\phi(z)\|_{\ell_2}^2 = f(z, z)$, and using the compactness of $\mathcal{Z}$ and the continuity assumption **A1** which imply $\sup_{z \in \mathcal{Z}} f(z, z) < \infty$.

We denote by r the rank of f, that is the largest $k \geq 1$ such that $\lambda_k^f > 0$, with $r \coloneqq \infty$ if $\lambda_k^f > 0$ for all $k \geq 1$. When $r < \infty$ we abuse notation slightly by writing

$$\phi(z) \coloneqq \left[(\lambda_1^f)^{1/2} u_1^f(z) \;\; (\lambda_2^f)^{1/2} u_2^f(z) \;\; \cdots \;\; (\lambda_r^f)^{1/2} u_r^f(z) \right]^\top. \tag{6}$$

We stress two points. First, the central purpose of the LMS model is to explain and describe manifold structure in data as a general statistical phenomenon. The breadth of this objective necessitates a flexible modelling paradigm and, except when considering examples, we do not make specific distributional or functional assumptions, such as Gaussianity. The assumptions in this paper involve more general concepts, such as continuity, smoothness or stationarity. Second, by contrast to much of the kernel methods literature, the kernel f is not chosen by the user here — it is unknown and derives from the ingredients of the LMS model.

3 Connecting statistical and geometric properties of the LMS model

In this section we explain how statistical properties of the LMS model allow us to connect the geometry of the data vectors, $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, which can be thought of as a point cloud in $\mathbb{R}^p$, to the structure of $\mathcal{M}$, and in turn the latent metric space $\mathcal{Z}$. This is important for two reasons. First, it shows how manifold structure in data emerges from elementary statistical properties of the LMS model, thus clarifying in what sense and why the Manifold Hypothesis holds. Second, it forms the basis for data analysis procedures we detail in section 4. We proceed in four main steps:

- Section 3.1 shows how inner-products between data vectors, say $\mathbf{Y}_i, \mathbf{Y}_j$, relate to inner products between $\phi(Z_i), \phi(Z_j)$. Since $\phi(Z_1), \ldots, \phi(Z_n)$ are i.i.d. and valued in $\mathcal{M}$, recall (5), this gives our first indication that the geometry of the point cloud $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ will reflect the shape of $\mathcal{M}$.
- Section 3.2 shows that under a simple distinguishability assumption, the feature map ϕ is a homeomorphism. Informally, this means we can think of $\mathcal{M}$ as being equivalent to $\mathcal{Z}$ up to some continuous, invertible distortion such as bending, twisting or stretching. Formally, we can say $\mathcal{M}$ is a topological manifold.
- Section 3.3 shows that when $\mathcal{Z}$ is a subset of Euclidean space, conditions closely related to weak stationarity of the random function X_j imply ϕ is an *isometry*. This means a very special form of geometric relationship holds between $\mathcal{M}$ and $\mathcal{Z}$, in which distances between points in $\mathcal{Z}$, say Z_i and Z_j, are faithfully represented by distances measured *along the manifold* $\mathcal{M}$ between $\phi(Z_i)$ and $\phi(Z_j)$, rather than by straight-line distances of the form $\|\phi(Z_i) - \phi(Z_j)\|_{\ell_2}$.
- Section 3.4 shows that if the kernel is sufficiently smooth, most of the structure of $\mathcal{M}$ is captured in a low-dimensional subspace. This hints towards the potential effectiveness of PCA (step 2 in the workflow) for manifold exploration.

Remarkably, we shall draw the conclusions in the second and third points above without any explicit knowledge of the eigenvalues and eigenfunctions which appear in the definition of ϕ, and which thus define $\mathcal{M}$.

3.1 Relating data inner products to feature map inner products

We have not made any assumptions about the functional form of $z \mapsto X_j(z)$, $j = 1, \ldots, p$, in the LMS model, other than **A1**. Nevertheless, the following proposition shows that a linear relationship holds between $\mathbf{Y}_i - \sigma\mathbf{E}_i$ and $\phi(Z_i)$.

Proposition 1. *Assume **A1**. Then under the LMS model with $r \in \{1, 2, \ldots, \} \cup \{\infty\}$, the matrix $\mathbf{W} \in \mathbb{R}^{p \times r}$ with elements*

$$\mathbf{W}_{jk} := \frac{1}{(p\lambda_k^f)^{1/2}} \int_{\mathcal{Z}} X_j(z) u_k^f(z) \mu(\mathrm{d}z) \tag{7}$$

satisfies

$$\mathbf{Y}_i \overset{m.s.}{=} p^{1/2}\mathbf{W}\phi(Z_i) + \sigma\mathbf{E}_i, \quad i = 1, \ldots, n, \qquad \mathbb{E}[\mathbf{W}^\top\mathbf{W}] = \mathbf{I}_r, \tag{8}$$

where $\mathbf{I}_r$ is the identity matrix with r rows and columns.

The qualification "$\overset{m.s.}{=}$" in (8) indicates that the infinite summations constituting the matrix-vector product $\mathbf{W}\phi(Z_i)$ in the case $r = \infty$ converge in the mean-square sense. The proof of proposition 1, in appendix C, entails a generalised form of Karhunen-Loève expansion of $X_1, \ldots, X_p$.

The identities in (8) can be interpreted as meaning that $p^{-1/2}\mathbf{Y}_i$ is a noisy, random projection of $\phi(Z_i)$. Indeed we can use (8) together with the defining properties of the LMS model in section 2 to describe the behaviour of the inner-product between $\mathbf{Y}_i, \mathbf{Y}_j \in \mathbb{R}^p$ when the randomness in $X_1, \ldots, X_p$ and $\mathbf{E}$ is averaged out:

$$\begin{aligned}\frac{1}{p}\mathbb{E}[\langle \mathbf{Y}_i, \mathbf{Y}_j \rangle | Z_i, Z_j] &= \langle \phi(Z_i), \mathbb{E}[\mathbf{W}^\top\mathbf{W}]\phi(Z_j) \rangle_{\ell_2} + 0 + 0 + \sigma^2 \frac{1}{p}\mathbb{E}[\langle \mathbf{E}_i, \mathbf{E}_j \rangle] \\ &= \langle \phi(Z_i), \phi(Z_j) \rangle_{\ell_2} + \sigma^2 \mathbf{I}[i = j].\end{aligned} \tag{9}$$

The quantity $p^{-1}\langle \mathbf{Y}_i, \mathbf{Y}_j \rangle$ is an arithmetic mean of p random variables. If, conditionally on Z_i, Z_j, the summands in $p^{-1}\langle \mathbf{Y}_i, \mathbf{Y}_j \rangle$ are weakly dependent and have moments bounded uniformly in p, then by a law of large numbers argument $p^{-1}\langle \mathbf{Y}_i, \mathbf{Y}_j \rangle$ will be close to its conditional expectation (9) with high probability when p is large (see proposition 6 in appendix C for details). Moreover, write $\mathbf{W} \equiv [\mathbf{W}_1|\cdots|\mathbf{W}_p]^\top$ and note from (7) that the only randomness in $\mathbf{W}_j$ arises from X_j. So if $X_1, \ldots, X_p$ were assumed weakly dependent, $\mathbf{W}_1, \ldots, \mathbf{W}_p$ would be too. Then, again assuming moments bounded uniformly in p, by a law of large numbers argument the sum $\sum_{j=1}^p \mathbf{W}_j \mathbf{W}_j^\top$ will be close to its expectation with high probability when p is large, i.e.,

$$\mathbf{W}^\top \mathbf{W} = \sum_{j=1}^{p} \mathbf{W}_j \mathbf{W}_j^\top \approx \mathbb{E}[\mathbf{W}^\top \mathbf{W}] = \mathbf{I}_r.$$

We therefore conclude that, subject to suitable weak dependence and moment conditions,

$$\left| p^{-1}\langle \mathbf{Y}_i, \mathbf{Y}_j \rangle - \langle \phi(Z_i), \phi(Z_j) \rangle_{\ell_2} - \sigma^2 \mathbf{I}[i=j] \right| \to 0, \quad \text{as} \quad p \to \infty, \tag{10}$$

in probability. In this sense the geometry of the collection of high-dimensional data vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ reflects that of $\phi(Z_1), \ldots, \phi(Z_n)$, subject to some distortion depending on the noise level σ. Moreover, if (10) holds, then $|p^{-1}\|\mathbf{Y}_i - \mathbf{Y}_j\|^2 - \|\phi(Z_i) - \phi(Z_j)\|_{\ell_2}^2 - 2\sigma^2| \to 0$ as $p \to \infty$.

In section 4.1, we shall complement the above reasoning with theorem 1 which shows that when the noise level σ is fixed, and $n \to \infty$ and $p/n \to \infty$ simultaneously, using PCA to reduce dimension of $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ allows $\phi(Z_1), \ldots, \phi(Z_n)$ to be recovered, up to an orthogonal transformation.

3.2 Relating distinguishability of latent variables to homeomorphism

A *homeomorphism* between two metric spaces is a mapping which is continuous, bijective and has a continuous inverse. If such a mapping exists the two metric spaces are said to be homeomorphic, or *topologically equivalent*. To develop some intuition, one can think about the case in which the metric spaces in question are subsets of the three-dimensional Euclidean world around us. In this situation mappings which qualify as homeomorphisms include transformations of shape by bending, twisting, stretching and folding, but not cutting, puncturing or joining [Bing, 1960]. Topological equivalence implies the two metric spaces in question must exhibit the same number of connected components, the same number of 1-dimensional loops and more generally the same number of k-dimensional "holes" as each other. Detecting such features using data is the purpose of persistent homology methods within the field of Topological Data Analysis [Carlsson, 2009, Chazal and Michel, 2021]. But there is more to a topological structure than its homology; for example, in the transcriptomics application (introduction and section 5.2), the hypothesised underlying structure has interesting, 'tree-like', topology but no interesting homology.

We shall now see that, with only a little more structure added to the LMS model, ϕ is homeomorphism between $\mathcal{Z}$ and $\mathcal{M}$, where the distance on $\mathcal{M}$ is $\|\cdot - \cdot\|_{\ell_2}$. The first requirement, continuity of ϕ, means that $d_{\mathcal{Z}}(z, z') \to 0$ implies $\|\phi(z) - \phi(z')\|_{\ell_2} \to 0$. This holds due to the identities:

$$\begin{aligned} \|\phi(z) - \phi(z')\|_{\ell_2}^2 &= \|\phi(z)\|_{\ell_2}^2 + \|\phi(z')\|_{\ell_2}^2 - 2\langle \phi(z), \phi(z') \rangle_{\ell_2} \\ &= f(z,z) + f(z',z') - 2f(z,z'), \end{aligned}$$

combined with continuity of f under **A1**. By its definition, $\phi : \mathcal{Z} \to \mathcal{M}$ is automatically surjective, and if ϕ is one-to-one, its inverse is automatically continuous due to a general result in the theory of metric spaces [Sutherland, 2009, Prop. 13.26] concerning the inverse of a continuous mapping with compact domain. The question of whether or not ϕ is a homeomorphism thus reduces to whether or not it is one-to-one. Consider the following assumption.

A2. *For each $z, z' \in \mathcal{Z}$ such that $z \neq z'$, $\sum_{j=1}^p \mathbb{E}[|X_j(z) - X_j(z')|^2] > 0$.*

Proposition 2. *Assume **A1**. Then $\phi : \mathcal{Z} \to \mathcal{M}$ is a homeomorphism if and only if **A2** holds.*

Assumption **A2** can be interpreted as a "distinguishability" condition, requiring that points in $\mathcal{Z}$ can be distinguished from each other in terms of the random functions X_j. This distinguishability

can also be stated in terms of the kernel f; in the proof of proposition 2 in appendix C, we show that **A2** is equivalent to the condition: for each $z, z' \in \mathcal{Z}$ such that $z \neq z'$, there exists $\xi \in \mathcal{Z}$ such that $f(z, \xi) \neq f(z', \xi)$.

The term *topological manifold* conventionally means some set such that each point in that set has a neighbourhood which is homeomorphic to some subset of Euclidean space. We note that the relationship between $\mathcal{M}$ and $\mathcal{Z}$ is of a similar nature, except that $\mathcal{M}$ is globally rather than only locally homeomorphic to $\mathcal{Z}$, and the metric space $\mathcal{Z}$ need not be Euclidean. Putting these differences aside, we shall simply call $\mathcal{M}$ a manifold from now on.

When $\mathcal{M}$ and $\mathcal{Z}$ are homeomorphic, they must have the same *covering dimension*—see [Pears, 1975, Ch.3] for background—this is an abstract topological notion dimension, which generalises the usual notion of dimension of Euclidean space. In this sense, we can say that when $\mathcal{Z}$ is low-dimensional, $\mathcal{M}$ is low-dimensional too.

3.3 Relating stationarity to isometry

Weak stationarity of any one of the random functions X_j in the LMS model would mean that:

- $\mathbb{E}[X_j(z)]$ is constant in z, and
- $\mathbb{E}[(X_j(z) - \mathbb{E}[X_j(z)])(X_j(z') - \mathbb{E}[X_j(z')])]$ is a function only of distance between z and z'.

If all the random functions $X_1, \ldots, X_p$ were to have this property, it would follow from the definition of f in (2) that $f(z, z')$ must also be a function only of distance between z and z'. We shall now see that this leads to an *isometric* relationship between $\mathcal{M}$ and $\mathcal{Z}$. To define isometry, it's convenient to work in the following setting:

A3. *$\mathcal{Z}$ is a compact subset of $\mathbb{R}^d$, and there exists a continuous path in $\mathcal{Z}$ of finite length between any two points in $\mathcal{Z}$.*

The precise mathematical definition of a path and its length are given in appendix C.1. In the setting of **A3**, we denote by $d^{\text{geo}}_{\mathcal{Z}}(z, z')$ the shortest path length, or geodesic distance, in $\mathcal{Z}$. This is the infimum of the lengths of all paths in $\mathcal{Z}$ with end-points z, z' (see appendix C.1 for details). If $\mathcal{Z}$ is convex, the shortest path between two points is a straight line and $d^{\text{geo}}_{\mathcal{Z}}(z, z') = \|z - z'\|_{\mathbb{R}^d}$. For $x, x' \in \mathcal{M}$, the shortest path length, or geodesic distance, in $\mathcal{M}$ is denoted $d^{\text{geo}}_{\mathcal{M}}(x, x')$ and defined analogously to $d^{\text{geo}}_{\mathcal{Z}}(z, z')$. Even when $\mathcal{Z}$ is convex, in general $\mathcal{M}$ is not convex and $d^{\text{geo}}_{\mathcal{M}}(x, x')$ is not equal to the straight-line distance $\|x - x'\|_{\ell_2}$.

We shall say *isometry* holds between $\mathcal{Z}$ and $\mathcal{M}$ if

$$d^{\text{geo}}_{\mathcal{M}}(\phi(z), \phi(z')) = d^{\text{geo}}_{\mathcal{Z}}(z, z'), \qquad \forall z, z' \in \mathcal{Z}. \tag{11}$$

Compared to homeomorphism, this isometry condition imposes more of a constraint on the relationship between $\mathcal{Z}$ and $\mathcal{M}$. One can interpret isometry as allowing ϕ to transform $\mathcal{Z}$ into $\mathcal{M}$ by bending, but not by stretching or compressing, since that would violate the equality of shortest path lengths.

The following proposition shows that isometry holds up to a scaling constant when, for z, z' close to each other, $f(z, z')$ depends only on the Euclidean distance between z and z'. In contrast, weak-sense stationarity involves the more stringent requirement that such dependence holds *for all* z, z'. Define $\mathcal{D} := \{(z, z); z \in \mathcal{Z}\} \subset \mathcal{Z} \times \mathcal{Z}$.

Proposition 3. *Assume **A2** and **A3**. If $f(z, z') = g(\|z - z'\|^2_{\mathbb{R}^d})$ for all z, z' in an open neighbourhood of $\mathcal{D}$ where g is twice continuously differentiable and $g'(0) < 0$, then*

$$d^{\text{geo}}_{\mathcal{M}}(\phi(z), \phi(z')) = \sqrt{-2g'(0)}\, d^{\text{geo}}_{\mathcal{Z}}(z, z'). \tag{12}$$

The following proposition complements proposition 3 by addressing the special case in which $\mathcal{Z}$ is a sphere.

Proposition 4. *Assume **A2**. If $\mathcal{Z} = \{z \in \mathbb{R}^d : \|z\|_{\mathbb{R}^d} = 1\}$ and $f(z, z') = g(\langle z, z' \rangle_{\mathbb{R}^d})$ for all z, z' in an open neighbourhood of $\mathcal{D}$ where g is twice continuously differentiable and $g'(1) > 0$, then*

$$d^{\text{geo}}_{\mathcal{M}}(\phi(z), \phi(z')) = \sqrt{g'(1)}\, d^{\text{geo}}_{\mathcal{Z}}(z, z'). \tag{13}$$

The proofs of propositions 3 and 4 are at the end of appendix C.1.

3.4 Relating smoothness to concentration within a low-dimensional subspace

When the latent domain $\mathcal{Z}$ is a subset of $\mathbb{R}^d$, we will say that f is smooth if it can be expressed as the restriction of a smooth function on $\mathbb{R}^d \times \mathbb{R}^d$ to $\mathcal{Z} \times \mathcal{Z}$. How smooth f is affects how much of the manifold $\mathcal{M}$ we can capture using only the first few coordinates. For some $s < r$, consider the truncated map

$$\phi_s(z) := \left[(\lambda_1^f)^{1/2} u_1^f(z) \quad \cdots \quad (\lambda_s^f)^{1/2} u_s^f(z) \quad 0 \quad \cdots \right]^\top.$$

The eigenvalues give us a measure of how well $\mathcal{M}_s := \phi_s(\mathcal{Z})$ approximates $\mathcal{M}$ through the mean square error

$$\mathbb{E}[\|\phi(Z_i) - \phi_s(Z_i)\|_{\ell_2}^2] = \sum_{k>s} \lambda_k^f \mathbb{E}\left[|u_k^f(Z_i)|^2\right] = \sum_{k>s} \lambda_k^f. \tag{14}$$

The rate of decay of the eigenvalues is known to be related to the smoothness of the kernel [Takhanov, 2023] and so, under smoothness assumptions, (14) tells us that for s suitably large the first s coordinates of ϕ can provide a good approximation to $\mathcal{M}$, even if $r = \infty$. When $s \leq p$, such smoothness also implies each vector $\mathbf{Y}_i$ will be concentrated within the (at most) s-dimensional subspace of $\mathbb{R}^p$ spanned by the first s columns of $\mathbf{W}$. Indeed recalling the identity $\mathbf{Y}_i \overset{m.s.}{=} p^{1/2}\mathbf{W}\phi(Z_i) + \sigma\mathbf{E}_i$ from proposition 1 we have:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{Y}_i - p^{1/2}\mathbf{W}\phi_s(Z_i)\|^2\right] &= \mathbb{E}\left[\|p^{1/2}\mathbf{W}\phi(Z_i) - p^{1/2}\mathbf{W}\phi_s(Z_i) + \sigma\mathbf{E}_i\|^2\right] \\ &= p\mathbb{E}[\|\phi(Z_i) - \phi_s(Z_i)\|_{\ell_2}^2] + \sigma^2\mathbb{E}[\|\mathbf{E}_i\|^2] = p\sum_{k>s}\lambda_k^f + p\sigma^2,\end{aligned}$$

where the independence of $\mathbf{W}$, $\phi(Z_i)$, $\mathbf{E}_i$, the second equality in (8) and the properties $\mathbb{E}[\mathbf{E}_{ij}] = 0$, $\mathbb{E}[\mathbf{E}_{ij}^2] = 1$ have been used. An LMS model with smooth kernel can therefore produce a data matrix $\mathbf{Y}$ which is 'approximately low-rank', a common feature of real data [Udell and Townsend, 2019], hinting that PCA may be a useful tool to help recover $\phi(Z_1), \ldots, \phi(Z_n)$.

3.5 A visual example

To illustrate some of the concepts from section 3, we consider a case in which $\mathcal{Z}$ is a torus embedded in $\mathbb{R}^3$, satisfying **A3** with $d = 3$. We take μ to be the uniform distribution on the torus, and $Z_1, \ldots, Z_{4000}$ simulated from μ are shown in figure 3. The colouring of the points in this figure emphasises that the torus is the Cartesian product of two circles, and the locations on the torus can be parameterised in terms of angles around these two circles.

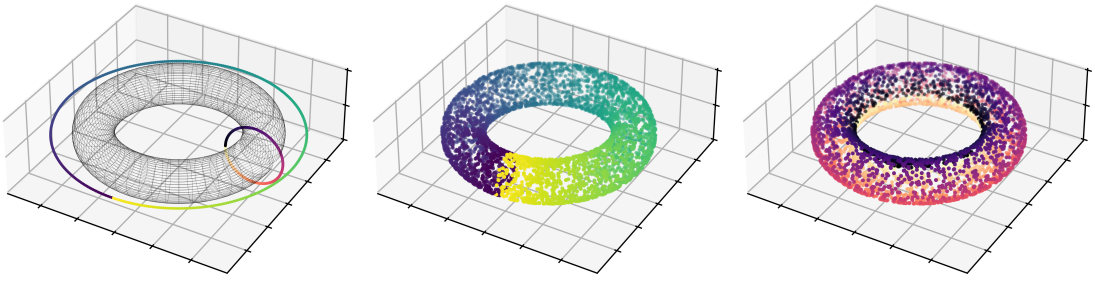

Figure 3: Torus example. Left: grey wireframe of $\mathcal{Z}$, a torus, with colour bars indicating coordinates with respect to two circles. Both the middle and right plots show the same $n = 4000$ points, $Z_1, \ldots, Z_{4000}$, which are sampled uniformly on the torus, coloured by their coordinates with respect to each of the two circles.

We assume $X_1 \ldots, X_p$ are i.i.d., zero-mean Gaussian processes with common covariance function $\exp(-\|z - z'\|_{\mathbb{R}^3}^2) = f(z, z')$, which satisfies **A1**. Figure 4 shows numerical approximations to

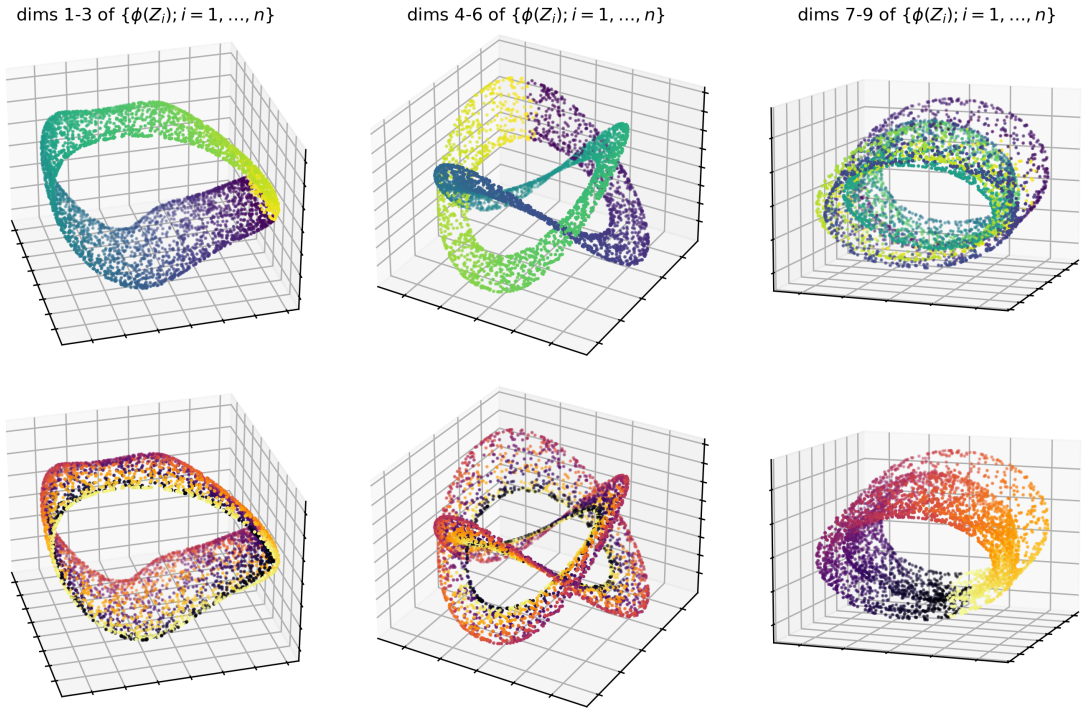

Figure 4: Torus example. Both the top and bottom rows show the first 9 dimensions of $\phi(Z_i)$, $i = 1, \ldots, 4000$. In each row, points are coloured according to the coordinates of the underlying points $Z_1, \ldots, Z_n$ with respect to the two circles shown in figure 3. Numerical scales are omitted to de-clutter the plots.

the first 1-3, 4-6 and 7-9 dimensions of $\phi(Z_i)$, $i = 1, \ldots, 4000$ (these approximations were obtained using PCA, the details of which are given later in section 4.1). The only difference between the two rows of plots in figure 4 is the colouring of the points; the colouring in the top row is the colouring of the corresponding points in the middle plot in figure 3, similarly the colouring in the bottom row matches that in the plot on the right of figure 3.

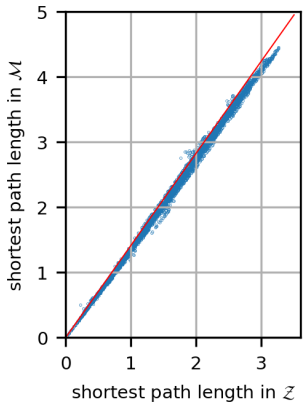

Figure 5: Torus example. Blue: numerical shortest path lengths between points in $\mathcal{M}$ vs. between the corresponding points in $\mathcal{Z}$. Red: theoretical scaling relationship $\sqrt{2}$.

It is clear from figure 4 that the global shape of $\mathcal{M}$, when viewed three dimensions at a time, is qualitatively different to the global shape of $\mathcal{Z}$. However, assumption **A2** holds, so by lemma 2, we know $\mathcal{M}$ is topologically equivalent to $\mathcal{Z}$, and by proposition 3, ϕ isometry holds, up to a scaling factor of $\sqrt{-g'(0)} = \sqrt{2}$ for the f in question. This tells us that shortest path lengths in $\mathcal{M}$

are equal to the corresponding path lengths $\mathcal{Z}$, up to a factor of $\sqrt{2}$. Figure 5 shows comparison of these shortest path lengths, computed numerically from the points in figures 4 and 3 using a nearest neighbour graph as detailed in section 4.4. We see a close approximation to the theoretical scaling factor of $\sqrt{2}$, shown by the red line.

Overall this example illustrates that if we are interested in discovering the topological or geometric structure of $\mathcal{Z}$ based on observations of $\mathcal{M}$, we should not pay attention to the global shape of $\mathcal{M}$ that we perceive visually, because that depends on both $\mathcal{Z}$ and ϕ. However, when homeomorphism holds, we can in principle recover the abstract topological structure of $\mathcal{Z}$ and its homological features such as number of connected components, number of holes, etc., from $\mathcal{M}$. Moreover, when isometry holds, at least up to a constant scaling factor, we can gain insight into the geometry of $\mathcal{Z}$ from shortest paths in $\mathcal{M}$.

4 Methodology

In this section, properties of the LMS model are used to explain and justify the workflow outlined in section 1. Discussion of the step 1. is postponed until after discussion of step 2.

4.1 Linear dimension reduction by PCA

Given data $\mathbf{Y} \in \mathbb{R}^{n\times p}$ and $s \leq \min\{p, n\}$, let the columns of $\mathbf{V}_{\mathbf{Y}} \in \mathbb{R}^{p\times s}$ be orthonormal eigenvectors associated with the s largest eigenvalues of $\mathbf{Y}^\top\mathbf{Y} \in \mathbb{R}^{p\times p}$. The *dimension-s PCA embedding* is the collection of vectors $\zeta_1, \ldots, \zeta_n$, defined by:

$$[\zeta_1|\cdots|\zeta_n]^\top \coloneqq \mathbf{Y}\mathbf{V}_{\mathbf{Y}}, \tag{15}$$

so for each $\zeta_i = \mathbf{V}_{\mathbf{Y}}^\top\mathbf{Y}_i$ is a vector in $\mathbb{R}^s$. These quantities are sometimes called principal component scores [Lee et al., 2010, Shen et al., 2012, Hellton and Thoresen, 2017]. When performing PCA in practice, one often centers the data about their sample mean. For simplicity of presentation we do not consider such centering here, although we do not require population centering, that is we do not assume $\mathbb{E}[\mathbf{Y}_i] = 0$.

The following assumptions about the LMS model are introduced to enable theoretical analysis of the PCA embedding.

A4. *The random functions $X_1, X_2, \ldots$ are independent.*

A5. $\sup_{j\geq 1}\sup_{z\in\mathcal{Z}}\mathbb{E}[|X_j(z)|^4] < \infty$ *and* $\sup_{j\geq 1}\sup_{i\geq 1}\mathbb{E}[|\mathbf{E}_{ij}|^4] < \infty$.

A6. *For each $p \geq 1$, the rank r of the mean correlation kernel f defined in (2) is finite, and r and $1/\lambda_r^f$ are bounded as $p \to \infty$.*

Theorem 1. *Assume **A1**, **A4**-**A6** and let r be as therein. Let $\zeta_1, \ldots, \zeta_n$ be the dimension-r PCA embedding of $\mathbf{Y} \in \mathbb{R}^{n\times p}$ under the LMS model. Then there exists a random orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{r\times r}$ depending on n and p such that*

$$\max_{i=1,\ldots,n}\left\|p^{-1/2}\mathbf{Q}\zeta_i - \phi(Z_i)\right\|_2 \in O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}} + \sqrt{\frac{n}{p}}\right) \tag{16}$$

as $n \to \infty$ and $p/n \to \infty$ simultaneously, where $\|\cdot\|_2$ is the Euclidean norm.

Theorem 1 is a corollary to a more detailed non-asymptotic concentration result for the PCA embedding, theorem 4, stated and proved in appendix D.

Interpretation of theorem 1

Theorem 1 implies that for any $\epsilon > 0$, the probability that $\max_{i=1,\ldots,n}\left\|p^{-1/2}\mathbf{Q}\zeta_i - \phi(Z_i)\right\|_2 > \epsilon$ converges to zero when n and p/n grow simultaneously. In that sense $\phi(Z_1), \ldots, \phi(Z_n)$ can be recovered from $p^{-1/2}\zeta_1, \ldots, p^{-1/2}\zeta_n$, up to an orthogonal transformation, i.e., a transformation which preserves distances and inner-products. We see that computing the PCA embedding

achieves a form of de-noising and signal extraction: each ζ_i depends on all three sources of randomness in the LMS model, but $\phi(Z_i)$ clearly depends only on the random latent variable Z_i.

The result has positive implications for different forms of unsupervised learning, such as clustering, topological data analysis or manifold learning in the regime $n, p/n \to \infty$. Viewed as sets, the point clouds $\{p^{-1/2}\zeta_i\}_{i=1,\dots,n}$ and $\{\phi(Z_i)\}_{i=1,\dots,n}$ converge to each other in Hausdorff distance, up to $\mathbf{Q}$, implying convergence of topological summaries such as persistence diagrams [Wasserman, 2018], and so on. Broadly speaking, we can consider $p^{-1/2}\zeta_1, \dots, p^{-1/2}\zeta_n$ as proxies for $\phi(Z_1), \dots, \phi(Z_n)$ when estimating features of $\mathcal{M}$. The LMS model then gives us a way to translate such estimates into statements about the latent domain $\mathcal{Z}$ (see section 3).

Discussion of assumptions A4-A6

In the proof of theorem 4 and hence theorem 1, the independence assumption **A4** and the moment assumption **A9** are used when analysing $p^{-1}\mathbf{Y}\mathbf{Y}^\top$ via matrix a polynomial moment concentration inequality of Paulin et al. [2016]. The moment assumption **A9** is not particularly restrictive. From a modelling point of view relaxing **A4** to some form of weak dependence or mixing condition would be desirable, but the authors do not know of any suitable polynomial moment matrix concentration inequalities which are applicable in that situation.

Concerning assumption **A6**, that f has finite rank: recall from section 3.4 that the eigenvalues tend to tail off quickly when f is smooth, in which case assumption **A6** might be taken to hold approximately. Moreover, if f is polynomial [Rubin-Delanchy, 2020] or piecewise polynomial, or if $\mathcal{Z}$ consists of finitely many points, then f has strictly finite rank. As a result, assumption **A6** is mild enough to include any function f which is obtainable from standard numerical or function approximation schemes (e.g. Taylor expansion, polynomial splines, etc).

The condition that r is bounded as $p \to \infty$ in **A6** can be understood as constraining the functional complexity of f as p grows. The condition that $1/\lambda_r^f$ is bounded as $p \to \infty$ means that the additive noise whose scale is specified by the constant σ cannot overwhelm the "signal" in the LMS model. If $X_1, X_2, \dots$ are identically distributed then f, and hence r and λ_r^f, are automatically constant in p. The assumption in theorem 1 that the dimension of the PCA embedding is equal to the finite kernel rank r is an idealisation, although a very common type of assumption in uniform consistency results for spectral embedding, e.g. Lyzinski et al. [2016].

4.2 Choosing the PCA dimension

Our model and theory motivate a new method for choosing the embedding dimension, $\hat{r}$. Before proceeding, we should make clear that the hat notation in $\hat{r}$ is meant loosely: we seek a choice which achieves a good bias/variance trade-off in practice, and this may or may not coincide with the true rank of the kernel, r. Moreover, we do not claim that there is a 'best' choice: different tasks benefit from different choices. In particular, if using PCA for prediction purposes we simply recommend cross-validation, as is common practice. For more exploratory analyses, as conducted here, we propose the following approach instead.

Assuming n is even split the data $\mathbf{Y}$ into two, $\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)} \in \mathbb{R}^{n/2 \times p}$, and for each candidate dimension ρ, take the orthogonal projection of the rows of $\mathbf{Y}^{(1)}$ onto the ρ principal eigenvectors of ${\mathbf{Y}^{(1)}}^\top \mathbf{Y}^{(1)}$ — the resulting $n/2$ vectors are p-dimensional, just constrained to a ρ-dimensional subspace. Next, measure how much this projection step has brought the first half closer to the second, using Wasserstein distance. Select $\hat{r}$ to be the ρ achieving the smallest distance. The procedure is described precisely in algorithm 1. The algorithm accommodates the possibility that there may be some maximal value $\rho_{\max} < \min(n, p)$ that one is willing to consider, e.g. for computational reasons.

When calculating the Wasserstein distance between two sets of n samples in $\mathbb{R}^p$, the overall computational complexity can be up to $O(n^2(p + n \log n))$. One way to reduce the burden of computing Wasserstein distances over very large dimensional point sets, is to reduce the data to $\rho_{\max} \ll p$ dimensions using PCA. Reducing the complexity further using approximations is well studied [Cuturi, 2013, Arjovsky et al., 2017, Bonneel et al., 2015, Arjovsky et al., 2017]. A variety of methods are available in the Python packages `POT` [Flamary et al., 2021] and `Geomloss` [Feydy et al., 2019].

To understand how $\hat{r}$ might relate to r, let us make a few simplifying assumptions. Suppose $r < \infty$, so that (with exact equality)

$$\mathbf{Y}_i = p^{1/2}\mathbf{W}\phi(Z_i) + \sigma\mathbf{E}_i,$$

and that the second Wasserstein distance is used, that is $d_\rho = \mathcal{W}_2(\mathbf{Y}^{(1)}\mathbf{\Pi}_\rho, \mathbf{Y}^{(2)})$ where

$$\mathcal{W}_2^2(\mathbf{A}, \mathbf{B}) \coloneqq \min_\pi \frac{1}{m}\sum \|\mathbf{A}_i - \mathbf{B}_{\pi(i)}\|_2^2, \quad \mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times d},$$

where $\mathbf{A}_i$ and $\mathbf{B}_i$ are the ith rows of $\mathbf{A}$ and $\mathbf{B}$, and where the minimum is over all permutations of the integers $1, \ldots, m$.

The second Wasserstein distance is particularly amenable to mathematical analysis because of the following property, which can be checked by direct calculation. If there exist $\hat{\mathbf{A}}_1, \ldots, \hat{\mathbf{A}}_m$ such that for all i, j we have $\langle \mathbf{A}_i - \hat{\mathbf{A}}_i, \mathbf{B}_j\rangle = 0$, then

$$\mathcal{W}_2^2(\mathbf{A}, \mathbf{B}) = \frac{1}{m}\|\mathbf{A} - \hat{\mathbf{A}}\|_{\mathrm{F}}^2 + \mathcal{W}_2^2(\hat{\mathbf{A}}, \mathbf{B}). \tag{17}$$

To see why algorithm 1 might reject overly values of ρ, suppose $\rho > r$ and consider the projection errors $\boldsymbol{\mathcal{E}}^{(1)} = \mathbf{Y}^{(1)}(\mathbf{\Pi}_\rho - \mathbf{\Pi}_r)$. With $\boldsymbol{\mathcal{E}}_i^{(1)}$ denoting the ith row of $\boldsymbol{\mathcal{E}}^{(1)}$ and the superscript (k) indicating random objects associated with $\mathbf{Y}^{(k)}$, suppose

$$\frac{1}{p}\langle\, \boldsymbol{\mathcal{E}}_i^{(1)}, \mathbf{E}_j^{(2)}\rangle \approx 0, \quad \text{and} \quad \frac{1}{p}\langle\, \boldsymbol{\mathcal{E}}_i^{(1)}, \sqrt{p}\mathbf{W}\phi(Z_j^{(2)})\rangle \approx 0, \quad i, j = 1, \ldots, n/2. \tag{18}$$

Then, using (17),

$$\frac{1}{p}d_\rho^2 \approx \frac{2}{np}\|\boldsymbol{\mathcal{E}}^{(1)}\|_{\mathrm{F}}^2 + \frac{1}{p}d_r^2,$$

where $\|\boldsymbol{\mathcal{E}}^{(1)}\|_{\mathrm{F}}^2$ is non-decreasing in $\rho - r$, and it follows that we should expect $d_\rho > d_r$ when ρ is large relative to r.

Why should the approximations in (18) hold? The first is the product of sample-splitting. If the $\mathbf{E}_i$ are statistically independent, then $\mathbf{E}^{(2)}$ is statistically independent of $\mathbf{V}_\rho^{(1)}$. Combined with the fact that the elements of $\mathbf{E}_i^{(2)}$ are mean-zero and unit-variance, we therefore expect the $p^{-1}\mathbf{E}_i^{(2)}$ to be approximately orthogonal to the subspace spanned by the columns of $\mathbf{V}_\rho^{(1)}$, *when p is large relative to n*. The second approximation seems to be reasonable as long as n and p are large, and we have confirmed this by simulation.

Now consider the case $\rho < r$. We have

$$d_\rho^2 = \mathcal{W}_2^2(\mathbf{Y}^{(1)}\mathbf{\Pi}_\rho, \mathbf{Y}^{(2)}) = \frac{2}{n}\|\mathbf{Y}^{(2)}(\mathbf{\Pi}_\rho - \mathbf{I}_p)\|_{\mathrm{F}}^2 + \mathcal{W}_2^2(\mathbf{Y}^{(1)}\mathbf{\Pi}_\rho, \mathbf{Y}^{(2)}\mathbf{\Pi}_\rho).$$

By the Eckart-Young theorem,

$$\|\mathbf{Y}^{(2)}(\mathbf{\Pi}_\rho - \mathbf{I}_p)\|_{\mathrm{F}}^2 \geq \sum_{k>\rho} \lambda_k(\mathbf{Y}^{(2)\top}\mathbf{Y}^{(2)}) \tag{19}$$

where $\lambda_k(\mathbf{Y}^{(2)\top}\mathbf{Y}^{(2)})$ is the kth largest eigenvalue of $\mathbf{Y}^{(2)\top}\mathbf{Y}^{(2)}$. The r.h.s. of (19) is non-decreasing in $r - \rho$. To see that the term $\mathcal{W}_2^2(\mathbf{Y}^{(1)}\mathbf{\Pi}_\rho, \mathbf{Y}^{(2)}\mathbf{\Pi}_\rho)$ converges to zero as $p/n, n \to \infty$, it is convenient to consider the case $\sigma = 0$. In this situation,

$$\frac{1}{p}\mathcal{W}_2^2(\mathbf{Y}^{(1)}\mathbf{\Pi}_\rho, \mathbf{Y}^{(2)}\mathbf{\Pi}_\rho) \leq \frac{1}{p}\mathcal{W}_2^2(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}) = \frac{1}{p}\mathcal{W}_2^2(\mathbf{\Phi}^{(1)}\mathbf{W}^\top, \mathbf{\Phi}^{(2)}\mathbf{W}^\top)$$

where $\mathbf{\Phi}^{(k)} = [\phi(Z_1^{(k)})|\cdots|\phi(Z_{n/2}^{(k)})]^\top \in \mathbb{R}^{n/2\times r}$.

Appealing to the same arguments as in section 3.1, as $p/n \to \infty$, $\frac{1}{p}\mathcal{W}_2^2(\mathbf{\Phi}^{(1)}\mathbf{W}^\top, \mathbf{\Phi}^{(2)}\mathbf{W}^\top)$ is concentrated about $\mathcal{W}_2^2(\mathbf{\Phi}^{(1)}, \mathbf{\Phi}^{(2)})$, and the latter converges to zero as $n \to \infty$ because the rows of $\mathbf{\Phi}^{(1)}$ and $\mathbf{\Phi}^{(2)}$ are i.i.d. random vectors in $\mathbb{R}^r$. It follows that we should expect $d_\rho > d_r$ when ρ is small relative to r, and so overall that d_ρ will have a minimum near $\rho = r$.

A general rule we could draw from these arguments, and which we see in practice, is that to recommend substantial dimension reduction the algorithm wants to see a large p relative to n, and noise. Conversely, if the noise level is low or if n is large relative to p, then d_ρ may keep decreasing with ρ, which we tend to interpret as contraindication against PCA.

Method comparison

We now explore the performance of Algorithm 1 in a few simulated examples. We consider four configurations, where each configuration refers to a choice of latent space $\mathcal{Z}$ and corresponding kernel f. In the first configuration, the latent space comprises six distinct elements. The latent spaces in the remaining configurations are different subsets of $\mathbb{R}^2$. In each configuration, we draw $n = 500$ points Z_i uniformly on $\mathcal{Z}$, and the resulting point sets are shown in figures 6a)1-4.

In the first configuration, we choose an arbitrary 6×6 positive-definite matrix to represent the kernel. In the second, $f(x, y) = (x^\top y + 1)^2$, which has rank 6; in the third, $f(x, y) = \{\cos(x^{(1)} - y^{(1)}) + \cos(x^{(2)} - y^{(2)}) + 2\}$, which has rank 5; and in the fourth, $f(x, y) = \exp(-\|x - y\|^2_{\mathbb{R}^2}/2)$, which has infinite rank.

We simulate a 500×1000 data matrix $\mathbf{Y}$ in each configuration, where the $p = 1000$ random functions are independent, zero mean Gaussian processes with the same covariance kernel f, and the errors $\mathbf{E}_{ij}$ are independent and standard normal.

In figures 6c)1-4 we show the Wasserstein error (log-scale), i.e., the distance computed in Algorithm 1, for dimensions up to $\rho_{\max} = 30$. Reassuringly, the optimum roughly coincides with the true rank of the kernel when finite (dashed black line, configurations 1-3) and at the same time it is interesting that a non-degenerate optimum is still found under infinite rank (configuration 4). If we lower the noise, the optimal dimension increases (figure 17, Appendix), reflecting the afore-mentioned bias/variance trade-off.

For comparison, we also show the dimensions selected using the ladle [Luo and Li, 2016] and elbow methods [Zhu and Ghodsi, 2006], as implemented in the R packages 'dimension' (on GitHub: https://github.com/WenlanzZ) and 'igraph' (on The Comprehensive R Archive Network), respectively. The ladle and Wasserstein methods seem to make similar choices, although as implemented the ladle method is computationally costly, which has precluded more simulations or going beyond $\max(n, p) = 1000$ to allow a more comprehensive comparison. We would advise against the elbow method for dimension selection under the LMS model, as it appears to favour dangerously low dimensions.

In configurations 3 and 4, there is isometry between $\mathcal{M}$ and $\mathcal{Z}$. As a result, we can aim to recover the path lengths in $\mathcal{Z}$ amongst $Z_1, \ldots, Z_n$ from $p^{-1/2}\zeta_1, \ldots, p^{-1/2}\zeta_n$ – see section 4.4 for details. This method yields infinite distances when the k-nearest neighbour graph isn't connected. Dealing with this issue in a systematic way is awkward, and we settled on the following solution. Picking ϵ as the 5% quantile of the $\mathbb{R}^2$ Euclidean distance matrix between the Z_i, we place an edge between any pair of points within distance ϵ, weighted by Euclidean distance, and approximate the geodesic distance between two points as the corresponding weighted graph distance. Any infinite distance remaining is replaced with the original Euclidean distance. The blue line in figures 6d)3-4 shows the entrywise mean square error between the estimated geodesic distance matrices of $Z_1, \ldots, Z_n$ and $p^{-1/2}\zeta_1, \ldots, p^{-1/2}\zeta_n$, for different choices of r. The optimum roughly coincides with the dimensions selected by the ladle and Wasserstein methods.

Because of the isometric relationship between $\mathcal{M}$ and $\mathcal{Z}$ in configurations 3 and 4, we might also hope that the persistence diagrams of their Rips filtrations would be similar. The red line in figures 6d)3-4 shows the bottleneck distance between the persistence diagrams of the Rips filtrations of $Z_1, \ldots, Z_n$ and $p^{-1/2}\zeta_1, \ldots, p^{-1/2}\zeta_n$, as implemented in the R package 'TDA', for different choices of r. In this metric, the optimal dimension (lowest bottleneck distance) is lower than that suggested by the ladle and Wasserstein methods, but we do not know to what extent this should be expected in general. The scales of the log-Wasserstein error, geodesic distance error, and bottleneck distance are not comparable and in figures 6d) we have recentered and rescaled the curves to make their maxima and minima agree.

In figures 6e)1-4 we show the persistence diagrams of the Rips filtrations of $p^{-1/2}\zeta_1, \ldots, p^{-1/2}\zeta_n$ computed on the basis of the dimension selected by the Wasserstein method (Algorithm 1), using the R package 'TDA'. Recall that in persistent homology the significance of a topological feature is quantified by its persistence, death minus birth, which is the vertical distance between the point (birth,death) to the diagonal $x = y$. Following Fasy et al. [2014] we draw a line parallel to $x = y$ to separate the signal from the noise, picking $y = x + 0.2$ by eye. In each figure, we report the number of connected components, $\hat{\beta}_0$, and holes, $\hat{\beta}_1$, estimated by this heuristic. The true corresponding values for $\mathcal{Z}$ are respectively (6,0), (1,8), (1,0), and (1,1).

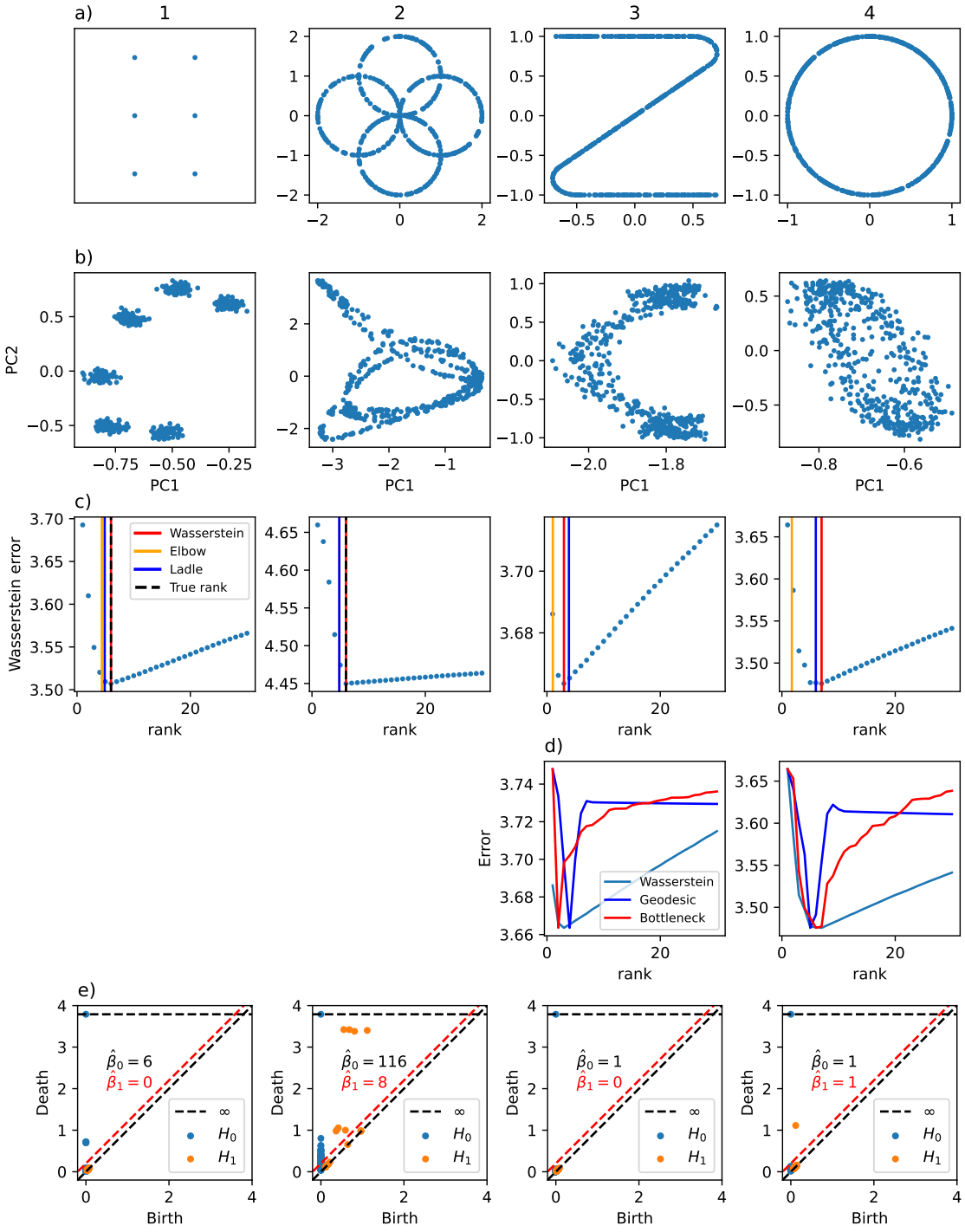

Figure 6: PCA dimension selection. Columns 1-4: different latent space/kernel configurations. 1-3 are finite rank, 4 infinite rank; configurations 3 and 4 are isometric. Row a: sampled positions ($n = 500$); b: first two principal components ($p = 1000$); c: the dimension selected by different methods, and the true rank when finite; d: error in geodesic distance and persistence diagram estimation (bottleneck distance) for the isometric configurations; e: persistence diagrams showing partial recovery of true topological features. Further details in main text.

Algorithm 1 PCA dimension selection

Input: data matrix $\mathbf{Y} \in \mathbb{R}^{n \times p}$;
maximal dimension $\rho_{\max} \leq \min(n, p)$, default $\rho_{\max} = \min(n, p)$.

1: Split the data as $\mathbf{Y}^{(1)} := \mathbf{Y}_{1:\lceil n/2 \rceil, 1:p}$, $\mathbf{Y}^{(2)} := \mathbf{Y}_{(\lceil n/2 \rceil + 1):n, 1:p}$
2: **for** $\rho \in \{1, ..., \rho_{\max}\}$ **do**
3: Let $\mathbf{V}_\rho^{(1)} \in \mathbb{R}^{p \times \rho}$ denote the matrix of orthogonal eigenvectors associated with the ρ largest eigenvalues of $\mathbf{Y}^{(1)\top}\mathbf{Y}^{(1)}$.
4: Project $\mathbf{Y}^{(1)}$ onto the linear span of the columns of $\mathbf{V}_\rho^{(1)}$, that is, compute: $\mathbf{Y}^{(1)}\mathbf{\Pi}_\rho$ where $\mathbf{\Pi}_\rho := \mathbf{V}_\rho^{(1)}\mathbf{V}_\rho^{(1)\top}$.
5: Compute Wasserstein distance d_ρ between $\mathbf{Y}^{(1)}\mathbf{\Pi}_\rho$ and $\mathbf{Y}^{(2)}$ as point sets in $\mathbb{R}^p$.
6: **end for**

Output: selected dimension $\hat{r} = \operatorname{argmin}\{d_\rho\}$.

4.3 Spherical projection

When performing data analysis we may wish to consider the assumption that f belongs to one of the families of kernels in proposition 3 or 4, because of their stationarity interpretation, and because the associated isometry properties would justify use of the PCA embedding to recover geometric features of $\mathcal{Z}$. However, all these kernels have the property that

$$p^{-1}\sum_{j=1}^{p}\mathbb{E}[|X_j(z)|^2] = f(z,z) = \text{const.}, \tag{20}$$

which from a modelling point of view may be restrictive. We shall now show that the spherically projected PCA embedding has a model-based interpretation which allows this restriction to be loosened.

Suppose we are given random functions $X_1, \ldots, X_p$ such that $f(z,z)$ is constant in $z \in \mathcal{Z}$. Without loss of generality assume this constant is 1. As usual let ϕ be the feature map associated with f. Note that in this situation $\mathcal{M}$ is a subset of the unit hypersphere $\{x \in \ell_2 : \|x\|_{\ell_2} = 1\}$. With $\alpha_1, \ldots, \alpha_n$ being i.i.d. random variables whose distribution is supported on a compact set $\mathcal{A} \subset \mathbb{R}_+$, define the model:

$$\mathbf{Y}_{ij} = \alpha_i X_j(Z_i) + \sigma\mathbf{E}_{ij}. \tag{21}$$

This can be viewed as a particular form of LMS model with extended latent space $\mathcal{Z}^{\text{ext}} := \mathcal{A} \times \mathcal{Z}$ and extended random functions $X_j^{\text{ext}}(\alpha, z) := \alpha X_j(z)$. Its mean correlation kernel is:

$$f^{\text{ext}}(\alpha, z, \alpha', z') := \frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\left[X_j^{\text{ext}}(\alpha, z)X_j^{\text{ext}}(\alpha', z')\right] = \alpha\langle\phi(z), \phi(z')\rangle_{\ell_2}\alpha',$$

the Mercer feature map of f^{ext} satisfies: $\phi^{\text{ext}}(\alpha, z) = \alpha\phi(z)$, and we have $p^{-1}\sum_{j=1}^{p}\mathbb{E}[|X_j^{\text{ext}}(\alpha, z)|^2] = \alpha^2$, allowing more flexibility than (20).

Now suppose assumption **A6** holds, let $\zeta_1, \ldots, \zeta_n$ be the dimension-r PCA embedding computed from $\mathbf{Y}$ under the extended LMS model (21) and consider the spherical projection $\zeta_i^{\text{sp}} := \zeta_i/\|\zeta_i\|_2$. Using the identities $\|\phi^{\text{ext}}(\alpha, z)\|_2 = \alpha\|\phi(z)\|_2 = \alpha f(z,z)^{1/2} = \alpha$, and applying the triangle inequality several times gives:

$$\begin{aligned}\|\mathbf{Q}\zeta_i^{\text{sp}} - \phi(Z_i)\|_2 &= \left\|\frac{p^{-1/2}\mathbf{Q}\zeta_i}{p^{-1/2}\|\zeta_i\|_2} - \frac{\phi^{\text{ext}}(\alpha_i, Z_i)}{\|\phi^{\text{ext}}(\alpha_i, Z_i)\|_2}\right\|_2 \\ &\leq 2\frac{\|p^{-1/2}\mathbf{Q}\zeta_i - \phi^{\text{ext}}(\alpha_i, Z_i)\|_2}{\alpha_i - \|p^{-1/2}\mathbf{Q}\zeta_i - \phi^{\text{ext}}(\alpha_i, Z_i)\|_2},\end{aligned}$$

where $\mathbf{Q}$ is any orthogonal matrix. Theorem 1 could therefore be applied to the LMS model (21) to establish that for the particular $\mathbf{Q}$ in that theorem, $\|p^{-1/2}\mathbf{Q}\zeta_i - \phi^{\text{ext}}(\alpha_i, Z_i)\|_2 \to 0$, which by the above inequality implies $\|\mathbf{Q}\zeta_i^{\text{sp}} - \phi(Z_i)\|_2 \to 0$. In summary, under the model (21), $\phi(Z_1), \ldots, \phi(Z_n)$ can be recovered from the spherically projected embedding $\zeta_1^{\text{sp}}, \ldots, \zeta_n^{\text{sp}}$, up to an orthogonal transformation.

4.4 Nearest neighbour graph construction

Constructing a nearest neighbour graph from the PCA embedding allows us to approximate topological and geometric features of $\mathcal{M}$ and hence $\mathcal{Z}$. In keeping with the workflow set out in section 1, we focus on the spherically projected embedding $\zeta_1^{\text{sp}}, \ldots, \zeta_n^{\text{sp}}$ but very similar considerations apply to the raw embedding $\zeta_1, \ldots, \zeta_n$. Noting that $\|\zeta_i^{\text{sp}}\|_2 = 1$ for all i, we denote by $d_{\mathbb{S}}(\zeta_i^{\text{sp}}, \zeta_j^{\text{sp}}) \coloneqq \arccos(\langle \zeta_i^{\text{sp}}, \zeta_j^{\text{sp}} \rangle_2)$ the circular arc distance on the unit hypersphere.

There are two popular types of nearest neighbour graph: the ϵ-nn and k-nn graphs, both of which are undirected, weighted graphs with n vertices, identified with $\zeta_1^{\text{sp}}, \ldots, \zeta_n^{\text{sp}}$. There is an edge between ζ_i^{sp} and ζ_j^{sp} in the ϵ-nn graph if $d_{\mathbb{S}}(\zeta_i^{\text{sp}}, \zeta_j^{\text{sp}}) \leq \epsilon$, and in the k-nn graph if ζ_i^{sp} is one of the k-nearest (with respect to $d_{\mathbb{S}}$) neighbours of ζ_j^{sp} or vice versa. In both types of graph, if there is an edge between ζ_i^{sp} and ζ_j^{sp} it is assigned weight $d_{\mathbb{S}}(\zeta_i^{\text{sp}}, \zeta_j^{\text{sp}})$. A number of algorithms for identifying nearest neighbours exactly or approximately are available, for example in the Python library `scikit-learn` [Pedregosa et al., 2011].

Recalling (33), nearest neighbour graph distances can be used to approximate shortest path lengths in $\mathcal{M}$:

$$d_{\mathcal{M}}^{\text{geo}}(\phi(Z_i), \phi(Z_j)) \approx \mathbf{D}_{\mathcal{M}}^{ij} \coloneqq \min_{x_1, \ldots, x_m} \{ d_{\mathbb{S}}(x_1, x_2) + \cdots + d_{\mathbb{S}}(x_{m-1}, x_m) \},$$

where the minimum is over all paths in the nearest neighbour graph connecting $x_1 = \zeta_i^{\text{sp}}$ and $x_m = \zeta_j^{\text{sp}}$. If there are no such paths, $\mathbf{D}_{\mathcal{M}}^{ij} = \infty$ by convention.

Various fast algorithms for computing shortest paths and shortest path lengths are available, for example in the Python library `NetworkX` [Hagberg et al., 2008], which can be parallelised `nx-parallel` (https://github.com/networkx/nx-parallel). If further speed-up in computing shortest paths length is needed, one might consider contraction hierarchies [Geisberger et al., 2008].

The use of nearest neighbour graphs to approximate path lengths on manifolds is well studied and is the first step in the Isomap procedure [Tenenbaum et al., 2000]. The theoretical accuracy of such approximations has been analysed by Bernstein et al. [2000], Trosset and Buyukbas [2020]. In particular Trosset and Buyukbas [2020] note that the k-nn graph is often preferred in practice although its analysis is more complicated. They also note that choosing a single value for ϵ or k is a difficult problem in general. Where possible in the examples of section 5 we compute and analyse the nearest neighbour graph over a range of values for ϵ or k, rather than selecting one single value. This approach is similar in spirit to the computation of ϵ-nn graphs over a range of ϵ values in persistent homology techniques [Carlsson, 2009, Chazal and Michel, 2021].

5 Examples

In the following three real data examples, we will assume the LMS model holds and explore hypotheses about the latent domain, $\mathcal{Z}$, feature map, ϕ, and the manifold underlying the data, $\mathcal{M} = \phi(\mathcal{Z})$. In all examples, we have access to background information. In some cases, such as the first hypothesis in the image and transcriptomics examples, the background points us towards generic hypotheses, such as '$\mathcal{M}$ is a loop' or '$\mathcal{M}$ is a tree'. In others, e.g. the second hypothesis in the image and transcriptomics examples, the information is used more explicitly to obtain trial values for $\mathcal{Z}$, realisations z_i of Z_i, and to estimate parts of the model.

In his famous book "Exploratory Data Analysis" [Tukey, 1977], Tukey observed that there was a readily accepted division in the process of criminal justice between "the search for the evidence — the responsibility of the police and other investigative forces — and the evaluation of the evidence's strength — a matter for juries and judges". In terms of this analogy, the workflow we present is directed towards the former activity — the search for clues, indications, appearances. In the examples to follow, we do not attempt to formally evaluate the strength of the evidence presented, beyond baseline comparisons against uniform models. We regard this confirmatory analysis as an important but distinct undertaking requiring different techniques.

The code and data used are available here: https://github.com/anniegray52/explore_manifold_hyp

5.1 Images

We return to the data set of images described in section 1. Recall there are $n = 72$ images, each consisting of $p = 110592$ grey-scale pixels, taken from angles $0, 5, 10, \ldots, 355$ degrees around the circumference of a circle. We will assume XY coordinates for the camera positions, $\cos(\theta_i), \sin(\theta_i)$ for each angle θ_i (converted to radians). In this context, we will first consider the hypothesis:

1. $\mathcal{Z}$ is a circle and ϕ is a homeomorphism. An informal implication is: The data lie close to a loop.

Finding this hypothesis to be tenable given the data, we will consider the stronger hypothesis:

2. $\mathcal{Z}$ is the circle of camera positions, $z_i = (\cos(\theta_i), \sin(\theta_i))$ are the (known) camera positions, and ϕ is a scaled isometry. An informal implication is: Distances along the loop correspond to distances along the circle between camera positions.

The first step of the workflow is to apply the dimension selection method. As per figure 7a), using $\rho_{\max} = 35$, this results in $\hat{r} = 11$. The kernel density estimate in 7b) demonstrates the variation in the magnitudes of the dimension-$\hat{r}$ PCA embedding vectors $\|\zeta_1\|_2, \ldots, \|\zeta_n\|_2$; in all subsequent steps we instead work with the spherically projected embedding $\zeta_1^{\text{sp}}, \ldots, \zeta_n^{\text{sp}}$ as per (4.3).

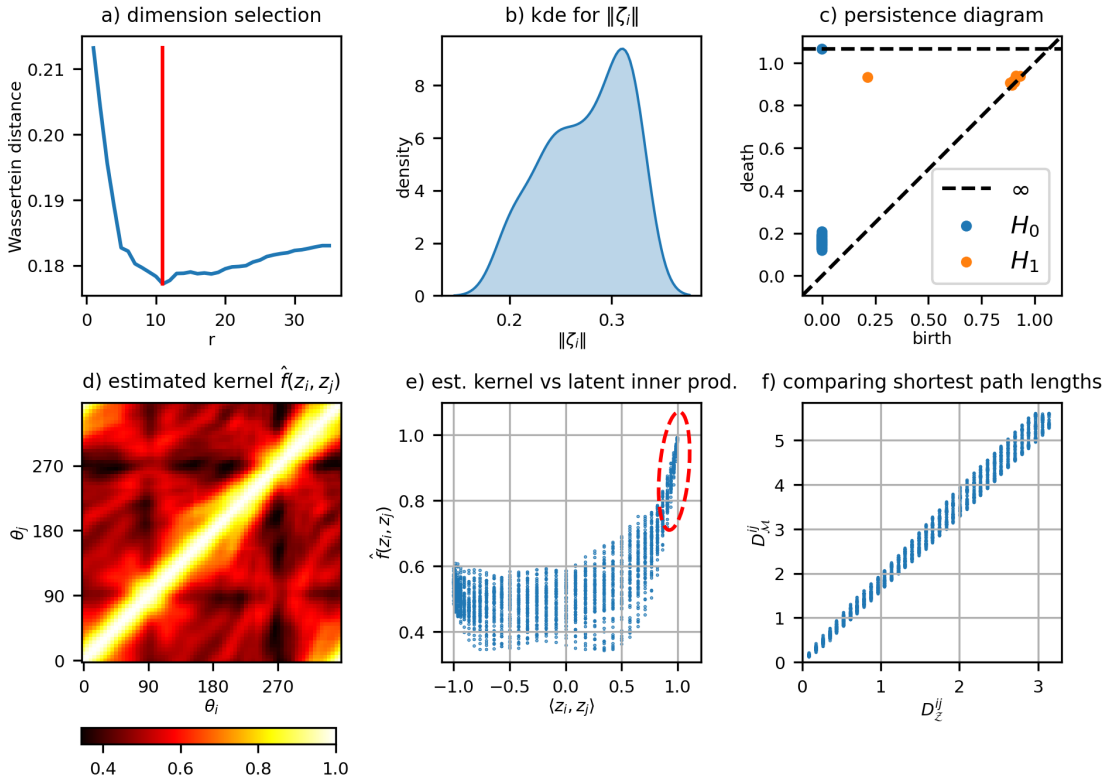

Figure 7: Images example. a) Wasserstein dimension selection; red vertical line indicates minimum at $\hat{r} = 11$. b) Kernel density estimate for the magnitudes of the PCA embedding vectors. c) Persistence diagram shows evidence of a single "loop" in the embedding. d) Estimated kernel as a function of latent positions in angular form $\theta_i = \arctan(z_i^{(2)}/z_i^{(1)})$. e) Estimated kernel as a function of latent inner product $\langle z_i, z_j \rangle$, the red dashed ellipse highlights $\hat{f}(z_i, z_j)$ in the region $\langle z_i, z_j \rangle \approx 1$. f) Evidence of a linear relationship between shortest path lengths computed from the nearest neighbour graph $\mathcal{G}$ (y-axis), and from the latent positions (x-axis).

We now consider the first hypothesis, which would be mathematically justified by assumption **A2**, with $\mathcal{Z}$ (say) the unit circle on $\mathbb{R}^2$ ($d_{\mathcal{Z}}$ the Euclidean metric). Then indeed ϕ would be

a homeomorphism and $\mathcal{M}$ would be topologically equivalent to a circle. Conveniently in this example, the presence of loops or holes in data point-clouds can be assessed using persistent homology techniques [Carlsson, 2009, Chazal and Michel, 2021]. Figure 7c) shows a persistence diagram computed from the spherically projected PCA embedding using the Python package `Ripser.py` [Tralie et al., 2018]. The blue dot on the horizontal dashed line is indicative of a single connected component with persists over a large range of length scales. The isolated single orange dot close to the horizontal dashed line is indicative of a single "loop" in the embedding, also persisting over a large range of length scales. We did not find any higher dimensional features, checking up to dimension 6, beyond which computation was prohibitive. This is all consistent with the hypothesis.

We now consider the second hypothesis, which by proposition 4 would be mathematically justified by the mean correlation kernel being of the form $f(z, z') = g(\langle z, z' \rangle)$ in some neighbourhood of $z = z'$ (or equivalently $\langle z, z' \rangle = 1$), and $g'(1) > 0$ (recall proposition 4). Assuming that $z_i = (\cos(\theta_i), \sin(\theta_i))$, figure 7d) shows $\hat{f}(z_i, z_j) \coloneqq \langle \zeta_i^{\text{sp}}, \zeta_j^{\text{sp}} \rangle$, which we regard as an estimator of $f(z_i, z_j)$, plotted as a function of θ_i and θ_j. The fairly constant width of the pronounced yellow/white diagonal stripe in this plot admits the interpretation that indeed $f(z, z') = g(\langle z, z' \rangle)$ in a neighbourhood of $z = z'$. To examine this in more detail, figure 7e) plots values of $\hat{f}(z_i, z_j)$ against $\langle z_i, z_j \rangle$ over all $i, j = 1, \ldots, n$. The red dashed ellipse highlights that $\hat{f}(z_i, z_j)$ is approximately an increasing function of $\langle z_i, z_j \rangle$ in a neighbourhood of $z_i = z_j$, which is consistent with $g'(1) > 0$. Informed by (13) we thus anticipate there is isometry between $\mathcal{M}$ and $\mathcal{Z}$, up a scaling factor of $g'(1)^{1/2}$. To see if the data allow for such a relationship we compute the k-nn graph $\mathcal{G}$ as per section 4.4 with $k = 2$. This is the natural choice for k if $\mathcal{M}$ is topologically equivalent to a circle. Figure 7f) shows shortest path lengths $\mathbf{D}_{\mathcal{M}}^{ij}$ in $\mathcal{G}$ plotted against shortest path lengths around the circle, denoted $\mathbf{D}_{\mathcal{Z}}^{ij}$, over $i, j = 1, \ldots, n$. The clear linear relationship is consistent with there being little deviation from isometry, up to a scaling factor, which by a straight line fit in figure 7f) we can estimate: $g'(1) \approx 3.18$.

5.2 Single-cell transcriptomics

We now revisit the planaria single-cell transcriptomics example introduced in section 1. Recall that here we have $p = 5821$ dimensional gene expression data in $n = 5000$ cells from adult planarians, and we also know cell-type labels for each of these cells, indicated by the different colours in figure 2. Adult planarians have a large number of pluripotent stem cells, known as neoblasts, that continuously differentiate into all adult cell types, resulting in a lineage tree that connects all the cells in the whole animal. We represent this lineage by a continuous tree (formally defined later) and suppose the cell types are named positions, c_i, on this tree.

In this context, we will first consider the hypothesis:

1. $\mathcal{Z}$ is a continuous tree and ϕ is a homeomorphism. An informal implication is: The data lie close to a tree.

Finding this hypothesis to be tenable given the data, we will consider the stronger hypothesis:

2. $\mathcal{Z}$ is the lineage tree, $z_i = c_i$ are the (known) cell types, and ϕ is a homeomorphism. Informally: the tree represents the lineage of the cell types.

In graph theory, a tree is an undirected graph in which any two vertices are connected by a unique path. We consider an analogue of this concept which reflects the continuous nature of cell differentiation. Inspired by definitions in [Morgan and Shalen, 1984, Janson, 2023], we say a metric space $\mathcal{Z}$ is a *continuous tree* if for any $z, z' \in \mathcal{Z}$ there exists a homeomorphism ψ between $[0, 1]$ and some subset of $\mathcal{Z}$ such that $\psi(0) = z$ and $\psi(1) = z'$ (this means a continuous path in $\mathcal{Z}$ exists between z and z'), and all such homeomorphisms have the same image (this means the path is unique).

Following the workflow in section 1, dimension selection with $\rho_{\max} = 50$ results in $\hat{r} = 14$. We then calculate the dimension-$\hat{r}$ embedding and its spherical projection $\zeta_1^{\text{sp}}, \ldots, \zeta_n^{\text{sp}}$. For the remainder of this section we refer to the latter as the PCA embedding.

We now consider the first hypothesis, which would be mathematically justified by assumption **A2**. Then $\mathcal{M}$ equipped with the ℓ_2 distance also qualifies as a continuous tree, as the composition

of two homeomorphisms is a homeomorphism. To gain some preliminary insight into the structure of the PCA embedding, figure 8a) shows, in red, a histogram of inner products between all distinct pairs of embedding points $\zeta_i^{\text{sp}}, \zeta_j^{\text{sp}}$. As a baseline comparison, we generated a random embedding consisting of the same number $n = 5000$ points uniformly distributed on the $\hat{r}$-dimensional, unit-radius hypersphere. Figure 8a) shows, in black, a histogram of inner-products between all distinct pairs of points in this random embedding. We see this black histogram is symmetrical and concentrated around 0. By contrast, the red histogram is not symmetrical and exhibits two peaks. The peak near an inner product value of 1 indicates a substantial proportion of pairs of points $\zeta_i^{\text{sp}}, \zeta_j^{\text{sp}}$ which are much closer together than is observed in the random embedding. Many other pairs $\zeta_i^{\text{sp}}, \zeta_j^{\text{sp}}$ have inner products between -0.5 and 0, indicating they are spread out on the hypersphere, but not in the same way that uniformly random points are spread out. On the basis of this preliminary check we see no reason to rule out tree structure in the PCA embedding. We

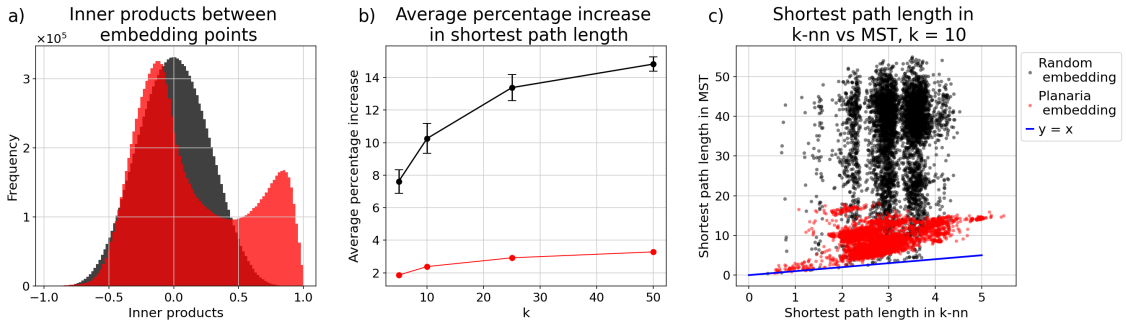

Figure 8: Single-cell transcriptomics example. a) histogram of inner products between distinct points in the PCA and random embeddings. b) average percentage increase in shortest path length in the minimum spanning tree compared to the k-nn graph, over different values of k. Results for the random embedding are shown in black, over 10 simulations with error bars indicated 2×standard error, c) comparing the shortest path lengths for samples in 10-nn graph and the MST.

now quantify how "tree-like" the PCA embedding is in two steps. First we compute the k-nn graph of the PCA embedding as per section 4.4, and its minimum spanning tree. The latter is obtained by removing edges from the k-nn graph until a tree is formed, in such a way that the total edge length of the tree is minimal. Various fast algorithms for computing minimum spanning trees are available, we used the Python library `NetworkX` [Hagberg et al., 2008]. The second step is to compare shortest path lengths in the k-nn graph to those in the minimum spanning tree. The shortest path length between any pair of vertices in the minimum spanning tree can only be greater than or equal to the shortest path length between those vertices in the k-nn graph. The percentage increase in shortest path length, when averaged over all pairs of vertices, serves as a univariate statistic which quantifies how tree-like the k-nn graph is. If the k-nn graph were a tree, this statistic would be exactly zero.

Figure 8b) shows the average percentage increase in shortest path length, as a function of k. The red line shows the results for the PCA embedding. The black line and error bars show the same quantity computed from repeated simulations of the random embedding, serving as a baseline for comparison. We see that across all values of k, the average percentage increase in shortest path length is much lower for the PCA embedding than for the random embedding. This indicates that the minimum spanning tree is a close approximation to the k-nn graph of the PCA embedding. To take a finer-grained look, figure 8c) shows shortest path lengths in the minimum spanning tree, versus in the k-nn graph with $k = 10$, for a sample of 5000 pairs of vertices. The blue "$y = x$" line indicates the lower bound on path length increase which would be achieved if the k-nn graph were a tree. Overall, the hypothesis that the planaria data are tree-like seems tenable given the findings in figure 8.

We now consider the second hypothesis. The left plot in figure 9 shows a visualisation of the minimum spanning tree derived from the k-nn graph with $k = 10$, with vertices coloured by the known cell type labels. This visualisation was obtained using the Scaleable Force Directed

Placement graph layout algorithm [Hu, 2005]. From the colouring by cell type, we see that biologically similar types, such as the three types of muscle cell, appear in localised branches of the tree.

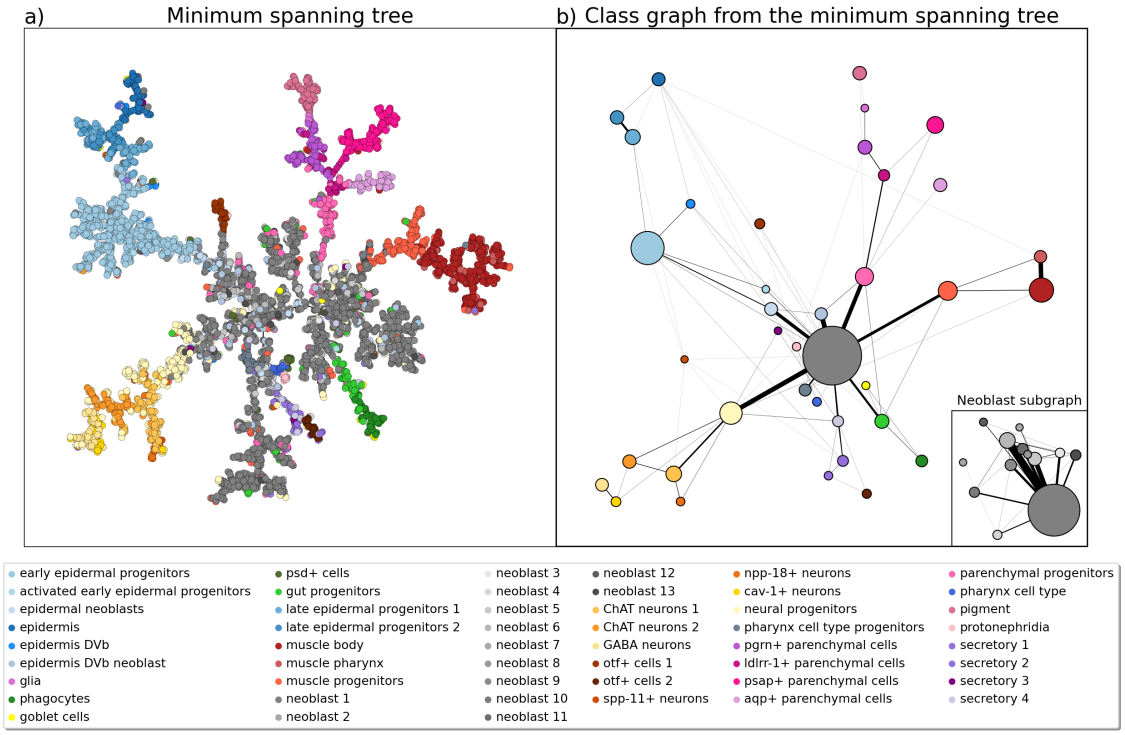

Figure 9: Single-cell transcriptomics example. a) minimum spanning tree computed from the spherically projected PCA embedding of the planaria data, colours indicate cell types. b) the class graph formed from the minimum spanning tree. All neoblast cell types are represented by a single dark grey node. The class subgraph consisting only of neoblast types is shown in the bottom right-hand corner inset.

We next construct a "class graph" which captures the relationships between cell types implied by the minimum spanning tree in figure 9a). In this class graph, each vertex corresponds to a cell type, and the undirected edge weight between any two vertices in the class graph is defined to be the total number of edges in the minimum spanning tree between cells of those two types.

The class graph is shown in figure 9b). The size of each node represents the total number of cells of that type. The thickness of the edges reflects their weights in the class graph, although for visual clarity we do not draw some edges with very low weights. Figure 9 elucidates cell development, tracing the lineage from stem cells to progenitors and differentiated cell populations: neurons, muscle cells, protonephridia, epidermis, and secretory cells.

The original paper of Plass et al. [2018] provides a consolidated tree, which amalgamates various evidence types. The overall structure aligns with our nearest neighbour approach, with branches for individual known cell types, however, discrepancies exist in the form of minor variations in the differentiated cell populations. For example, cav-1+ neurons connect to ChAT neurons 1 rather than neural progenitors. The connections from the muscle pharynx to the muscle progenitors and from the epidermis to the epidermal lineage, found based on marker gene analysis by Plass et al. [2018] coincide with the results from the nearest neighbour approach employed here. Acknowledging these differences, we refrain from delving further into minor disparities, given the current paper's intended scope.

For visual clarity, we draw a single node grouping together all the neoblast 1-13 cell types. The subgraph of the class graph corresponding to these neoblast types is shown in the inset of figure 9b), revealing a large number of neoblast 1 cells, linked by edges to most other neoblast cell types. This aligns with the results of the original authors, but contrasts with previous studies

[Van Wolfswinkel et al., 2014, Molina and Cebrià, 2021], which suggested distinct fates for various neoblast types. These disparities might be due to the unique ability of specialised neoblast cells to maintain pluripotency [Raz et al., 2021] or the sensitivity of the single-cell transcriptomic method, as in [Plass et al., 2018].

5.3 Temperature time series

In this example the raw data are time series of average daily temperatures in $n = 265$ towns and cities, on $p = 1450$ days. The data originate from the Berkeley Earth project [ber]. Our objective is to explore the relationship between temperature deviations and geographic locations of the towns and cities. To do so we take the ith data vector $\mathbf{Y}_i$ to be the temperature time series for town or city i centered about its long-run average. Thus geometry of the data point-cloud $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ as specified by the inner products $p^{-1}\langle \mathbf{Y}_i, \mathbf{Y}_j \rangle$ reflects the lag-zero cross-correlations amongst the time series.

In this context, we will first consider the hypothesis:

1. $\mathcal{Z}$ is a geographic region, $z_i = (\text{latitute}_i, \text{longitude}_i)$ are the (known) geographical locations of towns or cities, ϕ is a scaled isometry. An informal implication is: geodesic distances in $\mathcal{M}$ reflect geographical distances.

As a relaxation of the above we also ask if we can at least entertain:

2. $\mathcal{Z}$ is a geographic region, $z_i = (\text{latitute}_i, \text{longitude}_i)$ are the geographical locations of towns or cities, ϕ is a scaled isometry in certain subregions.

Following the workflow from section 1, figure 10a) shows the results of dimension selection with $\rho_{\max} = 70$, figure 10b) illustrates the variability of the magnitudes $\|\zeta_i\|_2$ of the non-projected embedding vectors, and we work henceforth with the spherically projected embedding as per section 4.3.

We now consider the first hypothesis, which by proposition 4 would be mathematically justified if the temperatures on a given day were stationary processes over Earth (a 'sphere'). If isometry between $\mathcal{Z}$ and $\mathcal{M}$ were to hold up to a scaling factor, then the k-nearest neighbours of z_i amongst $\{z_j; j \neq i\}$ would correspond to the k-nearest neighbours of $\phi(z_i)$ amongst $\{\phi(z_j); j \neq i\}$, with respect to $d^{\text{geo}}_{\mathcal{M}}$. In order to see if the data are consistent with the hypothesis of isometry, we therefore compute the proportion of edges in common between the embedding k-nn graph $\mathcal{G}$ (as per section 4.4), and the geographic k-nn graph defined by the known locations $z_1, \ldots, z_n$. Figure 10c) shows this proportion as a function of k. As a baseline to help interpret these results, we sampled n points uniformly from the $\hat{r}$-dimensional unit hypersphere, derived the k-nn graph from these points, then computed the proportion of edges in common with the geographic k-nn graph. This was repeated independently 100 times, and the resulting minimum, mean and maximum proportions of edges in common for each k are shown in red and black in figure 10c). The correspondence between $\mathcal{G}$ and the geographic k-nn graph is much better than under this uniform model. However, we see that as k increases up to 50, the embedding k-nn graph has about 70% of edges in common with the geographic k-nn graph, but increasing k further up to about $k = 130$ does not increase this percentage further. This plateauing suggests isometry does not hold.

We now consider the second hypothesis. The plateauing leaves open the possibility that there may be coincidence between the embedding and geographic k-nn graphs in some localised areas of $\mathcal{Z}$ but not in others. Indeed isometry as in (11) or (12) requires equality of shortest path lengths *for all* $z, z' \in \mathcal{Z}$. Figure 11 shows the locations of the towns and cities, and the edges in the embedding k-nn graph $\mathcal{G}$, with $k = 5$ chosen so that according to figure 10c) the embedding and geographic k-nn graphs have around 50% of edges in common. We see from figure 11 that in some regions, especially in central Europe, edges in the embedding k-nn graph generally correspond to geographic proximity, but elsewhere this correspondence does not hold. For example there are edges connecting Edinburgh, U.K., to cities in Norway which are not amongst its geographically nearest neighbours. Similarly, there are edges connecting Novorossiysk, Russia, to cities on the opposite shore of the Black Sea which are not amongst its geographically nearest neighbours. Conversely, geographic proximity does not always imply presence of an edge. For example, there

are no edges between Baia Mare, Romania, and two geographically close cities directly to the east, on the other side of the Carpathian mountains.

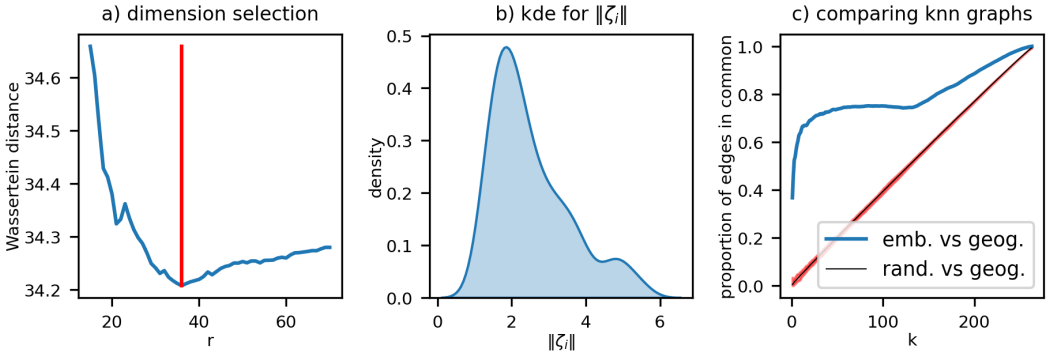

Figure 10: Temperatures example. a) Wasserstein dimension selection; red line indicates minimum at $\hat{r} = 36$. b) Kernel density estimate of the probability density of PC score magnitudes. c) The blue curve shows proportion of edges in common between embedding k-nn graph and geographic k-nn graph. The black line shows the mean proportion in common between the k-nn graph of a 100 uniformly random embeddings and the geographic k-nn graph. The red band indicates the range between maximum and minimum proportions across these 100 random embeddings.

Plotting the k-nn graph $\mathcal{G}$ in this way shows presence or absence of edges, but it doesn't convey the weight of these edges in the k-nn graph $\mathcal{G}$, which as per section 4.4, can approximate distances $d^{\text{geo}}_{\mathcal{M}}$. Since the embedding is of dimension $\hat{r} = 36$, it is challenging or perhaps impossible to construct a two-dimensional visualisation which faithfully conveys all aspects of its geometry. However, the visualisation task is much simpler if we choose some town or city, and then visualise the shortest paths in the embedding k-nn graph from that city to all other cities — the graph consisting of the union of all such paths is sometimes called a *shortest path tree*.

Figure 12 shows the shortest paths in $\mathcal{G}$ from Tallinn, Estonia, to all other towns and cities. Each such path is a sequence of towns or cities, and is visualised as a spline with knot points given by the locations of these towns and cities, with colour indicating length. Tallinn was chosen because of the different relationships between these shortest paths and geographic shortest paths which can be seen in different regions: the shortest paths in $\mathcal{G}$ which terminate at some towns and cities in central Europe, to the south-west of Tallinn, resemble geographic shortest paths, indicating a geometric relationship not far removed from isometry. By contrast, the red dots in figure 12 highlight the shortest path in $\mathcal{G}$ from Tallinn to Tripoli, Libya. This path passes through Sweden, Norway, the U.K., Ireland, France, Spain, back to France and then Italy. Clearly, this is not the geographically shortest path from Tallinn to Tripoli, indicating a strong deviation from isometry in these regions. Recalling from section 3.3 the relationship between weak stationarity and isometry, this deviation from isometry implies a pronounced lack of stationarity (with respect to geographic location) in these regions. This prompts us to consider what factors might disrupt temperature correlations along the geographically shortest path: for the case of Talinn to Tripoli, it seems reasonable to conjecture that changes in altitude, e.g. the Alps, and the Adriatic and Mediterranean seas might be such factors. We stress that our analysis by no means formally assesses the evidence that such factors are at play. Rather, it is through the process of inspecting figure 12 that we are led simply to consider such ideas as a step in data exploration.

6 Connections and conclusions

Geometric representation of high dimensional data. In their seminal JRSSB paper, Hall et al. [2005] introduced the perspective that, if $\mathbf{Y}_i$ and $\mathbf{Y}_j$ are i.i.d. random vectors whose elements satisfy suitable weak dependence and moment conditions, then $p^{-1}\|\mathbf{Y}_i - \mathbf{Y}_j\|^2$ converges to a constant as $p \to \infty$. This leads to a conclusion that i.i.d. high-dimensional data vectors tend to lie deterministically at the vertices of a simplex. The LMS model deviates from the assumption of i.i.d.

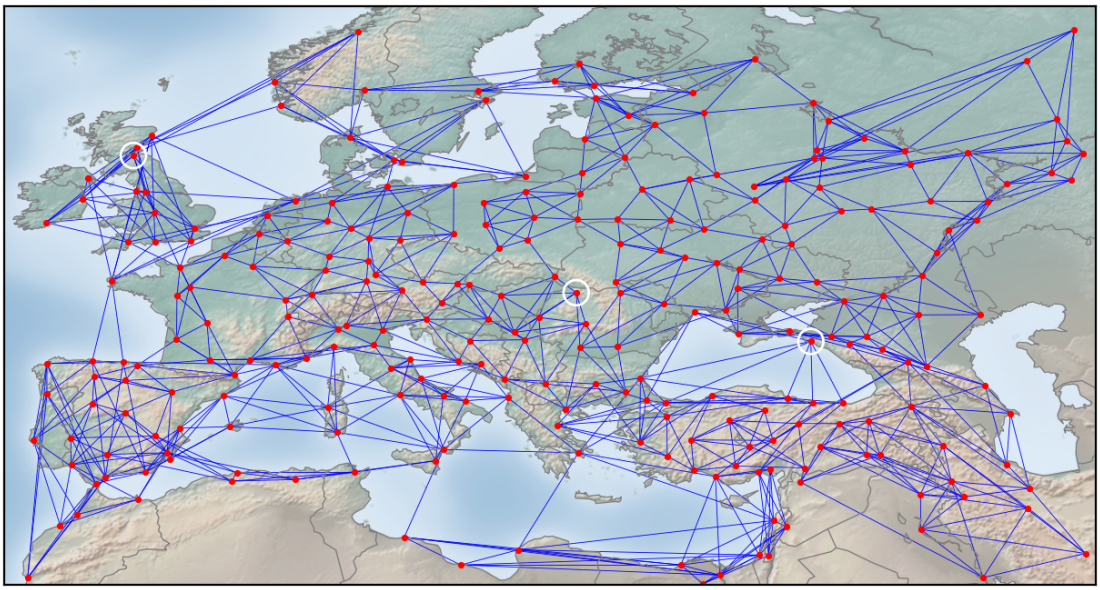

Figure 11: Temperatures example. Locations of towns and cities are shown in red. The blue lines correspond to edges in the embedding k-nn graph $\mathcal{G}$, with $k = 5$. The white circles highlight, from west to east: Edinburgh, U.K.; Baia Mare, Romania; and Novorossiysk, Russia.

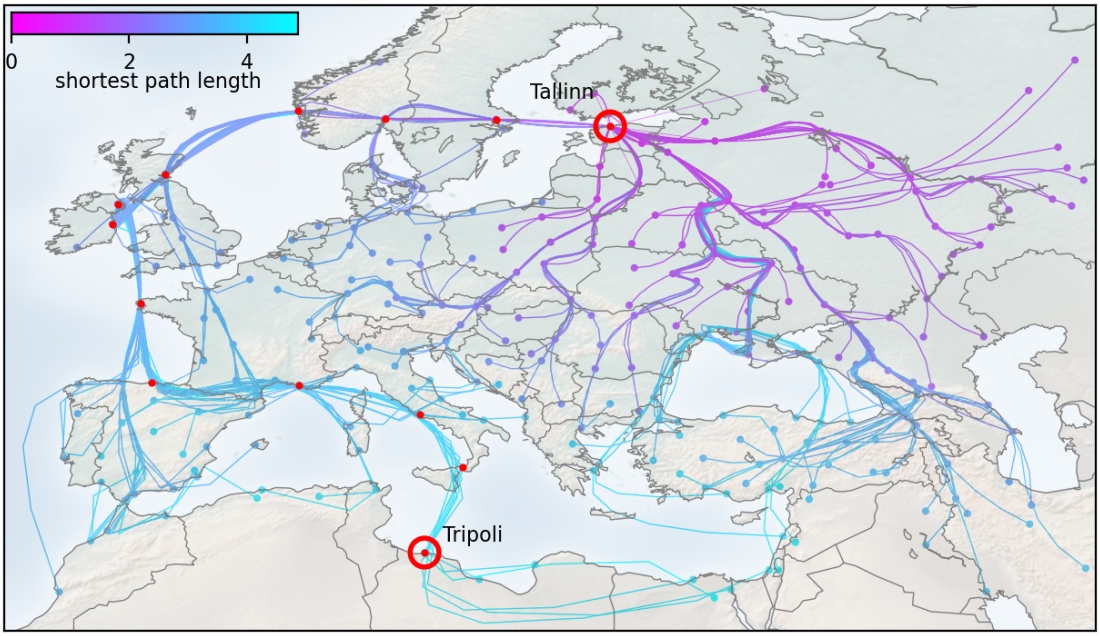

Figure 12: Temperatures example. Shortest paths in the embedding k-nn graph $\mathcal{G}$ from Tallinn, Estonia, to all other towns and cities. Each shortest path is visualised as a spline, with knot points given by the geographic locations of its constituent towns and cities. The red dots highlight the shortest path from Tallinn to Tripoli, Libya.

data vectors; under the LMS model, the "noise-free" data vectors $\mathbf{Y}_i - \sigma\mathbf{E}_i \equiv [X_1(Z_i) \cdots X_p(Z_i)]^\top$, $i = 1, \ldots, n$, are exchangeable but not independent. This dependence, combined with the latent variables $Z_1, ..., Z_n$, is key to the emergence of manifold structure in high dimensional data under the LMS model. In this way, the LMS model and our theoretical results extend and enrich the

geometric perspective of Hall et al. [2005], broadening the scope of high dimension low sample size (HDLSS) asymptotics [Shen et al., 2016, Aoshima et al., 2018].

Reflecting on the Manifold Hypothesis. Conventional interpretation of the Manifold Hypothesis as per the quote from [Cayton, 2005] in section 1 is that data vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \in \mathbb{R}^p$ are samples from some distribution supported on a manifold embedded in $\mathbb{R}^p$, perhaps subject to noise disturbances. Our analysis of the LMS model in section 3 provides a more nuanced perspective: $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are noisy, random projections of samples on a manifold $\mathcal{M}$; the manifold itself is a high-dimensional distortion of some latent domain $\mathcal{Z}$ and arises due to correlation over $\mathcal{Z}$. Under appropriate assumptions, $\mathcal{M}$ is homeomorphic or isometric to $\mathcal{Z}$.

Infinite exchangeable arrays. In how much generality is this perspective applicable? Inspired by remarks of Udell and Townsend [2019] in the context of latent variable models of low-rank matrices, we note the basic structure of the LMS model,

$$\mathbf{Y}_{ij} = X_j(Z_i) + \sigma \mathbf{E}_{ij}, \tag{22}$$

resembles a representation formula for exchangeable arrays due to Aldous [1981]: if $\mathbf{Y}$ is *any* infinite two-dimensional array of random variables such that permutations of its rows or columns do not alter the distribution of $\mathbf{Y}$, then there exists a function h such that the following equality in distribution holds

$$\mathbf{Y}_{ij} \overset{d}{=} h(\xi, Z_i, X_j, E_{ij}) \tag{23}$$

where ξ and the Z_i's, X_j's and E_{ij}'s are i.i.d. $\mathcal{U}[0,1]$-distributed random variables. Putting aside the fact that in the LMS model the rows of $\mathbf{Y}$ are exchangeable but the columns need not be, the resemblance between (22) and (23) indicates that the LMS model is rather general, albeit constrained to an additive form of error. The ability of PCA to extinguish noise, as characterised in theorem 1, seems closely tied to this additive structure.

PCA in high dimensions. The behaviour of PCA and principal component scores in high-dimensions has been the subject of intensive theoretical study, e.g., [Paul, 2007, Johnstone and Lu, 2009, Jung and Marron, 2009, Yata and Aoshima, 2009, Lee et al., 2010, Yata and Aoshima, 2012, Jung et al., 2012, Shen et al., 2013, 2016, Hellton and Thoresen, 2017]. A central theme in these works is analysis of the eigenvectors of the sample covariance matrix $n^{-1}\mathbf{Y}^\top\mathbf{Y} \in \mathbb{R}^{p\times p}$, which make up the columns of the matrix $\mathbf{V}_\mathbf{Y}$ appearing in (15). It is usually assumed that the data follow a spiked covariance model (a special case of the LMS model in which $\mathcal{Z}$ is Euclidean, f is linear and $X_1, \ldots, X_p$ are deterministic - see appendix B for details), with consideration given to various scaling relationships involving p, n and population covariance eigenvalues. In some situations with growing dimension $p \to \infty$, the sample covariance eigenvectors are inconsistent estimators of their population counterparts, such as when the eigenvalues are constant, n is growing and $p/n \to c > 0$ [Johnstone and Lu, 2009], or when n is fixed and the eigenvalues grow sublinearly as $p \to \infty$ [Jung and Marron, 2009, Jung et al., 2012].

Theorem 1 addresses high-dimensionality, but its proof does not entail establishing consistency of $\mathbf{V}_\mathbf{Y}$. Instead it starts with an elementary linear algebra argument (lemma 5 in section D) which shows that $p^{-1/2}\mathbf{Y}\mathbf{V}_\mathbf{Y} = \mathbf{U}_\mathbf{Y}\mathbf{\Lambda}_\mathbf{Y}^{1/2}$, where the columns of $\mathbf{U}_\mathbf{Y} \in \mathbb{R}^{n\times r}$ are orthonormal eigenvectors of $p^{-1}\mathbf{Y}\mathbf{Y}^\top \in \mathbb{R}^{n\times n}$ with associated eigenvalues on the diagonal of $\mathbf{\Lambda}_\mathbf{Y}$. The $\sqrt{n/p}$ term in (16) relates to the concentration behavior of the $n \times n$ matrix $p^{-1}\mathbf{Y}\mathbf{Y}^\top$ about its conditional expectation: $\mathbb{E}\left[(p^{-1}\mathbf{Y}\mathbf{Y}^\top)_{ij}|Z_i, Z_j\right] = p^{-1}\mathbb{E}\left[\langle \mathbf{Y}_i, \mathbf{Y}_j\rangle|Z_i, Z_j\right]$, c.f. (9). The $1/\sqrt{n}$ term in (16) concerns approximations to certain integrals with respect to μ, based on the samples $Z_1, \ldots, Z_n$, which arise when relating the rows of $\mathbf{U}_\mathbf{Y}\mathbf{\Lambda}_\mathbf{Y}^{1/2}$ to $\phi(Z_1), \ldots, \phi(Z_n)$.

The proof of theorem 4, from which theorem 1 is derived, relies heavily on matrix decomposition techniques used by Lyzinski et al. [2016] in the study of spectral embedding of random graphs under a random dot product model. The uniform (in $i = 1, \ldots, n$) nature of theorem 1 is directly inspired by the uniform consistency result of Lyzinski et al. [2016][Thm. 15], which is an instance of convergence with respect to the $2 \to \infty$ matrix norm, studied in detail by Cape et al. [2019]. We note more generally that singular vector estimation under low-rank assumptions is an active

area of research. As a recent example, Agterberg et al. [2022] obtained finite sample bounds and a Berry-Esseen type theorem for singular vectors under a model in which the signal is a deterministic low-rank matrix and heteroskedasticity and dependence is allowed in additive sub-Gaussian noise. Concurrent with the present work, Tong et al. [2025] have obtained uniform bounds for PCA matrix-denoising in a setting where a signal lies in a low-dimensional subspace and is observed in the presence of independent, high-dimensional sub-Gaussian noise. They obtain a lower-bound on the error of the denoised data matrix which indicates that PCA denoising is rate-optimal. In the work of Khoo et al. [2024], the same denoising techniques are discussed in the context of ingenious spectral methods for recovering temporal labels from noisy dynamical data, with links to manifold learning.

Truncated spectral embedding of graphs under a model with an infinite rank kernel was studied by Tang et al. [2013], but their results concern the Frobenius norm, hence a weaker, non-uniform measure of error than the $\max_{i=1,\ldots,n}$ error in theorem 1. It remains to be seen if a uniform consistency result similar to theorem 1 can be obtained for the LMS model with an infinite rank kernel.

Gaussian Process Latent Variable Models. The Gaussian Process Latent Variable Model (GPLVM) was devised by Lawrence [2003, 2005] in order to formulate dimension reduction as a statistical inference problem. The GPLVM can be viewed as a special case of the LMS model in which $\mathcal{Z}$ is a subset of $\mathbb{R}^d$, the random functions $X_1, \ldots, X_p$ are i.i.d., zero-mean Gaussian processes, and the elements of the noise matrix $\mathbf{E}$ are Gaussian. Under these assumptions the formula (8) from proposition 1, written in matrix form, is:

$$\mathbf{Y} \overset{m.s.}{=} p^{1/2}\mathbf{\Phi}\mathbf{W}^\top + \sigma\mathbf{E},$$

with $\mathbf{\Phi} \coloneqq [\phi(Z_1)|\cdots|\phi(Z_n)]^\top \in \mathbb{R}^{n\times r}$ and the elements of $\mathbf{E}$ and $p^{1/2}\mathbf{W} \in \mathbb{R}^{r\times p}$ are i.i.d. $\mathcal{N}(0,1)$. The latter property of $\mathbf{W}$ is equivalent to $X_1, \ldots, X_p$ being i.i.d., zero-mean Gaussian processes. After integrating out $\mathbf{W}$ and $\mathbf{E}$ analytically, the p columns of $\mathbf{Y}$ are i.i.d. with common distribution $\mathcal{N}(\mathbf{0}_n, \mathbf{\Phi}\mathbf{\Phi}^\top + \sigma^2\mathbf{I}_n)$. Lawrence [2005, App. B] considered maximum likelihood estimation of $\mathbf{\Phi}$ when $r < \infty$, showing that for the GPLVM and any orthogonal matrix $\mathbf{Q}$,

$$\widehat{\mathbf{\Phi}}^{\text{MLE}} \coloneqq \mathbf{U}_\mathbf{Y}(\mathbf{\Lambda}_\mathbf{Y} - \sigma^2\mathbf{I}_n)^{1/2}\mathbf{Q}^\top \tag{24}$$

is a maximum likelihood estimator of $\mathbf{\Phi}$. Here $\mathbf{\Lambda}_\mathbf{Y}$ and $\mathbf{U}_\mathbf{Y}$ are as in the above discussion of PCA in high dimensions. To make the connection to the PCA embedding $\zeta_1, \ldots, \zeta_n$ recall that

$$p^{-1/2}[\zeta_1|\cdots|\zeta_n]^\top \equiv p^{-1/2}\mathbf{Y}\mathbf{V}_\mathbf{Y} \equiv \mathbf{U}_\mathbf{Y}\mathbf{\Lambda}_\mathbf{Y}^{1/2}, \tag{25}$$

where the columns of $\mathbf{V}_\mathbf{Y}$ are orthonormal eigenvectors of $\mathbf{Y}^\top\mathbf{Y}$. Comparing (24) to (25), we can interpret the rows of $p^{1/2}\widehat{\mathbf{\Phi}}^{\text{MLE}}$ as a modified PCA embedding derived from eigen-decomposition of $p^{-1}\mathbf{Y}\mathbf{Y}^\top - \sigma^2\mathbf{I}_n$ instead of $p^{-1}\mathbf{Y}\mathbf{Y}^\top$. Computing $\widehat{\mathbf{\Phi}}^{\text{MLE}}$ in practice clearly requires σ^2 to be assumed known and $\mathbf{\Lambda}_\mathbf{Y} \succeq \mathbf{I}_r\sigma^2$. It may be possible to use some of our theoretical results to study consistency of $\widehat{\mathbf{\Phi}}^{\text{MLE}}$, but we leave this for future research.

Under the assumption that the kernel f belongs to a given parametric family Lawrence [2005] proposed maximum a-posteriori estimation of $Z_1, \ldots, Z_n$ as a form of nonlinear dimension reduction. Further developments of the GPLVM include hierarchical structures [Lawrence and Moore, 2007], variational inference [Titsias and Lawrence, 2010, Damianou et al., 2016], connections to Locally Linear Embedding [Lawrence, 2012, Roweis and Saul, 2000], pseudo-marginal MCMC techniques [Gadd et al., 2021], and handling computational scaleability and missing data [Lalchand et al., 2022].

Linear then nonlinear dimension reduction. Nonlinear dimension reduction techniques are designed to extract low-dimensional structure from data for purposes of exploration and visualisation. These methods were pioneered by Tenenbaum et al. [2000] and Roweis and Saul [2000], who devised Isomap and Local Linear Embedding, respectively; subsequent contributions include Semi-definite Embedding [Weinberger et al., 2004]; latent variable-based methods, [Lawrence,

2003, Saul, 2020]; Diffusion Maps [Coifman et al., 2005], Laplacian and Hessian Eigenmaps [Belkin and Niyogi, 2003, Donoho and Grimes, 2003]; Stochastic Neighbour Embedding (SNE and t-SNE) [Hinton and Roweis, 2002, Van der Maaten and Hinton, 2008] and Uniform Manifold Approximation and Projection (U-MAP). Several such methods are easily accessible through the massively popular Python package `scikit-learn` [Pedregosa et al., 2011], and their impact is exemplified by the fact that, at the time of writing, the t-SNE paper of Van der Maaten and Hinton [2008] has over $48,000$ citations according to Google Scholar. Each of these techniques work on different principles, but in broad terms, they take as input a set of points in high-dimensional Euclidean space, and output a set of points in low-dimensional Euclidean space in a way which is designed to minimise some measure of distortion of pairwise distances or inner-products.

It has been advocated in the literature to reduce data to tens or hundreds of dimensions using PCA as a preprocessing step, before applying nonlinear dimension reduction to obtain a two or three-dimensional representation. For example, in the context of t-SNE, Van der Maaten and Hinton [2008] state *"This speeds up the computation of pairwise distances between the data points and suppresses some noise without severely distorting the interpoint distances"*. Similar recommendations are given in [Van Der Maaten, 2014, Kobak and Berens, 2019, Saul, 2020]. Up until now however, there has been no detailed or rigorous statistical justification for this preprocessing.

Our model and theory explain why a) the manifold might be concentrated within a low-dimensional subspace (e.g. section 3.4); and b) why applying PCA can give us an extremely sharp view (i.e., uniformly consistent, theorem 1). This reinforces the message that nonlinear dimensionality reduction techniques need not be viewed as an alternative to PCA, but rather that the combination of the two, i.e., *linear then nonlinear dimensionality reduction*, may be particularly effective.

Incidentally, we recommend keeping the PCA embedding at hand to retain access to *ambient* properties of the manifold, for example, to estimate reach [Aamari et al., 2019], inspect inner products (figure 8), or evaluate the fidelity of low-dimensional visualisations such as t-SNE.

Exploratory Data Analysis. What are the limitations of the data analysis workflow we have proposed? This workflow is intentionally generic, and suitable for preliminary exploration of high-dimensional data and hypotheses about the data generating mechanism. It could serve as a first step before more detailed confirmatory analysis, in order to quantify uncertainty, perform formal hypothesis testing, or fit a parametric model, and so forth, but it clearly does not include those functionalities. We believe the methodology aligns with Tukey's philosophy of 'exploratory data analysis' [Tukey, 1977] although of course it is not 'model-free'. There is no contradiction there, and indeed several works have warned of the dangers of conflating these ideas [Hullman and Gelman, 2021].

References

Berkeley Earth. http://berkeleyearth.org. Accessed: 2021-05-27.

Eddie Aamari, Jisu Kim, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Estimating the reach of a manifold. *Electronic Journal of Statistics*, 13(1):1359 – 1399, 2019.

Joshua Agterberg, Zachary Lubberts, and Carey E Priebe. Entrywise estimation of singular vectors of low-rank matrices with heteroskedasticity and dependence. *IEEE Transactions on Information Theory*, 68(7):4618–4650, 2022.

David J Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

Makoto Aoshima, Dan Shen, Haipeng Shen, Kazuyoshi Yata, Yi-Hui Zhou, and James S Marron. A survey of high dimension low sample size asymptotics. *Australian & New Zealand journal of statistics*, 60(1):4–19, 2018.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Anil Aswani, Peter Bickel, and Claire Tomlin. Regression on manifolds: Estimation of the exterior derivative. *The Annals of Statistics*, 39(1):48–81, 2011.

Sivaraman Balakrishnan, Alesandro Rinaldo, Don Sheehy, Aarti Singh, and Larry Wasserman. Minimax rates for homology inference. In *Artificial Intelligence and Statistics*, pages 64–72. PMLR, 2012.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Mira Bernstein, Vin De Silva, John C Langford, and Joshua B Tenenbaum. Graph approximations to geodesics on embedded manifolds. https://users.math.msu.edu/users/iwenmark/Teaching/MTH995/Papers/MMod_BSLT00.pdf, 2000. online, accessed 14th March, 2022.

Peter J Bickel and Bo Li. Local polynomial regression on unknown manifolds. *Lecture Notes-Monograph Series*, pages 177–186, 2007.

R. H. Bing. Topological equivalence. *The American Mathematical Monthly*, 67(7):4–7, 1960. ISSN 00029890, 19300972. URL http://www.jstor.org/stable/2308625.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45, 2015.

Pratik Prabhanjan Brahma, Dapeng Wu, and Yiyuan She. Why deep learning works: A manifold disentanglement perspective. *IEEE transactions on neural networks and learning systems*, 27 (10):1997–2008, 2015.

Christoph Bregler and Stephen M Omohundro. Nonlinear manifold learning for visual speech recognition. In *Proceedings of IEEE International Conference on Computer Vision*, pages 494–499. IEEE, 1995.

Dmitri Burago. *A course in metric geometry*, volume 33. American Mathematical Soc., 2001.

Joshua Cape, Minh Tang, and Carey E Priebe. The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *The Annals of Statistics*, 47(5):2405–2439, 2019.

Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2): 255–308, 2009.

Kevin M Carter, Raviv Raich, and Alfred O Hero III. On local intrinsic dimension estimation and its applications. *IEEE Transactions on Signal Processing*, 58(2):650–663, 2009.

Lawrence Cayton. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, 12(1-17):1, 2005.

Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence*, 4, 2021.

Frédéric Chazal, Leonidas J Guibas, Steve Y Oudot, and Primoz Skraba. Persistence-based clustering in Riemannian manifolds. *Journal of the ACM (JACM)*, 60(6):1–38, 2013.

Ming-Yen Cheng and Hau-tieng Wu. Local linear regression on manifolds and its geometric interpretation. *Journal of the American Statistical Association*, 108(504):1421–1434, 2013.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems*, 35:25683–25696, 2022.

Ronald R Coifman, Stephane Lafon, Ann B Lee, Mauro Maggioni, Boaz Nadler, Frederick Warner, and Steven W Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences*, 102(21): 7426–7431, 2005.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26, 2013.

Andreas C. Damianou, Michalis K. Titsias, and Neil D. Lawrence. Variational inference for latent variables and uncertain inputs in Gaussian Processes. *Journal of Machine Learning Research*, 17(42):1–62, 2016.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:2406–2422, 2022.

Alex Diaz-Papkovich, Luke Anderson-Trocmé, Chief Ben-Eghan, and Simon Gravel. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS Genetics*, 15(11):e1008432, 2019.

David L Donoho and Carrie Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003.

Herbert Edelsbrunner and John Harer. Persistent homology — a survey. *Contemporary Mathematics*, 453:257–282, 2008.

Ahmed Elhag, Yuyang Wang, Josh Susskind, and Miguel Angel Bautista Martin. Manifold diffusion fields. In *NeurIPS Workshop*, 2023. URL https://arxiv.org/abs/2305.15586.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not All Language Model Features Are One-Dimensionally Linear. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Confidence sets for persistence diagrams. *The Annals of Statistics*, pages 2301–2339, 2014.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690, 2019.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *Journal of Machine Learning Research*, 22 (78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

C Gadd, Sara Wade, and AA Shah. Pseudo-marginal Bayesian inference for Gaussian process latent variable models. *Machine Learning*, 110(6):1105–1143, 2021.

Richard J Gardner, Erik Hermansen, Marius Pachitariu, Yoram Burak, Nils A Baas, Benjamin A Dunn, May-Britt Moser, and Edvard I Moser. Toroidal topology of population activity in grid cells. *Nature*, 602(7895):123–128, 2022.

Robert Geisberger, Peter Sanders, Dominik Schultes, and Daniel Delling. Contraction hierarchies: Faster and simpler hierarchical routing in road networks. In *Experimental Algorithms: 7th International Workshop, WEA 2008 Provincetown, MA, USA, May 30-June 1, 2008 Proceedings 7*, pages 319–333. Springer, 2008.

Christopher R Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Manifold estimation and singular deconvolution under Hausdorff loss. *The Annals of Statistics*, 40 (2):941–963, 2012a.

Christopher R Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Minimax manifold estimation. *Journal of Machine Learning Research*, 13:1263–1291, 2012b.

Jan-Mark Geusebroek, Gertjan J Burghouts, and Arnold WM Smeulders. The Amsterdam library of object images. *International Journal of Computer Vision*, 61(1):103–112, 2005. URL https://aloi.science.uva.nl. Retrieved March 2022.

Annie Gray, Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. Hierarchical clustering with dot products recovers hidden tree structure. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations ICLR*, 2024.

Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using NetworkX. Technical report, Los Alamos National Lab. (LANL), Los Alamos, NM (United States), 2008.

Peter Hall, James Stephen Marron, and Amnon Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(3):427–444, 2005.

Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J Zico Kolter, Ruslan Salakhutdinov, and Stefano Ermon. Manifold preserving guided diffusion. *arXiv preprint arXiv:2311.16424*, 2023.

Matthias Hein and Jean-Yves Audibert. Intrinsic dimensionality estimation of submanifolds in $\mathbb{R}^d$. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pages 289–296, 2005.

Kristoffer H Hellton and Magne Thoresen. When and why are principal component scores a good tool for visualizing high-dimensional data? *Scandinavian Journal of Statistics*, 44(3):581–597, 2017.

Geoffrey Hinton and Sam T Roweis. Stochastic neighbor embedding. In *Advanced in Neural Information Processing Systems (NeurIPS)*, volume 15, pages 833–840. Citeseer, 2002.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020.

Yifan Hu. Efficient, high-quality force-directed graph drawing. *Mathematica journal*, 10(1):37–71, 2005.

Jessica Hullman and Andrew Gelman. Designing for interactive exploratory data analysis requires theories of graphical inference. *Harvard Data Science Review*, 3(3):10–1162, 2021.

Svante Janson. Real trees. *arXiv preprint arXiv:2303.07920*, 2023.

Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

Sungkyu Jung and J Stephen Marron. PCA consistency in high dimension, low sample size context. *The Annals of Statistics*, 37(6B):4104–4130, 2009.

Sungkyu Jung, Arusharka Sen, and JS Marron. Boundary behavior in high dimension, low sample size asymptotics of PCA. *Journal of Multivariate Analysis*, 109:190–203, 2012.

Balázs Kégl. Intrinsic dimension estimation using packing numbers. *Advances in Neural Information Processing Systems (NeurIPS)*, 15, 2002.

Yuehaw Khoo, Xin T Tong, Wanjie Wang, and Yuguan Wang. Temporal label recovery from noisy dynamical data. *arXiv:2406.13635*, 2024.

Dmitry Kobak and Philipp Berens. The art of using t-sne for single-cell transcriptomics. *Nature communications*, 10(1):5416, 2019.

Samory Kpotufe. k-NN regression adapts to local intrinsic dimension. *Advances in neural information processing systems*, 24, 2011.

Samory Kpotufe and Sanjoy Dasgupta. A tree-based regressor that adapts to intrinsic dimension. *Journal of Computer and System Sciences*, 78(5):1496–1515, 2012.

Vidhi Lalchand, Aditya Ravuri, and Neil D Lawrence. Generalised GPLVM with stochastic variational inference. In *International Conference on Artificial Intelligence and Statistics*, pages 7841–7864. PMLR, 2022.

Oscar Lao, Timothy T Lu, Michael Nothnagel, Olaf Junge, Sandra Freitag-Wolf, Amke Caliebe, Miroslava Balascakova, Jaume Bertranpetit, Laurence A Bindoff, David Comas, et al. Correlation between genetic and geographic structure in Europe. *Current Biology*, 18(16):1241–1248, 2008.

Neil Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6(11), 2005.

Neil D Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 2, page 5, 2003.

Neil D Lawrence. A unifying probabilistic perspective for spectral dimensionality reduction: Insights and new models. *Journal of Machine Learning Research*, 13:1609–1638, 2012.

Neil D Lawrence and Andrew J Moore. Hierarchical Gaussian process latent variable models. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pages 481–488, 2007.

Chan-Su Lee and Ahmed Elgammal. Human motion synthesis by motion manifold learning and motion primitive segmentation. In *International Conference on Articulated Motion and Deformable Objects*, pages 464–473. Springer, 2006.

Seunggeun Lee, Fei Zou, and Fred A Wright. Convergence and prediction of principal component scores in high-dimensional settings. *Annals of statistics*, 38(6):3605, 2010.

Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in Neural Information Processing systems (NeurIPS)*, 17, 2004.

Lizhen Lin, Niu Mu, Pokman Cheung, and David Dunson. Extrinsic Gaussian processes for regression and classification on manifolds. *Bayesian Analysis*, 14(3):887–906, 2019.

Matthias Löffler, Anderson Y Zhang, and Harrison H Zhou. Optimality of spectral clustering in the Gaussian mixture model. *The Annals of Statistics*, 49(5):2506–2530, 2021.

Wei Luo and Bing Li. Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika*, 103(4):875–887, 2016.

Vince Lyzinski, Minh Tang, Avanti Athreya, Youngser Park, and Carey E Priebe. Community detection and classification in hierarchical stochastic blockmodels. *IEEE Transactions on Network Science and Engineering*, 4(1):13–26, 2016.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

J Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Transactions Royal Soc.*, 209:4–415, 1909.

Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. The origins of representation manifolds in large language models. *arXiv preprint arXiv:2505.18235*, 2025.

M Dolores Molina and Francesc Cebrià. Decoding stem cells: An overview on planarian stem cell heterogeneity and lineage progression. *Biomolecules*, 11(10):1532, 2021.

Kevin R Moon, Jay S Stanley III, Daniel Burkhardt, David van Dijk, Guy Wolf, and Smita Krishnaswamy. Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Current Opinion in Systems Biology*, 7:36–46, 2018.

John W Morgan and Peter B Shalen. Valuations, trees, and degenerations of hyperbolic structures, i. *Annals of Mathematics*, 120(3):401–476, 1984.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *The Journal of Machine Learning Research*, 21 (1):7018–7055, 2020.

Mu Niu, Pokman Cheung, Lizhen Lin, Zhenwen Dai, Neil Lawrence, and David Dunson. Intrinsic Gaussian processes on complex constrained domains. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):603–627, 2019.

Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1):419–441, 2008.

John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, 2008.

Neal Patwari and Alfred O Hero. Manifold learning algorithms for localization in wireless sensor networks. In *2004 IEEE international conference on acoustics, speech, and signal processing*, volume 3, pages iii–857. IEEE, 2004.

Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.

Daniel Paulin, Lester Mackey, and Joel A Tropp. Efron–Stein inequalities for random matrices. *The Annals of Probability*, 44(5):3431–3473, 2016.

A.R. Pears. *Dimension Theory of General Spaces*. Cambridge University Press, 1975.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jakiw Pidstrigach. Score-based generative models detect manifolds. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 35852–35865, 2022.

Mireya Plass, Jordi Solana, F Alexander Wolf, Salah Ayoub, Aristotelis Misios, Petar Glažar, Benedikt Obermayer, Fabian J Theis, Christine Kocks, and Nikolaus Rajewsky. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, 360(6391): eaaq1723, 2018.

Robert Pless and Richard Souvenir. A survey of manifold learning for images. *IPSJ Transactions on Computer Vision and Applications*, 1:83–94, 2009.

Amelie A Raz, Omri Wurtzel, and Peter W Reddien. Planarian stem cells specify fate yet retain potency during the cell cycle. *Cell Stem Cell*, 28(7):1307–1322, 2021.

Michael Reutlinger and Gisbert Schneider. Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery. *Journal of Molecular Graphics and Modelling*, 34:108–117, 2012.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

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

Lawrence K Saul. A tractable latent variable model for nonlinear dimensionality reduction. *Proceedings of the National Academy of Sciences*, 117(27):15403–15408, 2020.

Dan Shen, Haipeng Shen, Hongtu Zhu, and JS Marron. High dimensional principal component scores and data visualization. *arXiv preprint arXiv:1211.2679*, 2012.

Dan Shen, Haipeng Shen, Hongtu Zhu, and JS Marron. Surprising asymptotic conical structure in critical sample eigen-directions. *arXiv preprint arXiv:1303.6171*, 2013.

Dan Shen, Haipeng Shen, and James S Marron. A general framework for consistency of principal component analysis. *Journal of Machine Learning Research*, 17(150):1–34, 2016.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

Jan Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Your diffusion model secretly knows the dimension of the data manifold. *arXiv preprint arXiv:2212.12611*, 2022.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Wilson A Sutherland. *Introduction to metric and topological spaces*. Oxford University Press, 2nd edition, 2009.

Rustem Takhanov. On the speed of uniform convergence in Mercer's theorem. *Journal of Mathematical Analysis and Applications*, 518(2):126718, 2023.

Minh Tang, Daniel L Sussman, and Carey E Priebe. Universally consistent vertex classification for latent positions graphs. *The Annals of Statistics*, 41(3):1406–1430, 2013.

Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Michalis Titsias and Neil D Lawrence. Bayesian Gaussian process latent variable model. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 844–851. JMLR Workshop and Conference Proceedings, 2010.

Xin T Tong, Wanjie Wang, and Yuguan Wang. Uniform error bound for pca matrix denoising. *Bernoulli*, 31(3):2251–2275, 2025.

Christopher Tralie, Nathaniel Saul, and Rann Bar-On. Ripser.py: A lean persistent homology library for python. *The Journal of Open Source Software*, 3(29):925, Sep 2018. doi: 10.21105/joss.00925. URL https://doi.org/10.21105/joss.00925.

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Michael W Trosset and Gokcen Buyukbas. Rehabilitating Isomap: Euclidean representation of geodesic structure. *arXiv preprint arXiv:2006.10858*, 2020.

John W Tukey. *Exploratory data analysis*. Addison-Wesley series in behavioral science. Addison-Wesley Pub. Co., Reading, Mass, 1977. ISBN 0201076160.

Madeleine Udell and Alex Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019.

Laurens Van Der Maaten. Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*, 15(1):3221–3245, 2014.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.

Josien C Van Wolfswinkel, Daniel E Wagner, and Peter W Reddien. Single-cell analysis reveals functionally distinct classes within the planarian stem cell compartment. *Cell stem cell*, 15(3): 326–339, 2014.

Larry Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5: 501–532, 2018.

Kilian Q Weinberger, Fei Sha, and Lawrence K Saul. Learning a kernel matrix for nonlinear dimensionality reduction. In *Proceedings of the twenty-first international conference on Machine learning*, page 106, 2004.

Nick Whiteley, Annie Gray, and Patrick Rubin-Delanchy. Matrix factorisation and the interpretation of geodesic distance. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

Wang Xuejun, Hu Shuhe, Yang Wenzhi, and Shen Yan. On complete convergence for weighted sums of-mixing random variables. *Journal of Inequalities and Applications*, 2010:1–13, 2010.

Yun Yang and David B Dunson. Bayesian manifold regression. *The Annals of Statistics*, 44(2): 876–905, 2016.

Kazuyoshi Yata and Makoto Aoshima. PCA consistency for non-Gaussian data in high dimension, low sample size context. *Communications in Statistics: Theory and Methods*, 38(16-17):2634–2652, 2009.

Kazuyoshi Yata and Makoto Aoshima. Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *Journal of multivariate analysis*, 105(1): 193–215, 2012.

Kazuyoshi Yata and Makoto Aoshima. Geometric consistency of principal component scores for high-dimensional mixture models and its application. *Scandinavian Journal of Statistics*, 47(3): 899–921, 2020.

Mu Zhu and Ali Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51(2):918–930, 2006.

A Supporting results for section 2

The following version of Mercer's theorem can be found in [Steinwart and Christmann, 2008, Thm 4.49].

Theorem 2 (Mercer's theorem)**.** *Let $\mathcal{Z}$ be a compact metric space and let $f : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, be a symmetric, positive semi-definite, continuous function. Let μ be a finite Borel measure supported on $\mathcal{Z}$. Then there exists a countable collection of nonnegative real numbers $(\lambda_k^f)_{k\geq 1}$, $\lambda_1^f \geq \lambda_2^f \geq \ldots$ and $\mathbb{R}$-valued functions $(u_k^f)_{k\geq 1}$ which are orthonormal in $L_2(\mu)$, such that:*

$$f(z, z') = \sum_{k=1}^{\infty} \lambda_k^f u_k^f(z) u_k^f(z'), \quad z, z' \in \mathcal{Z},$$

where the convergence is absolute and uniform.

B Special cases of the LMS model

Spiked covariance model

The spiked covariance model [Johnstone, 2001, Paul, 2007] is the de facto standard model under which to study the theoretical properties of PCA, and is derived as follows. Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be a matrix of random variables such that $\mathbb{E}[\mathbf{X}^\top \mathbf{X}]$ has rank r. Consider the eigendecomposition $n^{-1}\mathbb{E}[\mathbf{X}^\top \mathbf{X}] = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$, where $\mathbf{V} \in \mathbb{R}^{p\times r}$, and define $\mathbf{Z} \coloneqq \mathbf{X}\mathbf{V}\mathbf{\Lambda}^{-1/2}$. We have

$$\mathbb{E}\left[\mathbf{Z}^\top \mathbf{Z}\right] = n\mathbf{I}_r, \qquad \mathbf{V}^\top \mathbf{V} = \mathbf{I}_r, \tag{26}$$

and $\mathbf{X} = \mathbf{Z}\mathbf{\Lambda}^{1/2}\mathbf{V}^\top$, a.s., where the latter equality can be checked by verifying

$$\mathbb{E}[\|\mathbf{X} - \mathbf{Z}\mathbf{\Lambda}^{1/2}\mathbf{V}^\top\|_{\mathrm{F}}^2] = \operatorname{tr}\mathbb{E}[(\mathbf{X} - \mathbf{Z}\mathbf{\Lambda}^{1/2}\mathbf{V}^\top)^\top(\mathbf{X} - \mathbf{Z}\mathbf{\Lambda}^{1/2}\mathbf{V}^\top)] = 0.$$

The spiked covariance model takes the form:

$$\mathbf{Y} = \mathbf{Z}\mathbf{\Lambda}^{1/2}\mathbf{V}^\top + \sigma\mathbf{E},$$

where the elements of $\mathbf{E} \in \mathbb{R}^{n\times p}$ are usually assumed to be zero-mean, unit variance and uncorrelated. The rows of $\mathbf{Z} \in \mathbb{R}^{n\times r}$ are called individual-specific random effects, and are usually assumed to be i.i.d. The following proposition shows that a spiked covariance model of precisely this form is a special case of the LMS model.

Proposition 5. *For any $r < \infty$, let the rows of $\mathbf{Z} \in \mathbb{R}^{n\times r}$ be i.i.d. random vectors such that the first equality in (26) holds, let $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_r)$ where $\lambda_1, \ldots, \lambda_r$ are any strictly positive real numbers and let $\mathbf{V} = [v_1|\cdots|v_p]^\top \in \mathbb{R}^{p\times r}$ be any deterministic matrix such that the second equality in (26) holds. Then, if $\mathbf{Y}$ follows the LMS model specified by:*

$$\mathcal{Z} \subset \mathbb{R}^r, \qquad [Z_1|\cdots|Z_n]^\top \coloneqq \mathbf{Z} \qquad X_j(z) \coloneqq \left\langle v_j, \mathbf{\Lambda}^{1/2} z \right\rangle, \tag{27}$$

the mean correlation kernel associated with this LMS model is:

$$f(z, z') = \frac{1}{p}\left\langle z, \mathbf{\Lambda} z' \right\rangle,$$

which has rank r, $\lambda_k^f = \lambda_k/p$, $[u_1^f(z) \cdots u_r^f(z)]^\top = z \in \mathbb{R}^r$ and the following identity holds:

$$\mathbf{Y} = \mathbf{Z}\mathbf{\Lambda}^{1/2}\mathbf{V}^\top + \sigma\mathbf{E}.$$

Proof. The claimed expression for $f(z, z')$ holds by substituting the definition of $X_j(z)$ in (27) into the definition $f(z, z') \coloneqq p^{-1}\sum_{j=1}^p \mathbb{E}[X_j(z)X_j(z')]$ and using the assumption of the proposition that $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_r$. The eigenfunctions $[u_1^f(z) \cdots u_r^f(z)]^\top = z \in \mathbb{R}^r$ are orthonormal due to the assumption $\mathbb{E}\left[\mathbf{Z}^\top\mathbf{Z}\right] = n\mathbf{I}_r$ and the i.i.d. nature of the rows of $\mathbf{Z}$. The expression for $\mathbf{Y}$ in the statement holds by substituting (27) into the definition of $\mathbf{Y}$ under the LMS model, i.e., $\mathbf{Y}_{ij} = X_j(Z_i) + \sigma\mathbf{E}_{ij}$. □

The relationship between the spiked covariance model (SCM) and the LMS model can thus be summarised as follows:

- the metric space $(\mathcal{Z}, d_{\mathcal{Z}})$ in the LMS model generalises the Euclidean domain of individual-specific random effects in the spiked covariance model;
- the eigenfunctions u_k^f, $k \geq 1$, in the LMS model generalise the linear dependence on individual-specific random effects in the SCM;
- the random functions X_j, $j = 1, \ldots, p$, in the LMS model generalise the deterministic, linear functions $\left\langle v_j, \mathbf{\Lambda}^{1/2} z \right\rangle$, $j = 1, \ldots, p$, which in light of (27) are implicit in the SCM;
- the LMS model allows for possibly infinite rank, generalising the finite-rank nature of the SCM.

Finite mixture model

Consider the case where $\mathcal{Z}$ has finitely many elements, say $\mathcal{Z} = \{1, \ldots, m\}$. For the following discussion it is not important that we take these elements to be the numbers $1, \ldots, m$, any m distinct abstract elements will do. In this situation the LMS model is a form of finite mixture model with random mixture centres. Indeed we see from:

$$\mathbf{Y}_{ij} = X_j(Z_i) + \sigma \mathbf{E}_{ij}$$

that $[X_1(z) \cdots X_p(z)]$ can be interpreted as the p-dimensional random centre of a mixture component labeled by $z \in \mathcal{Z}$, and the latent variable Z_i indicates which mixture component the ith row of the data matrix $\mathbf{Y}$ is drawn from. The simple form of the noise in the LMS model constrains the generality of this mixture model: recall the elements of $\mathbf{E}$ are independent across columns; elements in the same column but distinct rows are uncorrelated; all elements are unit variance.

To make $\mathcal{Z}$ into a metric space we consider the discrete metric $d_{\mathcal{Z}}(z, z') \coloneqq 0$ for $z = z'$, otherwise $d_{\mathcal{Z}}(z, z') \coloneqq 1$. The kernel f is specified by the matrix $\mathbf{F} \in \mathbb{R}^{m \times m}$ with entries

$$\mathbf{F}_{kl} \coloneqq \frac{1}{p} \sum_{j=1}^{p} \mathbb{E}[X_j(k) X_j(l)], \quad k, l \in \{1, \ldots, m\}.$$

In this situation **A1** and **A6** hold immediately, and $r \leq m$.

Topological equivalence of $\mathcal{M}$ and $\mathcal{Z}$ in this situation would mean that $\mathcal{M}$ consists of m distinct points $\{\phi(1), \ldots, \phi(m)\}$, each associated with exactly one element of $\mathcal{Z}$. If such topological equivalence were to hold then theorem 1 would tell us that the PCA embedding vectors will be clustered around the m distinct points $\{\mathbf{Q}^{-1}\phi(1), \ldots, \mathbf{Q}^{-1}\phi(m)\}$, with specifically $p^{-1/2}\zeta_i$ being close to $\mathbf{Q}^{-1}\phi(Z_i)$.

To verify topological equivalence it remains to check **A2** holds. To this end, suppose that $r = m$, i.e. $\mathbf{F}$ is full rank. Then it is not possible that any two rows of $\mathbf{F}$ are identical. That is, for $k, l \in \{1, \ldots m\}$ such that $k \neq l$, there must exist some $\xi \in \{1, \ldots, m\}$ such that $f(k, \xi) = \mathbf{F}_{k\xi} \neq \mathbf{F}_{l\xi} = f(l, \xi)$. Thus assumption **A2** is satisfied and hence $\mathcal{M}$ is topologically equivalent to $\mathcal{Z}$ if $r = m$.

In practical terms, we therefore see that in order to organise the n rows of $\mathbf{Y}$ into m clusters, one can first reduce dimension to $r = m$ by computing the PCA embedding and then apply some clustering technique to those embedding vectors. This two-step procedure of PCA followed by clustering, sometimes described as spectral clustering, is very popular in the practice of high-dimensional data analysis and is exactly what Yata and Aoshima [2020] recommend in the conclusion of their study of PCA embedding for mixture models in a regime where the number of samples is fixed and the dimension tends to infinity. It is already known that PCA, albeit under slightly different variations and assumptions, allows for "perfect clustering" in high-dimensional mixture models [Löffler et al., 2021, Agterberg et al., 2022].

To illustrate the behaviour of the LMS model and PCA embedding in this context, we consider a case in which $\mathcal{Z} = \{1, 2, 3\}$ and μ is the uniform distribution on $\mathcal{Z}$; for each $j = 1, \ldots, p$,

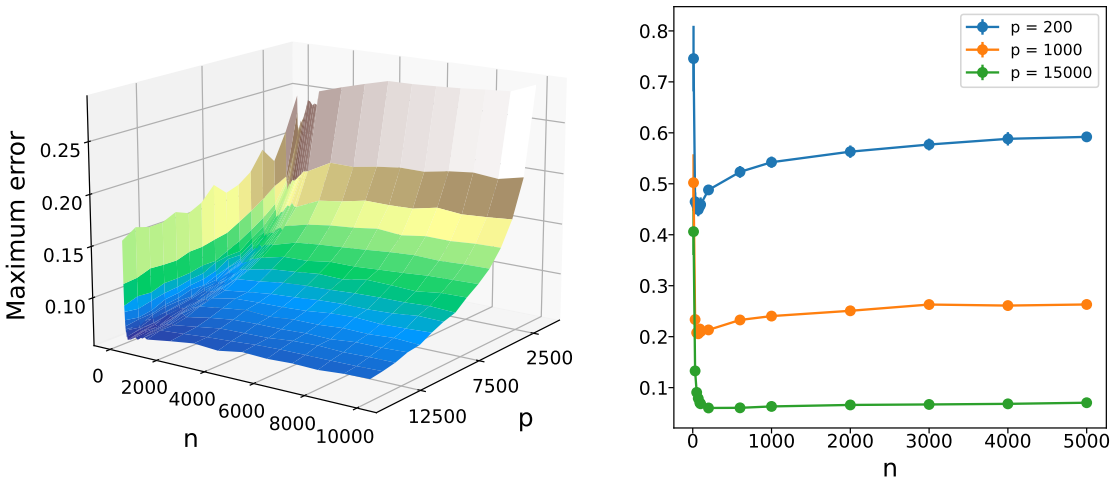

Figure 13: Mixture model example. Left: maximum error $\max_{i\neq j}\left|p^{-1/2}\|\zeta_i-\zeta_j\|_2-\|\phi(Z_i)-\phi(Z_j)\|_{\ell_2}\right|$, averaged over 50 independent realisations from the model, as a function of n and p. Right: the same error for $p=200,1000,15000$, as a function of n.

$[X_j(1)\ X_j(2)\ X_j(3)]^\top \sim \mathcal{N}(\mathbf{0},\mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is full-rank; and the elements of $\mathbf{E}$ are independent and identically distributed $\mathcal{N}(0,1)$ with $\sigma=1$. Figure 13 shows the error

$$\max_{i\neq j}\left|p^{-1/2}\|\zeta_i-\zeta_j\|_2-\|\phi(Z_i)-\phi(Z_j)\|_{\ell_2}\right|,$$

averaged over 50 independent realisations from the model. The plot on the left of the figure indicates that over the ranges considered, for fixed n the error decreases as p increases. Theorem 1 is not informative about the converse situation, when p is fixed and n increases: in this regime, the condition of theorem 4 involving a lower bound on n will eventually be satisfied, but the condition involving a lower bound on p/n will eventually be violated. We examine this in the right plot of figure 13. We see that for fixed p, as n increases the error initially quickly decreases, but then the appears to very slowly increase $n \gg p$. We conjecture the former and is related to the $1/\sqrt{n}$ term in (16).

Figure 14 illustrates how this error performance relates to the clustering of the PCA embedding vectors. When n is fixed, we see that as p increases the embedding vectors are increasingly tightly clustered around $\phi(1),\phi(2),\phi(3)$, in keeping with theorem 1. When p is fixed, we see that three clusters of embedding vectors are clearly discernible, but the clusters appear not to shrink as n grows.

Overall we conclude that, whilst theorem 1 shows that both n and p/n being large is sufficient to drive the error to zero, our numerical results suggest that for fixed p the error does not explode as n grows, and even when $n \gg p$ it may be that the PCA embedding still conveys the topological or geometric structure of $\mathcal{M}$ and hence $\mathcal{Z}$.

C Proofs and supporting material for section 3

Proof of proposition 1. Define

$$\widetilde{\mathbf{W}}_{jk} \coloneqq \int_{\mathcal{Z}} X_j(z)u_k^f(z)\mu(\mathrm{d}z), \tag{28}$$

and note that

$$\widetilde{\mathbf{W}}_{jk} = p^{1/2}(\lambda_k^f)^{1/2}\mathbf{W}_{jk}, \tag{29}$$

where $\mathbf{W}_{jk}$ is defined in (7).

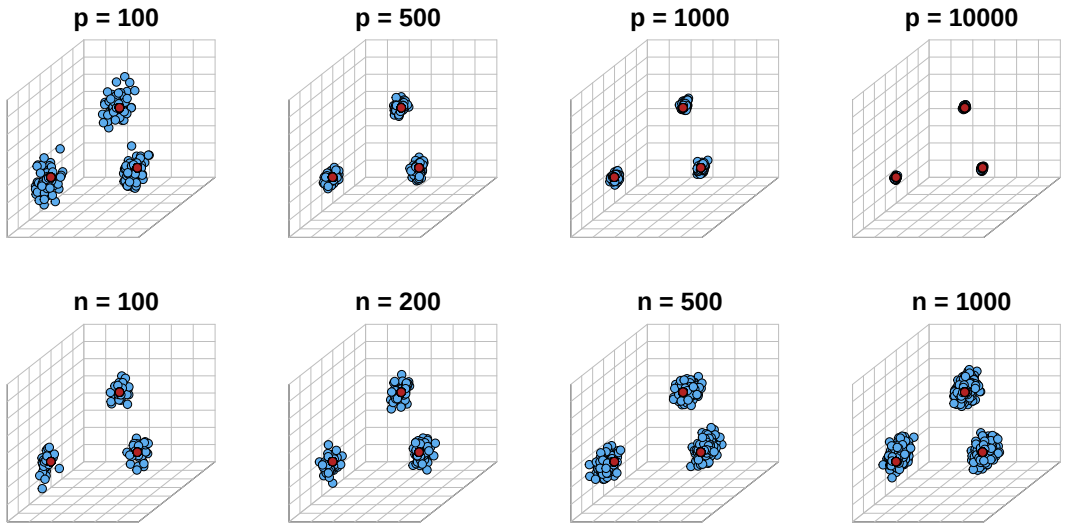

Figure 14: Mixture model example. PCA embedding $\{p^{-1/2}\zeta_1, \ldots, p^{-1/2}\zeta_n\}$ (blue dots) and $\phi(1), \phi(2), \phi(3)$ (red dots). Top row: n fixed to 200 and p varying. Bottom row p fixed to 200 and n varying.

Pick any $r_0 < r$ and recall $r \in \{1, 2 \ldots, \} \cup \{\infty\}$ is the number of nonzero eigenvalues $(\lambda_k^f)_{k \geq 1}$. We claim that, for any $z \in \mathcal{Z}$, the following equality holds:

$$\frac{1}{p} \sum_{j=1}^{p} \mathbb{E}\left[\left| X_j(z) - \sum_{k=1}^{r_0} u_k^f(z) \widetilde{\mathbf{W}}_{jk} \right|^2 \right] = f(z,z) - \sum_{k=1}^{r_0} \lambda_k^f |u_k^f(z)|^2. \tag{30}$$

To verify the equality (30), observe:

$$\begin{aligned}
&\frac{1}{p} \sum_{j=1}^{p} \mathbb{E}\left[\left| X_j(z) - \sum_{k=1}^{r_0} u_k^f(z) \widetilde{\mathbf{W}}_{jk} \right|^2 \right] \\
&= \frac{1}{p} \sum_{j=1}^{p} \mathbb{E}\left[|X_j(z)|^2 \right] - \frac{2}{p} \sum_{j=1}^{p} \mathbb{E}\left[X_j(z) \sum_{k=1}^{r_0} u_k^f(z) \widetilde{\mathbf{W}}_{jk} \right] \\
&\quad + \frac{1}{p} \sum_{j=1}^{p} \sum_{k=1}^{r_0} \sum_{\ell=1}^{r_0} \mathbb{E}\left[\widetilde{\mathbf{W}}_{jk} \widetilde{\mathbf{W}}_{j\ell} \right] u_k^f(z) u_\ell^f(z) \\
&= f(z,z) - 2 \sum_{k=1}^{r_0} u_k^f(z) \int_{\mathcal{Z}} f(z,z') u_k^f(z') \mu(\mathrm{d}z') \\
&\quad + \sum_{k=1}^{r_0} \sum_{\ell=1}^{r_0} u_k^f(z) u_\ell^f(z) \int_{\mathcal{Z}} \int_{\mathcal{Z}} f(z',z'') u_k^f(z') u_\ell^f(z'') \mu(\mathrm{d}z') \mu(\mathrm{d}z'') \\
&= f(z,z) - 2 \sum_{k=1}^{r_0} \lambda_k |u_k^f(z)|^2 + \sum_{k=1}^{r_0} \lambda_k^f |u_k^f(z)|^2 \\
&= f(z,z) - \sum_{k=1}^{r_0} \lambda_k^f |u_k^f(z)|^2,
\end{aligned}$$

where the second equality uses (28) and $f(z,z') = p^{-1} \sum_{j=1}^{p} \mathbb{E}[X_j(z) X_j(z')]$, and the third equality uses the fact that $(u_k^f, \lambda_k^f)_{k \geq 1}$, by definition, are $L_2(\mu)$-orthonormal eigenfunctions and eigenvalues of the integral operator associated with the kernel f and the measure μ.

By Mercer's theorem (theorem 2) the r.h.s. of (30) converges to zero as $r_0 \to r$, uniformly in z. Each of the summands on the l.h.s. of (30) is nonnegative, so they must also converge to zero uniformly in z. Using this uniform convergence and the fact that for any $j = 1, \ldots, p$ and $i = 1, \ldots, n$, the pair of random variables X_j and Z_i are statistically independent, we have:

$$\lim_{r_0 \to r} \mathbb{E}\left[\left|X_j(Z_i) - \sum_{k=1}^{r_0} u_k^f(Z_i)\widetilde{\mathbf{W}}_{jk}\right|^2\right] = \lim_{r_0 \to r} \int_{\mathcal{Z}} \mathbb{E}\left[\left|X_j(z) - \sum_{k=1}^{r_0} u_k^f(z)\widetilde{\mathbf{W}}_{jk}\right|^2\right] \mu(\mathrm{d}z)$$
$$\leq \lim_{r_0 \to r} \sup_z \mathbb{E}\left[\left|X_j(z) - \sum_{k=1}^{r_0} u_k^f(z)\widetilde{\mathbf{W}}_{jk}\right|^2\right] = 0. \tag{31}$$

Noting the identity (29) and recalling $\phi(z) = [(\lambda_1^f)^{1/2} u_1^f(z)\ (\lambda_2^f)^{1/2} u_2^f(z)\ \cdots]^\top$, we find that (31) can equivalently be written:

$$X_j(Z_i) \overset{m.s.}{=} p^{1/2} \langle \phi(Z_i), \mathbf{W}_j \rangle_{\ell_2},$$

where $\mathbf{W}_j$ is the jth row of $\mathbf{W}$. This completes the proof of the first identity in (8).

The second identity in (8) follows from the fact that $(u_k^f, \lambda_k^f)_{k \geq 1}$ are orthonormal eigenfunctions/values:

$$\sum_{j=1}^{p} \mathbb{E}[\mathbf{W}_{jk}\mathbf{W}_{j\ell}] = \frac{1}{(\lambda_k^f \lambda_\ell^f)^{1/2}} \int_{\mathcal{Z}} \int_{\mathcal{Z}} u_k^f(z) \frac{1}{p} \sum_{j=1}^{p} \mathbb{E}\left[X_j(z) X_j(z')\right] u_\ell^f(z') \mu(\mathrm{d}z') \mu(\mathrm{d}z)$$
$$= \frac{\lambda_\ell^f}{(\lambda_k^f \lambda_\ell^f)^{1/2}} \int_{\mathcal{Z}} u_k^f(z) u_\ell^f(z) \mu(\mathrm{d}z) = \begin{cases} 1, & k = \ell \\ 0, & k \neq \ell \end{cases}.$$

□

We introduce the following assumption in order to prove proposition 6 below.

A7. *For mixing coefficients φ satisfying $\sum_{k \geq 1} \varphi^{1/2}(k) < \infty$ and all $z, z' \in \mathcal{Z}$, the sequence $\{(X_j(z), X_j(z')); j \geq 1\}$ is φ-mixing.*

Assumption **A9** is stated in section D.

Proposition 6. *Assume **A1**, **A9** and **A7**, and let $q \geq 1$ and φ be as therein. Then there exists a constant $C(\varphi)$ depending only on φ such that for any $\delta > 0$ and any i, j,*

$$\mathbb{P}\left(\left|p^{-1} \langle \mathbf{Y}_i, \mathbf{Y}_j \rangle - \langle \phi(Z_i), \phi(Z_j) \rangle_{\ell_2} - \sigma^2 \mathbf{I}[i = j]\right| \geq \delta \middle| Z_i, Z_j\right) \leq \frac{1}{\delta^{2q}} \frac{1}{p^q} C(\varphi) M(q, \sigma)$$

where

$$M(q, \sigma) := \sup_{j \geq 1} \sup_{z \in \mathcal{Z}} \mathbb{E}\left[|X_j(z)|^{4q}\right]$$
$$+ \sigma \sup_{i,j \geq 1} \mathbb{E}\left[|\mathbf{E}_{ij}|^{2q}\right] \sup_{j \geq 1} \sup_{z \in \mathcal{Z}} \mathbb{E}\left[|X_j(z)|^{2q}\right]$$
$$+ \sigma^2 \sup_{i,j \geq 1} \mathbb{E}\left[|\mathbf{E}_{ij}|^{4q}\right].$$

Proof. Fix any i, j and consider the decomposition:

$$p^{-1} \langle \mathbf{Y}_i, \mathbf{Y}_j \rangle - \langle \phi(Z_i), \phi(Z_j) \rangle_{\ell_2} - \sigma^2 \mathbf{I}[i = j] = \sum_{k=1}^{4} \Delta_k$$

where

$$\Delta_1 := p^{-1} \langle \mathbf{X}(Z_i), \mathbf{X}(Z_j) \rangle - f(Z_i, Z_j)$$
$$\Delta_2 := p^{-1} \sigma \langle \mathbf{X}(Z_i), \mathbf{E}_j \rangle$$
$$\Delta_3 := p^{-1} \sigma \langle \mathbf{X}(Z_j), \mathbf{E}_i \rangle$$
$$\Delta_4 := p^{-1} \sigma^2 \langle \mathbf{E}_i, \mathbf{E}_j \rangle - \sigma^2 \mathbf{I}[i = j]$$

and $\mathbf{X}(z) \coloneqq [X_1(z) \dots X_p(z)]^\top$.

Writing Δ_1 as

$$\Delta_1 = \frac{1}{p}\sum_{k=1}^{p} \Delta_{1,k}, \qquad \Delta_{1,k} \coloneqq X_k(Z_i)X_k(Z_j) - \mathbb{E}\left[X_k(Z_i)X_k(Z_j)\middle| Z_i, Z_j\right].$$

we see that Δ_1 is an arithmetic mean of p random variables, each of which is conditionally mean-zero given Z_i, Z_j.

For Δ_2, we have

$$\Delta_2 = \frac{\sigma}{p}\sum_{k=1}^{p} \Delta_{2,k}, \qquad \Delta_{2,k} \coloneqq X_k(Z_i)\mathbf{E}_{jk},$$

By definition of the LMS model, the three collections of random variables, $(Z_1, \dots, Z_n)$, $(X_1, \dots, X_p)$ and $(\mathbf{E}_1, \dots, \mathbf{E}_n)$ are mutually independent, and the elements of each vector $\mathbf{E}_j \in \mathbb{R}^p$ are mean-zero and independent. Therefore given Z_i, Z_j and $X_1, \dots, X_p$, Δ_2 is an arithmetic mean of conditionally independent and conditionally mean-zero random variables. The same decomposition holds for Δ_3, with i, j interchanged.

For Δ_4, we have

$$\Delta_4 = \frac{\sigma^2}{p}\sum_{1\le k\le p} \Delta_{4,k}, \qquad \Delta_{4,k} \coloneqq \mathbf{E}_{ik}\mathbf{E}_{jk} - \mathbf{I}[i=j].$$

Recalling from the definition of the LMS model that the elements of $\mathbf{E}$ are mean zero, unit-variance, uncorrelated across rows, and independent across columns, we see that Δ_4 is a sum of p mean-zero and mutually independent random variables.

The proof proceeds by using a moment inequality for mixing random variables [Xuejun et al., 2010][Lemma 1.7] to bound $\mathbb{E}[|\Delta_1|^{2q}|Z_i, Z_j]$, and using the Marcinkiewicz–Zygmund inequality to bound $\mathbb{E}[|\Delta_2|^{2q}|Z_i, Z_j, X_1, \dots, X_p]$, $\mathbb{E}[|\Delta_3|^{2q}|Z_i, Z_j, X_1, \dots, X_p]$ and $\mathbb{E}[|\Delta_4|^{2q}]$. These bounds are then combined and Markov's inequality applied. The details are similar to [Gray et al., 2023][Proof of proposition 2], so are omitted.

□

Proof of proposition 2. As explained above the statement of proposition 2, we only need to show that **A2** holds if and only if ϕ is one-to-one. For any $z, z' \in \mathcal{Z}$ consider the identities:

$$\begin{aligned}
\|\phi(z) - \phi(z')\|_{\ell_2}^2 &= \|\phi(z)\|_{\ell_2}^2 + \|\phi(z')\|_{\ell_2}^2 - 2\langle\phi(z), \phi(z')\rangle_{\ell_2} \\
&= f(z,z) + f(z',z') - 2f(z,z') \\
&= \frac{1}{p}\sum_{j=1}^{p}\mathbb{E}[|X_j(z)|^2] + \frac{1}{p}\sum_{j=1}^{p}\mathbb{E}[|X_j(z')|^2] - \frac{2}{p}\sum_{j=1}^{p}\mathbb{E}[X_j(z)X_j(z')] \\
&= \frac{1}{p}\sum_{j=1}^{p}\mathbb{E}[|X_j(z) - X_j(z')|^2].
\end{aligned}$$

Hence $\phi(z) \neq \phi(z')$ if and only if $\sum_{j=1}^{p}\mathbb{E}[|X_j(z) - X_j(z')|^2] > 0$, showing that **A2** is equivalent to ϕ being one-to-one, as required to complete the proof of the proposition.

To go further let us also show that ϕ being one-to-one is equivalent to the condition:

$$\text{for any } z, z' \in \mathcal{Z} \text{ such that } z \neq z', \text{ there exists } \xi \text{ such that } f(z,\xi) \neq f(z',\xi). \tag{32}$$

We first show that (32) implies ϕ is one-to-one. We prove the contrapositive to this statement. So suppose that ϕ is not one-to-one. Then there must exist $z \neq z' \in \mathcal{Z}$ such that $\phi(z) = \phi(z')$. This implies that for any ξ in $\mathcal{Z}$, $f(z,\xi) = \langle\phi(z), \phi(\xi)\rangle_{\ell_2} = \langle\phi(z'), \phi(\xi)\rangle_{\ell_2} = f(z',\xi)$, which is the converse of (32).

In the other direction, suppose the converse of (32) holds, i.e., the exists $z \neq z'$ such that $f(z,\xi) = f(z',\xi)$ for all ξ. By considering the cases $\xi = z$ and $\xi = z'$ we find $f(z,z) = f(z,z') = f(z',z')$. In turn,

$$\|\phi(z) - \phi(z')\|_{\ell_2}^2 = f(z,z) + f(z',z') - 2f(z,z') = 0,$$

i.e., $\phi(z) = \phi(z')$, and hence ϕ is not one-to-one. □

C.1 Proofs and supporting material for section 3.3

The purpose of this section is to state some definitions and intermediate results, building towards the proofs of propositions 3 and 4. Recall that the term "continuous path" was used in **A3**. From henceforth we just say "path" for short.

The following definitions are standard in metric geometry [Burago, 2001]. For $x, x' \in \mathcal{M}$, a *path* in $\mathcal{M}$ with end-points x, x' is a continuous function $\gamma : [0,1] \to \mathcal{M}$ such that $\gamma_0 = x$ and $\gamma_1 = x'$, where $\mathcal{M}$ is equipped with the distance $\|\cdot - \cdot\|_{\ell_2}$. With $n \geq 1$, a non-decreasing sequence $t_0, \ldots, t_n$ such that $t_0 = 0$ and $t_n = 1$, is called a *partition*. Given a path γ and a partition $\mathcal{P} = (t_0, \ldots, t_n)$, define $\chi(\gamma, \mathcal{P}) \coloneqq \sum_{k=1}^{n} \|\gamma_{t_k} - \gamma_{t_{k-1}}\|_{\ell_2}$. The *length* of γ is $L(\gamma) \coloneqq \sup_{\mathcal{P}} \chi(\gamma, \mathcal{P})$, where the supremum is over all possible partitions.

When $\mathcal{Z}$ is a subset of $\mathbb{R}^d$, a *path* η in $\mathcal{Z}$ with end-points z, z' is a continuous function $\eta : [0,1] \to \mathcal{Z}$ such that $\eta_0 = z, \eta_1 = z'$, and with $\chi(\eta, \mathcal{P}) \coloneqq \sum_{k=1}^{n} \|\eta_{t_k} - \eta_{t_{k-1}}\|_{\mathbb{R}^d}$ the length of η is $L(\eta) \coloneqq \sup_{\mathcal{P}} \chi(\eta, \mathcal{P})$.

The *shortest path lengths*, also known as *geodesic distances*, in $\mathcal{M}$ and $\mathcal{Z}$ are:

$$d^{\text{geo}}_{\mathcal{M}}(x, x') \coloneqq \inf_{\gamma : \gamma_0 = x, \gamma_1 = x'} L(\gamma) \qquad d^{\text{geo}}_{\mathcal{Z}}(z, z') \coloneqq \inf_{\eta : \eta_0 = z, \eta_1 = z'} L(\eta), \tag{33}$$

where the infima are over all paths in respectively $\mathcal{M}$ and $\mathcal{Z}$ with the indicated end-points.

A8. *Assume **A3** holds and with d as therein, additionally assume there exists a closed ball $\widetilde{\mathcal{Z}} \subset \mathbb{R}^d$ centered on the origin such that: $\mathcal{Z} \subset \widetilde{\mathcal{Z}}$; the definition of $f(z, z')$ can be extended from $\mathcal{Z} \times \mathcal{Z}$ to $\widetilde{\mathcal{Z}} \times \widetilde{\mathcal{Z}}$; f is C^2 on $\widetilde{\mathcal{Z}} \times \widetilde{\mathcal{Z}}$ and the matrix $\mathbf{H}_\xi \in \mathbb{R}^{d \times d}$ with elements:*

$$(\mathbf{H}_\xi)_{ij} \coloneqq \left. \frac{\partial^2 f}{\partial z_i \partial z'_j} \right|_{(\xi, \xi)}$$

is positive-definite for all $\xi \in \mathcal{Z}$.

The statement of the following theorem, from [Whiteley et al., 2021], is paraphrased slightly in order to match the assumptions of interest here.

Theorem 3 (Whiteley et al. [2021], Thm 1.)**.** *Assume **A1**, **A2** and **A8**. Then ϕ is a bi-Lipschitz homeomorphism between $\mathcal{Z}$ and $\mathcal{M}$. Let x, x' be any two points in $\mathcal{M}$, and let γ be any path in $\mathcal{M}$ of finite length, with end-points x, x'. Define $\eta : [0,1] \to \mathcal{Z}$ by $\eta_t \coloneqq \phi^{-1}(\gamma_t)$. Then η is a path in $\mathcal{Z}$ with $L(\eta) < \infty$. For any $\epsilon > 0$ there exists a partition $\mathcal{P}_\epsilon$ such that for any partition $\mathcal{P} = (t_0, \ldots, t_n)$ satisfying $\mathcal{P}_\epsilon \subseteq \mathcal{P}$,*

$$\left| L(\gamma) - \sum_{k=1}^{n} \left\langle \eta_{t_k} - \eta_{t_{k-1}}, \mathbf{H}_{\eta_{t_{k-1}}} (\eta_{t_k} - \eta_{t_{k-1}}) \right\rangle^{1/2} \right| \leq \epsilon. \tag{34}$$

Proof of proposition 3. Under the assumptions of the proposition, by direct calculation $\mathbf{H}_\xi = -2g'(0)\mathbf{I}_d$ for all $\xi \in \mathcal{Z}$ and **A8** holds.

Fix any $z, z' \in \mathcal{Z}$ and let η be any finite length path in $\mathcal{Z}$ with these end-points. By theorem 3, ϕ is Lipschitz, so γ defined by $\gamma_t \coloneqq \phi(\eta_t)$ has finite length. Define $x \coloneqq \phi(z)$, $x' \coloneqq \phi(z')$. Applying theorem 3, we have from (34) that for any $\epsilon > 0$ there exists a partition $\mathcal{P}_\epsilon$ such that for any $\mathcal{P} = (t_0, \ldots, t_n)$ satisfying $\mathcal{P}_\epsilon \subseteq \mathcal{P}$,

$$\left| L(\gamma) - \sqrt{-2g'(0)}\chi(\eta, \mathcal{P}) \right| \leq \epsilon. \tag{35}$$

Also, using the definition of path length $L(\eta)$ and the triangle inequality, there exists $\widetilde{\mathcal{P}}_\epsilon$ such that for any partition $\mathcal{P}$ satisfying $\widetilde{\mathcal{P}}_\epsilon \subseteq \mathcal{P}$, we have:

$$|L(\eta) - \chi(\eta, \mathcal{P})| \leq \epsilon. \tag{36}$$

Choosing $\mathcal{P}$ to be the union of $\mathcal{P}_\epsilon$ and $\widetilde{\mathcal{P}}_\epsilon$, i.e., if $\tau \in \mathcal{P}_\epsilon$ or $\widetilde{\mathcal{P}}_\epsilon$, then $\tau \in \mathcal{P}$, we find that (35) and (36) are satisfied simultaneously. Since ϵ was arbitrarily small, we find that $L(\gamma) = \sqrt{-2g'(0)}L(\eta)$.

By theorem 3, ϕ is a bi-Lipschitz homeomorphism, so $\tilde{\gamma}_t = \phi(\tilde{\eta}_t)$ defines a bijection between the set of finite-length paths $\tilde{\gamma}$ in $\mathcal{Z}$ with end-points $\phi(z), \phi(z')$ and the set of finite length paths $\tilde{\eta}$ in $\mathcal{Z}$ with end-points z, z'. Therefore by taking the infimum over η on both sides of $L(\gamma) = \sqrt{-2g'(0)}L(\eta)$ where γ is defined by $\gamma_t = \phi(\eta_t)$ as above, we find that

$$d^{\text{geo}}_{\mathcal{M}}(\phi(z), \phi(z')) = \sqrt{-2g'(0)}d^{\text{geo}}_{\mathcal{Z}}(z, z') \tag{37}$$

as required. □

Proof of proposition 4. For the $\mathcal{Z}$ in question, we have $g(\langle z, z'\rangle_{\mathbb{R}^d}) = g(1 - \|z - z'\|^2_{\mathbb{R}^d}/2)$. The proof is completed by applying proposition 3 and using the chain rule of differentiation. □

D Proof and supporting results for theorem 1

Theorem 1 is a corollary to theorem 4. The proofs of both these theorems are in section D.2. Section D.1 contains definitions and notation used throughout section D. Various intermediate results used in the proof of theorem 4 are given in sections D.3-D.6.

The following assumption is a more detailed version of **A5**.

A9. *For some* $q \geq 1$, $\sup_{j\geq 1}\sup_{z\in\mathcal{Z}}\mathbb{E}[|X_j(z)|^{4q}] < \infty$ *and* $\sup_{j\geq 1}\sup_{i\geq 1}\mathbb{E}[|\mathbf{E}_{ij}|^{4q}] < \infty$.

Theorem 4. *Assume* ***A1****,* ***A4****,* ***A6*** *and* ***A9****, and let* $q \geq 1$ *and* $r < \infty$ *be as therein. For* $\min(p, n) \geq r$*, let* $\mathbf{Y} \in \mathbb{R}^{n\times p}$ *follow the LMS model from section 2 and let* $\zeta_1, \ldots, \zeta_n$ *be the dimension-r PCA embedding. Then there exists a random orthogonal matrix* $\mathbf{Q} \in \mathbb{R}^{r\times r}$ *depending on* n *and* p *such that for any* $\delta \in (0, 1)$ *and* $\epsilon \in (0, 1]$*, if*

$$n \geq c_1\sigma^2 r^{1/2}\left(1 \vee \frac{\sigma^2 r^{1/2}}{\epsilon^2}\right) \vee \log\left(\frac{r}{\delta}\right) \quad \text{and} \quad \frac{p}{n} \geq c_2(q)\frac{r}{\delta^{1/q}\epsilon^2},$$

then

$$\max_{i=1,\ldots,n}\left\|p^{-1/2}\mathbf{Q}\zeta_i - \phi(Z_i)\right\|_2 \leq \epsilon$$

with probability at least $1 - \delta$*. Here* c_1 *and* $c_2(q)$ *are constants depending on the suprema in* ***A9*** *and the quantity* $\inf_{p\geq 1}\lambda^f_r$ *which is strictly positive under* ***A6****; and* $\|\cdot\|_2$ *is the Euclidean norm.*

D.1 Definitions and preliminaries

Throughout section D the probability measure μ in the LMS model is considered fixed, $(\lambda^f_k, u^f_k)_{k\geq 1}$ are as in section 2, and assumption **A6** is taken to hold, so that the rank of f is finite, i.e., $r < \infty$.

D.1.1 Notation concerning vectors and matrices in general

We notationally index the eigenvalues of a generic symmetric matrix $\mathbf{A}$ in a non-increasing but otherwise arbitrary order $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \cdots$. For a vector x with elements x_i, $\|x\|_\infty := \max_i |x_i|$ and $\|x\|_2 := \sqrt{\sum_i |x_i|^2}$, and the spectral norm and Frobenius norm of matrices are denoted $\|\cdot\|_2$ and $\|\cdot\|_F$.

D.1.2 Some matrices of interest

Let the matrix $\mathbf{\Phi} \in \mathbb{R}^{n\times r}$ be defined by

$$\mathbf{\Phi} := [\phi(Z_1)|\cdots|\phi(Z_n)]^\top,$$

Let $\mathbf{\Lambda_Y} \in \mathbb{R}^{r\times r}$ be the diagonal matrix with diagonal elements the eigenvalues $\lambda_1(p^{-1}\mathbf{YY}^\top)$ $\ldots, \lambda_r(p^{-1}\mathbf{YY}^\top)$, and let $\mathbf{U_Y} \in \mathbb{R}^{n\times r}$ have as its columns orthonormal eigenvectors associated with these eigenvalues. Since $\mathbf{\Phi} \in \mathbb{R}^{n\times r}$ and $r \leq \min(p, n)$, the matrix $\mathbf{\Phi\Phi}^\top$ has rank at most r. Let $\mathbf{\Lambda_\Phi} \in \mathbb{R}^{r\times r}$ be the diagonal matrix with diagonal elements which are the eigenvalues $\lambda_1(\mathbf{\Phi\Phi}^\top), \ldots, \lambda_r(\mathbf{\Phi\Phi}^\top)$, and let $\mathbf{U_\Phi} \in \mathbb{R}^{n\times r}$ have as its columns orthonormal eigenvectors associated with these eigenvalues. Let $\mathbf{F}_1\mathbf{\Sigma}\mathbf{F}_2^\top$ denote the full singular value decomposition of $\mathbf{U}_\mathbf{\Phi}^\top\mathbf{U_Y}$ and define the random orthogonal matrix $\mathbf{F}_\star := \mathbf{F}_1\mathbf{F}_2^\top$.

D.1.3 Some events of interest

With U_j denoting the jth column of $\mathbf{U}_{\mathbf{\Phi}}$, define:

$$
\begin{aligned}
A_1(\epsilon) &\coloneqq \left\{\|p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top - \sigma^2\mathbf{I}_n\|_2 \le \epsilon n\right\} \\
A_2(\epsilon) &\coloneqq \bigcap_{i=1}^{n} B_{\mathbf{Y},i}(\epsilon) \cap \bigcap_{i=1}^{r} B_{\mathbf{\Phi},i}(\epsilon) \\
A_3(\epsilon) &\coloneqq \left\{\max_{j=1,\dots,r} \|(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top - \sigma^2\mathbf{I}_n)U_j\|_\infty \le \epsilon n^{1/2}\right\} \\
A_{\text{rank}} &\coloneqq \left\{\text{rank}(\mathbf{Y}\mathbf{Y}^\top) \ge r\right\} \cap \left\{\text{rank}(\mathbf{\Phi}\mathbf{\Phi}^\top) = r\right\} \\
B_{\mathbf{Y},i}(\epsilon) &\coloneqq \begin{cases} \left\{\lambda_i^f(1-\epsilon) \le \frac{1}{n}\lambda_i(p^{-1}\mathbf{Y}\mathbf{Y}^\top) \le \lambda_i^f(1+\epsilon)\right\}, & 1 \le i \le r, \\ \left\{\frac{1}{n}\lambda_i(p^{-1}\mathbf{Y}\mathbf{Y}^\top) \le \epsilon\lambda_r^f\right\}, & r+1 \le i \le n. \end{cases} \\
B_{\mathbf{\Phi},i}(\epsilon) &\coloneqq \left\{(1-\epsilon)\lambda_i^f \le \frac{1}{n}\lambda_i(\mathbf{\Phi}\mathbf{\Phi}^\top) \le (1+\epsilon)\lambda_i^f\right\}, \qquad 1 \le i \le r.
\end{aligned}
$$

D.2 Proofs of theorems 4 and 1

Proof of theorem 4. Let $\mathbf{F}_1\mathbf{\Sigma}\mathbf{F}_2^\top$ be the full singular value decomposition of $\mathbf{U}_{\mathbf{\Phi}}^\top\mathbf{U}_{\mathbf{Y}}$ and define the random orthogonal matrix $\mathbf{F}_\star \coloneqq \mathbf{F}_1\mathbf{F}_2^\top$. On the event A_{rank} we have $\mathbf{U}_{\mathbf{\Phi}}\mathbf{\Lambda}_{\mathbf{\Phi}}\mathbf{U}_{\mathbf{\Phi}}^\top = \mathbf{\Phi}\mathbf{\Phi}^\top$, and applying lemma 4 we find there exists a random orthogonal matrix $\widehat{\mathbf{Q}}$ such that $\mathbf{U}_{\mathbf{\Phi}}\mathbf{\Lambda}_{\mathbf{\Phi}}^{1/2} = \mathbf{\Phi}\widehat{\mathbf{Q}}$, hence $[\mathbf{U}_{\mathbf{\Phi}}\mathbf{\Lambda}_{\mathbf{\Phi}}^{1/2}\mathbf{F}_\star]_i = \phi(Z_i)^\top\mathbf{Q}$ for all $i = 1,\dots n$, where $\mathbf{Q} \coloneqq \widehat{\mathbf{Q}}\mathbf{F}_\star$ is orthogonal and $[\cdot]_i$ denotes the ith row of a matrix. Lemma 5 shows that $[\mathbf{U}_{\mathbf{Y}}\mathbf{\Lambda}_{\mathbf{Y}}^{1/2}]_i = p^{-1/2}\zeta_i$. Combining these observations we have shown that on the event A_{rank},

$$
\|p^{-1/2}\mathbf{Q}\zeta_i - \phi(Z_i)\|_2 = \|[\mathbf{U}_{\mathbf{Y}}\mathbf{\Lambda}_{\mathbf{Y}}^{1/2} - \mathbf{U}_{\mathbf{\Phi}}\mathbf{\Lambda}_{\mathbf{\Phi}}^{1/2}\mathbf{F}_\star]_i\|_2, \qquad i = 1,\dots,n. \tag{38}
$$

Now fix any $\epsilon_1 > 0$, $\epsilon_2 \in (0, 1/2)$ and $\epsilon_3 > 0$. Note that the event A_{rank} is a superset of $A_2(\epsilon_2)$ and thus $A_1(\epsilon_1)\cap A_2(\epsilon_2)\cap A_3(\epsilon_3) \subseteq A_{\text{rank}}$. Throughout the remainder of the proof of theorem 4 we shall establish various identities and inequalities involving random variables, random matrices, etc; all such identifies and inequalities to be understood as holding on the event $A_1(\epsilon_1)\cap A_2(\epsilon_2)\cap A_3(\epsilon_3)$, although we shall avoid making this explicit in our notation in order to avoid repetition. For example, for two random matrices say $\mathbf{A}$ and $\mathbf{B}$, we write "$\mathbf{A} = \mathbf{B}$" as shorthand for "$\mathbf{A}(\omega) = \mathbf{B}(\omega)$ for all $\omega \in A_1(\epsilon_1) \cap A_2(\epsilon_2) \cap A_3(\epsilon_3)$" and similarly for two random variables say X, Y, we write "$X \le Y$" as shorthand for "$X(\omega) \le Y(\omega)$ for all $\omega \in A_1(\epsilon_1) \cap A_2(\epsilon_2) \cap A_3(\epsilon_3)$".

Noting that on the event A_{rank}, the matrices $\mathbf{\Lambda}_{\mathbf{Y}}^{-1/2}$ and $\mathbf{\Lambda}_{\mathbf{\Phi}}^{-1/2}$ are well-defined, let us introduce:

$$
\begin{aligned}
\mathbf{C}_1 &\coloneqq \mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}}^{1/2} - \mathbf{\Lambda}_{\mathbf{\Phi}}^{1/2}\mathbf{F}_\star \\
\mathbf{C}_2 &\coloneqq (\mathbf{U}_{\mathbf{\Phi}}^\top\mathbf{U}_{\mathbf{Y}} - \mathbf{F}_\star)\mathbf{\Lambda}_{\mathbf{Y}}^{1/2} \\
\mathbf{C}_3 &\coloneqq \mathbf{U}_{\mathbf{Y}} - \mathbf{U}_{\mathbf{\Phi}}\mathbf{F}_\star = \mathbf{U}_{\mathbf{Y}} - \mathbf{U}_{\mathbf{\Phi}}\mathbf{U}_{\mathbf{\Phi}}^\top\mathbf{U}_{\mathbf{Y}} + \mathbf{U}_{\mathbf{\Phi}}(\mathbf{U}_{\mathbf{\Phi}}^\top\mathbf{U}_{\mathbf{Y}} - \mathbf{F}_\star) \\
\mathbf{D}_1 &\coloneqq \mathbf{U}_{\mathbf{\Phi}}\mathbf{C}_1 \\
\mathbf{D}_2 &\coloneqq \mathbf{U}_{\mathbf{\Phi}}\mathbf{C}_2 \\
\mathbf{D}_3 &\coloneqq (\mathbf{I} - \mathbf{U}_{\mathbf{\Phi}}\mathbf{U}_{\mathbf{\Phi}}^\top)(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{C}_3\mathbf{\Lambda}_{\mathbf{Y}}^{-1/2} \\
\mathbf{D}_4 &\coloneqq -\mathbf{U}_{\mathbf{\Phi}}\mathbf{U}_{\mathbf{\Phi}}^\top(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{U}_{\mathbf{\Phi}}\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}}^{-1/2} \\
\mathbf{D}_5 &\coloneqq (p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{U}_{\mathbf{\Phi}}(\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}}^{-1/2} - \mathbf{\Lambda}_{\mathbf{\Phi}}^{-1/2}\mathbf{F}_\star)
\end{aligned}
$$

We now claim that:

$$
\mathbf{U}_{\mathbf{Y}}\mathbf{\Lambda}_{\mathbf{Y}}^{1/2} - \mathbf{U}_{\mathbf{\Phi}}\mathbf{\Lambda}_{\mathbf{\Phi}}^{1/2}\mathbf{F}_\star = (p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{U}_{\mathbf{\Phi}}\mathbf{\Lambda}_{\mathbf{\Phi}}^{-1/2}\mathbf{F}_\star + \sum_{i=1}^{5}\mathbf{D}_i, \tag{39}
$$

which up to some notational differences, is the same decomposition used by Lyzinski et al. [2016, Proof of Thm 18.] in the analysis of spectral methods for community detection in graphs. To verify the decomposition (39), observe:

$$
\begin{aligned}
\mathbf{U}_{\mathbf{Y}}\mathbf{\Lambda}_{\mathbf{Y}}^{1/2} - \mathbf{U}_{\mathbf{\Phi}}\mathbf{\Lambda}_{\mathbf{\Phi}}^{1/2}\mathbf{F}_\star &= \mathbf{U}_{\mathbf{Y}}\mathbf{\Lambda}_{\mathbf{Y}}^{1/2} - \mathbf{U}_{\mathbf{\Phi}}\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}}^{1/2} \\
&\quad + \mathbf{U}_{\mathbf{\Phi}}\mathbf{C}_1 \\
&= (\mathbf{I}_n - \mathbf{U}_{\mathbf{\Phi}}\mathbf{U}_{\mathbf{\Phi}}^\top)\mathbf{U}_{\mathbf{Y}}\mathbf{\Lambda}_{\mathbf{Y}}^{1/2} \\
&\quad + \mathbf{U}_{\mathbf{\Phi}}\mathbf{C}_2 \\
&\quad + \mathbf{U}_{\mathbf{\Phi}}\mathbf{C}_1 \\
&= (\mathbf{I}_n - \mathbf{U}_{\mathbf{\Phi}}\mathbf{U}_{\mathbf{\Phi}}^\top)(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{U}_{\mathbf{Y}}\mathbf{\Lambda}_{\mathbf{Y}}^{-1/2} \qquad (40) \\
&\quad + \mathbf{U}_{\mathbf{\Phi}}\mathbf{C}_2 \\
&\quad + \mathbf{U}_{\mathbf{\Phi}}\mathbf{C}_1 \\
&= (p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{U}_{\mathbf{\Phi}}\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}}^{-1/2} \\
&\quad - \mathbf{U}_{\mathbf{\Phi}}\mathbf{U}_{\mathbf{\Phi}}^\top(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{U}_{\mathbf{\Phi}}\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}}^{-1/2} \\
&\quad + (\mathbf{I}_n - \mathbf{U}_{\mathbf{\Phi}}\mathbf{U}_{\mathbf{\Phi}}^\top)(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{C}_3\mathbf{\Lambda}_{\mathbf{Y}}^{-1/2} \\
&\quad + \mathbf{U}_{\mathbf{\Phi}}\mathbf{C}_2 \\
&\quad + \mathbf{U}_{\mathbf{\Phi}}\mathbf{C}_1 \\
&= (p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{U}_{\mathbf{\Phi}}\mathbf{\Lambda}_{\mathbf{\Phi}}^{-1/2}\mathbf{F}_\star \\
&\quad + (p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{U}_{\mathbf{\Phi}}(\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}}^{-1/2} - \mathbf{\Lambda}_{\mathbf{\Phi}}^{-1/2}\mathbf{F}_\star) \\
&\quad - \mathbf{U}_{\mathbf{\Phi}}\mathbf{U}_{\mathbf{\Phi}}^\top(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{U}_{\mathbf{\Phi}}\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}}^{-1/2} \\
&\quad + (\mathbf{I}_n - \mathbf{U}_{\mathbf{\Phi}}\mathbf{U}_{\mathbf{\Phi}}^\top)(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{C}_3\mathbf{\Lambda}_{\mathbf{Y}}^{-1/2} \\
&\quad + \mathbf{U}_{\mathbf{\Phi}}\mathbf{C}_2 \\
&\quad + \mathbf{U}_{\mathbf{\Phi}}\mathbf{C}_1 \\
&= (p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{U}_{\mathbf{\Phi}}\mathbf{\Lambda}_{\mathbf{\Phi}}^{-1/2}\mathbf{F}_\star + \mathbf{D}_5 + \mathbf{D}_4 + \mathbf{D}_3 + \mathbf{D}_2 + \mathbf{D}_1 \qquad (41)
\end{aligned}
$$

where (40) holds because $\mathbf{U}_{\mathbf{Y}}\mathbf{\Lambda}_{\mathbf{Y}}^{1/2} = p^{-1}\mathbf{Y}^\top\mathbf{Y}\mathbf{U}_{\mathbf{Y}}\mathbf{\Lambda}_{\mathbf{Y}}^{-1/2}$ and $\mathbf{U}_{\mathbf{\Phi}}\mathbf{U}_{\mathbf{\Phi}}^\top\mathbf{\Phi}\mathbf{\Phi}^\top = \mathbf{\Phi}\mathbf{\Phi}^\top$.

The proof proceeds by bounding the Frobenius norm of each matrix $\mathbf{D}_i$, $i = 1, \ldots, 5$. Using lemma 2,

$$
\begin{aligned}
\|\mathbf{D}_1\|_F &= \|\mathbf{C}_1\|_F \\
&\leq \frac{r^{1/2}}{2n^{1/2}(1-\epsilon_2)^{1/2}(\lambda_r^f)^{1/2}}\left[n\frac{(\epsilon_1 + n^{-1}\sigma^2)^2}{\lambda_r^f(1-2\epsilon_2)}\left(1 + 2\frac{\lambda_1^f}{\lambda_r^f}\left(\frac{1+\epsilon_2}{1-2\epsilon_2}\right)\right) + n\epsilon_1 + \sigma^2\right] \\
&= \frac{r^{1/2}n^{1/2}(\epsilon_1 + n^{-1}\sigma^2)}{2(1-\epsilon_2)^{1/2}(\lambda_r^f)^{1/2}}\left[\frac{(\epsilon_1 + n^{-1}\sigma^2)}{\lambda_r^f(1-2\epsilon_2)}\left(1 + 2\frac{\lambda_1^f}{\lambda_r^f}\left(\frac{1+\epsilon_2}{1-2\epsilon_2}\right)\right) + 1\right]. \qquad (42)
\end{aligned}
$$

Using lemma 1,

$$
\begin{aligned}
\|\mathbf{D}_2\|_F &\leq r^{1/2}\|\mathbf{C}_2\|_2 \\
&= r^{1/2}n^{1/2}[\lambda_1^f(1+\epsilon_2)]^{1/2}\left[\frac{\epsilon_1 + n^{-1}\sigma^2}{\lambda_r^f(1-2\epsilon_2)}\right]^2. \qquad (43)
\end{aligned}
$$

Again using lemma 1 and the fact that $\mathbf{U_Y} - \mathbf{U_\Phi}\mathbf{U_\Phi}^\top\mathbf{U_Y} = (\mathbf{U_Y}\mathbf{U_Y}^\top - \mathbf{U_\Phi}\mathbf{U_\Phi}^\top)\mathbf{U_Y}$,

$$\begin{aligned}
\|\mathbf{D}_3\|_F &\leq 2r^{1/2}\|p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top\|_2\|\mathbf{C}_3\|_2\|\mathbf{\Lambda}_\mathbf{Y}^{-1/2}\|_2 \\
&\leq 2r^{1/2}\frac{(\epsilon_1 n + \sigma^2)}{n^{1/2}\left[\lambda_r^f(1-\epsilon_2)\right]^{1/2}}\left(\|\mathbf{U_Y}\mathbf{U_Y}^\top - \mathbf{U_\Phi}\mathbf{U_\Phi}^\top\|_2 + \|\mathbf{U_\Phi}^\top\mathbf{U_Y} - \mathbf{F}_\star\|_2\right) \\
&\leq 2r^{1/2}n^{1/2}\frac{(\epsilon_1 + n^{-1}\sigma^2)^2}{\left[\lambda_r^f(1-\epsilon_2)\right]^{3/2}}\left(1 + \frac{\epsilon_1 + n^{-1}\sigma^2}{\lambda_r^f(1-\epsilon_2)}\right). \qquad (44)
\end{aligned}$$

Directly:

$$\begin{aligned}
\|\mathbf{D}_4\|_F &\leq r^{1/2}\|\mathbf{D}_4\|_2 \\
&\leq r^{1/2}\|p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top\|_2\|\mathbf{\Lambda}_\mathbf{Y}^{-1/2}\|_2 \\
&\leq r^{1/2}\frac{(\epsilon_1 n + \sigma^2)}{n^{1/2}\left[\lambda_r^f(1-\epsilon_2)\right]^{1/2}} \\
&= r^{1/2}n^{1/2}\frac{(\epsilon_1 + n^{-1}\sigma^2)}{\left[\lambda_r^f(1-\epsilon_2)\right]^{1/2}}. \qquad (45)
\end{aligned}$$

Using lemma 2,

$$\begin{aligned}
\|\mathbf{D}_5\|_F &= \|(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{U_\Phi}(\mathbf{F}_\star\mathbf{\Lambda}_\mathbf{Y}^{-1/2} - \mathbf{\Lambda}_\mathbf{\Phi}^{-1/2}\mathbf{F}_\star)\|_F \\
&\leq r^{1/2}\|p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top\|_2\|\mathbf{F}_\star\mathbf{\Lambda}_\mathbf{Y}^{-1/2} - \mathbf{\Lambda}_\mathbf{\Phi}^{-1/2}\mathbf{F}_\star\|_F \\
&\leq r^{1/2}(\epsilon_1 n + \sigma^2)\frac{\|\mathbf{F}_\star\mathbf{\Lambda}_\mathbf{Y} - \mathbf{\Lambda}_\mathbf{\Phi}\mathbf{F}_\star\|_F}{2n^{3/2}(\lambda_r^f)^{3/2}(1-\epsilon_2)^{3/2}} \\
&\leq \frac{rn^2(\epsilon_1 + n^{-1}\sigma^2)}{2n^{3/2}(\lambda_r^f)^{3/2}(1-\epsilon_2)^{3/2}}\left[\frac{(\epsilon_1 + n^{-1}\sigma^2)^2}{\lambda_r^f(1-2\epsilon_2)}\left(1 + 2\frac{\lambda_1^f}{\lambda_r^f}\left(\frac{1+\epsilon_2}{1-2\epsilon_2}\right)\right) + \epsilon_1 + \frac{\sigma^2}{n}\right] \\
&= \frac{rn^{1/2}(\epsilon_1 + n^{-1}\sigma^2)^2}{2(\lambda_r^f)^{3/2}(1-\epsilon_2)^{3/2}}\left[\frac{(\epsilon_1 + n^{-1}\sigma^2)}{\lambda_r^f(1-2\epsilon_2)}\left(1 + 2\frac{\lambda_1^f}{\lambda_r^f}\left(\frac{1+\epsilon_2}{1-2\epsilon_2}\right)\right) + 1\right]. \qquad (46)
\end{aligned}$$

Having obtained the above bounds on $\|\mathbf{D}_i\|_F$, for $i = 1, \ldots, 5$, we turn to the first term on the r.h.s. of (39). Writing $[\cdot]_i$ to indicate the ith row of a matrix,

$$\begin{aligned}
&\max_{i=1,\ldots,n}\|[(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{U_\Phi}\mathbf{\Lambda}_\mathbf{\Phi}^{-1/2}\mathbf{F}_\star]_i\|_2 \qquad (47) \\
&= \max_{i=1,\ldots,n}\|[(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{U_\Phi}\mathbf{\Lambda}_\mathbf{\Phi}^{-1/2}]_i\|_2 \\
&\leq \frac{1}{n^{1/2}(\lambda_r^f)^{1/2}(1-\epsilon_2)^{1/2}}\max_{i=1,\ldots,n}\|[(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{U_\Phi}]_i\|_2 \\
&\leq \frac{r^{1/2}}{n^{1/2}(\lambda_r^f)^{1/2}(1-\epsilon_2)^{1/2}}\max_{j=1,\ldots,r}\|(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)U_j\|_\infty \\
&\leq \frac{r^{1/2}(\epsilon_3 + \sigma^2/n^{1/2})}{(\lambda_r^f)^{1/2}(1-\epsilon_2)^{1/2}}. \qquad (48)
\end{aligned}$$

where U_j is the jth column of $\mathbf{U_\Phi}$.

Recall that at the start of the proof we fixed arbitrary values $\epsilon_1 > 0$, $\epsilon_2 \in (0, 1/2)$ and $\epsilon_3 > 0$. We now need to work with a specific numerical value for ϵ_2, so let us take it to be $1/4$. Elementary manipulations of the bounds (42)-(46) then show that there exists $\tilde{c}_0$ depending only

on the constants $c_\lambda^{\max}, c_\lambda^{\min}$ in lemma 7 such that

$$\|\mathbf{D}_1\|_F \le \tilde{c}_0 r^{1/2} n^{1/2} \left(\epsilon_1 + \frac{\sigma^2}{n}\right)\left(\epsilon_1 + \frac{\sigma^2}{n} + 1\right)$$
$$\|\mathbf{D}_2\|_F \le \tilde{c}_0 r^{1/2} n^{1/2} \left(\epsilon_1 + \frac{\sigma^2}{n}\right)^2$$
$$\|\mathbf{D}_3\|_F \le \tilde{c}_0 r^{1/2} n^{1/2} \left(\epsilon_1 + \frac{\sigma^2}{n}\right)^2\left(\epsilon_1 + \frac{\sigma^2}{n} + 1\right)$$
$$\|\mathbf{D}_4\|_F \le \tilde{c}_0 r^{1/2} n^{1/2} \left(\epsilon_1 + \frac{\sigma^2}{n}\right)$$
$$\|\mathbf{D}_5\|_F \le \tilde{c}_0 r n^{1/2} \left(\epsilon_1 + \frac{\sigma^2}{n}\right)^2\left(\epsilon_1 + \frac{\sigma^2}{n} + 1\right).$$

Now assuming

$$n \ge 2\sigma^2 r^{1/2} \tag{49}$$

i.e, $n^{-1}\sigma^2 r^{1/2} \le 1/2$, and assuming

$$\epsilon_1 r^{1/2} \le 1/2 \tag{50}$$

we have

$$\left(\epsilon_1 + \frac{\sigma^2}{n}\right) r^{1/2} \le 1.$$

Applying this inequality (recalling $r \ge 1$) in the above bounds on $\|\mathbf{D}_i\|_F$, for $i = 1, 2, 3, 5$ and allowing $\tilde{c}_0$ to increase where necessary we obtain:

$$\max_{i=1,\dots,5} \|\mathbf{D}_i\|_F \le \tilde{c}_0 r^{1/2} n^{1/2} \left(\epsilon_1 + \frac{\sigma^2}{n}\right)$$

Combining this estimate with (48) and again allowing $\tilde{c}_0$ to increase as needed,

$$\begin{aligned}\max_{i=1,\dots,n} \|[\mathbf{U}_\mathbf{Y}\mathbf{\Lambda}_\mathbf{Y}^{1/2} - \mathbf{U}_\mathbf{\Phi}\mathbf{\Lambda}_\mathbf{\Phi}^{1/2}\mathbf{F}_\star]_i\|_2 &\le \max_{i=1,\dots,n} \|[(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{U}_\mathbf{\Phi}\mathbf{\Lambda}_\mathbf{\Phi}^{-1/2}\mathbf{F}_\star]_i\|_2 + \sum_{i=1}^5 \|\mathbf{D}_i\|_F \\ &\le r^{1/2}\tilde{c}_0 n^{1/2}\left(\epsilon_1 + \frac{\sigma^2}{n}\right) + r^{1/2}\tilde{c}_0\epsilon_3. \end{aligned} \tag{51}$$

Now fix any $\epsilon \in (0, 1]$ and let us strengthen (49) to

$$n \ge \left(2\sigma^2 r^{1/2}\right) \vee \left(\frac{9}{\epsilon^2}\tilde{c}_0^2 r\sigma^4\right) \tag{52}$$

so that $r^{1/2}\tilde{c}_0 n^{-1/2}\sigma^2 \le \epsilon/3$. Then setting $\epsilon_1 := \epsilon/(3n^{1/2}r^{1/2}\tilde{c}_0)$ (which satisfies (50) since $\tilde{c}_0 \ge 1$), $\epsilon_3 := \epsilon/(3r^{1/2}\tilde{c}_0)$ and recalling that we have already chosen $\epsilon_2 := 1/4$ we have as a consequence of (51),

$$\begin{aligned}&\mathbb{P}\left(\max_{i=1,\dots,n} \|[\mathbf{U}_\mathbf{Y}\mathbf{\Lambda}_\mathbf{Y}^{1/2} - \mathbf{U}_\mathbf{\Phi}\mathbf{\Lambda}_\mathbf{\Phi}^{1/2}\mathbf{F}_\star]_i\|_2 \le \epsilon\right) \\ &\qquad \ge 1 - \mathbb{P}(A_1(\epsilon/[3n^{1/2}r^{1/2}\tilde{c}_0])^c) - \mathbb{P}(A_2(1/4)^c) - \mathbb{P}(A_3(\epsilon/[3r^{1/2}\tilde{c}_0])^c).\end{aligned}$$

Now fix any $\delta \in (0, 1)$. By lemma 9, proposition 7 and lemma 11, there exists constants $\tilde{c}_1(q)$, $\tilde{c}_2$ and $\tilde{c}_3(q)$ (depending only on the constants $c_\lambda^{\max}$, $c_\lambda^{\min}$ from lemma 7 and the constants $c_X(2q)$, $c_E(2q)$ from lemma 9) such that

$$\frac{p}{n} \ge \tilde{c}_1(q)^{1/q}\frac{r}{\delta^{1/q}\epsilon^2} \quad \Rightarrow \quad \mathbb{P}(A_1(\epsilon/[3n^{1/2}r^{1/2}\tilde{c}_0])^c) \le \frac{\delta}{3}.$$
$$n \ge \tilde{c}_2\left[\sigma^2 \vee \log\left(\frac{r}{\delta}\right)\right] \text{ and } p \ge \frac{\tilde{c}_2}{\delta^{1/q}} \quad \Rightarrow \quad \mathbb{P}(A_2(1/4)^c) \le \frac{\delta}{3}.$$
$$\frac{p}{n^{1/q}} \ge \tilde{c}_3(q)^{1/q}\frac{r^{1+1/q}}{\delta^{1/q}\epsilon^2} \quad \Rightarrow \quad \mathbb{P}(A_3(\epsilon/[3r^{1/2}\tilde{c}_0])^c) \le \frac{\delta}{3}.$$

Combining these conditions with (52) and appropriately defining c_1 and c_2 gives the conditions in the statement of the theorem. Recalling (38), the proof is complete. □

Proof of theorem 1. If **A5** holds, then **A9** holds with $q = 1$. We may then apply theorem 4 in the case $q = 1$, and in order for the lower bound conditions on n and p/n in the statement of theorem 4 to be satisfied for some given δ and ϵ, it is sufficient that:

$$n \geq \frac{-\check{c}_1 \log \delta}{\epsilon^2} \qquad \text{and} \qquad \frac{p}{n} \geq \frac{\check{c}_2}{\epsilon^2 \delta}, \tag{53}$$

for suitable constants $\check{c}_1 > 0$ and $\check{c}_2 > 0$ depending on $\sigma, c_1, c_2(q)$ and $\sup_{p \geq 1} r$, noting the latter supremum is finite under **A6**.

To complete the proof we need to show that for any $\delta \in (0, 1)$ there exists $\epsilon_0 > 0$ and $M > 0$ such that if $(1/\sqrt{n} + \sqrt{n/p})^{-1} > M$, then:

$$\mathbb{P}\left[\max_{i=1,\ldots,n} \|p^{-1/2}\mathbf{Q}\zeta_i - \phi(Z_i)\|_2 > \epsilon_0 \left(\frac{1}{\sqrt{n}} + \sqrt{\frac{n}{p}}\right)\right] < \delta. \tag{54}$$

So to proceed, fix any $\delta \in (0, 1)$, define $\epsilon_0 \coloneqq \sqrt{-\check{c}_1 \log \delta} \vee \sqrt{\check{c}_2/\delta}$, $M \coloneqq \epsilon_0$ and $\epsilon \coloneqq \epsilon_0(1/\sqrt{n} + \sqrt{n/p})$.

Assume that $(1/\sqrt{n} + \sqrt{n/p})^{-1} \geq M$ and notice that in this situation $\epsilon \in (0, 1]$, which is a requirement of theorem 4. It follows from the definition of ϵ_0 that:

$$\epsilon_0^2 \geq -\check{c}_1 \log \delta \geq \frac{-\check{c}_1 \log \delta}{\left(1 + n/\sqrt{p}\right)^2} = \frac{-\check{c}_1 \log \delta}{n\left(1/\sqrt{n} + \sqrt{n/p}\right)^2},$$

and rearranging then using the above definition of ϵ gives:

$$n \geq \frac{-\check{c}_1 \log \delta}{\epsilon^2},$$

i.e., the first inequality in (53) holds. Similarly

$$\epsilon_0^2 \geq \frac{\check{c}_2}{\delta} \geq \frac{\check{c}_2}{\left(\sqrt{p}/n + 1\right)^2 \delta} = \frac{\check{c}_2}{\frac{p}{n}\left(1/\sqrt{n} + \sqrt{n/p}\right)^2 \delta}$$

hence

$$\frac{p}{n} \geq \frac{\check{c}_2}{\epsilon^2 \delta},$$

i.e., the second inequality in (53) holds. Thus by theorem 4,

$$\mathbb{P}\left[\max_{i=1,\ldots,n} \|p^{-1/2}\mathbf{Q}\zeta_i - \phi(Z_i)\|_2 > \left(\frac{1}{\sqrt{n}} + \sqrt{\frac{n}{p}}\right)\epsilon_0\right] < \delta.$$

which is (54).

□

D.3 Matrix estimates

Lemma 1. *Assume **A1** and **A6**. Then for any $\epsilon_1 > 0$ and $\epsilon_2 \in (0, 1/2)$, on the event*

$$A_1(\epsilon_1) \cap A_2(\epsilon_2)$$

we have

$$\|\mathbf{U}_\mathbf{Y}\mathbf{U}_\mathbf{Y}^\top - \mathbf{U}_\mathbf{\Phi}^\top\mathbf{U}_\mathbf{\Phi}\|_2 \leq \frac{\epsilon_1 + n^{-1}\sigma^2}{\lambda_r^f(1 - 2\epsilon_2)}$$

and

$$\|\mathbf{U}_\mathbf{\Phi}^\top\mathbf{U}_\mathbf{Y} - \mathbf{F}_\star\|_2 \leq \left[\frac{\epsilon_1 + n^{-1}\sigma^2}{\lambda_r^f(1 - 2\epsilon_2)}\right]^2.$$

Proof. In outline, the proof follows Lyzinski et al. [2016, Proof of Prop. 16], although we work with the spectral rather than Frobenius norm. On the event in the statement we have:

$$|\lambda_r(\mathbf{\Phi}\mathbf{\Phi}^\top) - \lambda_{r+1}(p^{-1}\mathbf{Y}\mathbf{Y}^\top)| \geq n\lambda_r^f(1-2\epsilon_2) > 0$$

and with σ_i denoting the ith singular value of $U_{\mathbf{\Phi}}^\top U_{\mathbf{Y}}$ and $\sigma_i = \cos(\theta_i)$, the Davis-Kahan $\sin(\theta)$ theorem gives:

$$\begin{aligned}\|\mathbf{U}_{\mathbf{Y}}\mathbf{U}_{\mathbf{Y}}^\top - \mathbf{U}_{\mathbf{\Phi}}^\top\mathbf{U}_{\mathbf{\Phi}}\|_2 = \max_i |\sin(\theta_i)| &\leq \frac{\|p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top\|_2}{|\lambda_r(\mathbf{\Phi}\mathbf{\Phi}^\top) - \lambda_{r+1}(p^{-1}\mathbf{Y}\mathbf{Y}^\top)|} \\ &\leq \frac{\epsilon_1 + n^{-1}\sigma^2}{\lambda_r^f(1-2\epsilon_2)}. \end{aligned} \tag{55}$$

Therefore

$$\begin{aligned}\|\mathbf{U}_{\mathbf{\Phi}}^\top\mathbf{U}_{\mathbf{Y}} - \mathbf{F}_\star\|_2 &= \|\mathbf{F}_1\mathbf{\Sigma}\mathbf{F}_2^\top - \mathbf{F}_1\mathbf{F}_2^\top\|_2 \\ &= \|\mathbf{F}_1(\mathbf{\Sigma} - \mathbf{I}_r)\mathbf{F}_2^\top\|_2 \\ &= \|\mathbf{\Sigma} - \mathbf{I}_r\|_2 \\ &= \max_{i=1,\ldots,r} |1-\sigma_i| \\ &\leq \max_{i=1,\ldots,r} |1-\sigma_i^2| = \max_{i=1,\ldots,r} |\sin(\theta_i)|^2 \\ &\leq \left[\frac{\epsilon_1 + n^{-1}\sigma^2}{\lambda_r^f(1-2\epsilon_2)}\right]^2 \end{aligned}$$

where for the first inequality uses $\|\mathbf{U}_{\mathbf{\Phi}}^\top\mathbf{U}_{\mathbf{Y}}\|_2 \leq 1$ and the second inequality is from (55). □

Lemma 2. *Assume **A1** and **A6**. For any $\epsilon_1 > 0$, $\epsilon_2 \in (0, 1/2)$, on the event*

$$A_1(\epsilon_1) \cap A_2(\epsilon_2)$$

we have

$$\|\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}} - \mathbf{\Lambda}_{\mathbf{\Phi}}\mathbf{F}_\star\|_F \leq r^{1/2}\left[n\frac{(\epsilon_1 + n^{-1}\sigma^2)^2}{\lambda_r^f(1-2\epsilon_2)}\left(1 + 2\frac{\lambda_1^f}{\lambda_r^f}\left(\frac{1+\epsilon_2}{1-2\epsilon_2}\right)\right) + n\epsilon_1 + \sigma^2\right],$$

$$\|\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}}^{1/2} - \mathbf{\Lambda}_{\mathbf{\Phi}}^{1/2}\mathbf{F}_\star\|_F \leq \frac{\|\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}} - \mathbf{\Lambda}_{\mathbf{\Phi}}\mathbf{F}_\star\|_F}{2n^{1/2}(1-\epsilon_2)^{1/2}(\lambda_r^f)^{1/2}},$$

$$\|\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}}^{-1/2} - \mathbf{\Lambda}_{\mathbf{\Phi}}^{-1/2}\mathbf{F}_\star\|_F \leq \frac{\|\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}}^{1/2} - \mathbf{\Lambda}_{\mathbf{\Phi}}^{1/2}\mathbf{F}_\star\|_F}{n(1-\epsilon_2)\lambda_r^f}.$$

Proof. Using a decomposition idea from [Lyzinski et al., 2016, proof of lemma 17], with

$$\mathbf{R} \coloneqq \mathbf{U}_{\mathbf{Y}} - \mathbf{U}_{\mathbf{\Phi}}\mathbf{U}_{\mathbf{\Phi}}^\top\mathbf{U}_{\mathbf{Y}},$$

we have

$$\begin{aligned}\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}} - \mathbf{\Lambda}_{\mathbf{\Phi}}\mathbf{F}_\star = (\mathbf{F}_\star - \mathbf{U}_{\mathbf{\Phi}}^\top\mathbf{U}_{\mathbf{Y}})\mathbf{\Lambda}_{\mathbf{Y}} &+ \mathbf{U}_{\mathbf{\Phi}}^\top(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{R} \\ &+ \mathbf{U}_{\mathbf{\Phi}}^\top(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{U}_{\mathbf{\Phi}}\mathbf{U}_{\mathbf{\Phi}}^\top\mathbf{U}_{\mathbf{Y}} \\ &+ \mathbf{\Lambda}_{\mathbf{\Phi}}(\mathbf{U}_{\mathbf{\Phi}}^\top\mathbf{U}_{\mathbf{Y}} - \mathbf{F}_\star)\end{aligned}$$

hence

$$\|\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}} - \mathbf{\Lambda}_{\mathbf{\Phi}}\mathbf{F}_\star\|_2 \leq \|\mathbf{U}_{\mathbf{\Phi}}^\top\mathbf{U}_{\mathbf{Y}} - \mathbf{F}_\star\|_2(\|\mathbf{\Lambda}_{\mathbf{Y}}\|_2 + \|\mathbf{\Lambda}_\Phi\|_2) \tag{56}$$

$$+ \|\mathbf{U}_{\mathbf{\Phi}}^\top(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{R}\|_2 \tag{57}$$

$$+ \|\mathbf{U}_{\mathbf{\Phi}}^\top(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{U}_{\mathbf{\Phi}}\mathbf{U}_{\mathbf{\Phi}}^\top\mathbf{U}_{\mathbf{Y}}\|_2 \tag{58}$$

For the term on the r.h.s. of (56), on the event in the statement of the present lemma and using lemma 1 we have:

$$\|\mathbf{U}_{\mathbf{\Phi}}^\top \mathbf{U}_{\mathbf{Y}} - \mathbf{F}_\star\|_2(\|\mathbf{\Lambda}_{\mathbf{Y}}\|_2 + \|\mathbf{\Lambda}_\Phi\|_2) \leq \left[\frac{\epsilon_1 + n^{-1}\sigma^2}{\lambda_r^f(1-2\epsilon_2)}\right]^2 2n\lambda_1^f(1+\epsilon_2).$$

For the term in (57), using $\mathbf{R} = (\mathbf{U}_{\mathbf{Y}}\mathbf{U}_{\mathbf{Y}}^\top - \mathbf{U}_{\mathbf{\Phi}}^\top\mathbf{U}_{\mathbf{\Phi}})\mathbf{U}_{\mathbf{Y}}$, we have again on the event in the statement of the present lemma and using lemma 1,

$$\begin{aligned}\|\mathbf{U}_{\mathbf{\Phi}}^\top(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{R}\|_2 &\leq \|p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top\|_2\|\mathbf{R}\|_2 \\ &\leq (\|p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top - \sigma^2\mathbf{I}_n\|_2 + \sigma^2)\|\mathbf{U}_{\mathbf{Y}}\mathbf{U}_{\mathbf{Y}}^\top - \mathbf{U}_{\mathbf{\Phi}}^\top\mathbf{U}_{\mathbf{\Phi}}\|_2 \\ &\leq (\epsilon_1 n + \sigma^2)\left(\frac{\epsilon_1 + n^{-1}\sigma^2}{\lambda_r^f(1-2\epsilon_2)}\right) = n\frac{(\epsilon_1 + n^{-1}\sigma^2)^2}{\lambda_r^f(1-2\epsilon_2)}.\end{aligned}$$

For the term in (58),

$$\begin{aligned}\|\mathbf{U}_{\mathbf{\Phi}}^\top(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top)\mathbf{U}_{\mathbf{\Phi}}\mathbf{U}_{\mathbf{\Phi}}^\top\mathbf{U}_{\mathbf{Y}}\|_2 &\leq \left(\|(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top - \sigma^2\mathbf{I}_n)\|_2 + \sigma^2\right)\|\mathbf{U}_{\mathbf{\Phi}}^\top\mathbf{U}_{\mathbf{Y}}\|_2 \\ &\leq n\epsilon_1 + \sigma^2.\end{aligned}$$

The bound on $\|\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}} - \mathbf{\Lambda}_{\mathbf{\Phi}}\mathbf{F}_\star\|_F$ given in the statement holds by combining the above spectral norm bounds.

For the bound on $\|\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}}^{1/2} - \mathbf{\Lambda}_{\mathbf{\Phi}}^{1/2}\mathbf{F}_\star\|_F$ we use the fact that the elements of $\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}}^{1/2} - \mathbf{\Lambda}_{\mathbf{\Phi}}^{1/2}\mathbf{F}_\star$ can be written:

$$\begin{aligned}(\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}}^{1/2} - \mathbf{\Lambda}_{\mathbf{\Phi}}^{1/2}\mathbf{F}_\star)_{ij} &= (\mathbf{F}_\star)_{ij}\lambda_j(p^{-1}\mathbf{Y}\mathbf{Y}^\top)^{1/2} - \lambda_i(\mathbf{\Phi}\mathbf{\Phi}^\top)^{1/2}(\mathbf{F}_\star)_{ij} \\ &= (\mathbf{F}_\star)_{ij}\frac{[\lambda_j(p^{-1}\mathbf{Y}\mathbf{Y}^\top) - \lambda_i(\mathbf{\Phi}\mathbf{\Phi}^\top)]}{\lambda_j(p^{-1}\mathbf{Y}\mathbf{Y}^\top)^{1/2} + \lambda_i(\mathbf{\Phi}\mathbf{\Phi}^\top)^{1/2}}\end{aligned}$$

hence

$$|(\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}}^{1/2} - \mathbf{\Lambda}_{\mathbf{\Phi}}^{1/2}\mathbf{F}_\star)_{ij}| \leq \frac{|(\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}} - \mathbf{\Lambda}_\Phi\mathbf{F}_\star)_{ij}|}{2n^{1/2}(1-\epsilon_2)^{1/2}(\lambda_r^f)^{1/2}},$$

and so

$$\|\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}}^{1/2} - \mathbf{\Lambda}_{\mathbf{\Phi}}^{1/2}\mathbf{F}_\star\|_F \leq \frac{\|\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}} - \mathbf{\Lambda}_\Phi\mathbf{F}_\star\|_F}{2n^{1/2}(1-\epsilon_2)^{1/2}(\lambda_r^f)^{1/2}}.$$

The bound on $\|\mathbf{F}_\star\mathbf{\Lambda}_{\mathbf{Y}}^{-1/2} - \mathbf{\Lambda}_{\mathbf{\Phi}}^{-1/2}\mathbf{F}_\star\|_F$ in the statement is obtained in a similar manner using the fact that for any $a, b > 0$, $a^{-1/2} - b^{-1/2} = (b^{1/2} - a^{1/2})/(a^{1/2}b^{1/2})$. □

D.4 Some linear algebra

Lemma 3. *For any $m_1, m_2 \geq 1$, $\mathbf{A} \in \mathbb{R}^{m_2 \times m_1}$, $q \leq \min\{m_1, m_2\}$ and strictly positive real numbers $\lambda_1, \ldots, \lambda_q$,*
a) there exists $\mathbf{U} \in \mathbb{R}^{m_2 \times q}$ such that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_q$ and $\mathbf{A}\mathbf{A}^\top\mathbf{U} = \mathbf{U}\mathbf{\Lambda}$, if and only if there exists $\mathbf{V} \in \mathbb{R}^{m_1 \times q}$ such that $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_q$ and $\mathbf{A}^\top\mathbf{A}\mathbf{V} = \mathbf{V}\mathbf{\Lambda}$, where $\mathbf{\Lambda} \coloneqq \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$;
b) when $\mathbf{V}$ with the properties stated in part a) exists, a choice of $\mathbf{U}$ which has the properties stated in part a) is $\mathbf{U} = \mathbf{A}\mathbf{V}\mathbf{\Lambda}^{-1/2}$;
c) $\lambda_i(\mathbf{A}^\top\mathbf{A}) = \lambda_i(\mathbf{A}\mathbf{A}^\top)$, for $i = 1, \ldots, \min\{m_1, m_2\}$.
d) the rank of $\mathbf{A}^\top\mathbf{A}$ is equal to that of $\mathbf{A}\mathbf{A}^\top$;

Proof. Assume the existence of $\mathbf{V}$ with the properties stated in part a). Taking $\mathbf{U} \coloneqq \mathbf{A}\mathbf{V}\mathbf{\Lambda}^{-1/2}$ we have

$$\begin{aligned}\mathbf{U}^\top\mathbf{U} &\coloneqq \mathbf{\Lambda}^{-1/2}\mathbf{V}^\top\mathbf{A}^\top\mathbf{A}\mathbf{V}\mathbf{\Lambda}^{-1/2} \\ &= \mathbf{\Lambda}^{-1/2}\mathbf{V}^\top\mathbf{V}\mathbf{\Lambda}\mathbf{\Lambda}^{-1/2} \\ &= \mathbf{\Lambda}^{-1/2}\mathbf{\Lambda}\mathbf{\Lambda}^{-1/2} = \mathbf{I}_q\end{aligned}$$

and

$$\begin{aligned}\mathbf{A}\mathbf{A}^\top\mathbf{U} &= \mathbf{A}\mathbf{A}^\top\mathbf{A}\mathbf{V}\mathbf{\Lambda}^{-1/2} \\ &= \mathbf{A}\mathbf{V}\mathbf{\Lambda}\mathbf{\Lambda}^{-1/2} \\ &= \mathbf{U}\mathbf{\Lambda}.\end{aligned}$$

The implication in the other direction for part a) holds by interchanging $\mathbf{A}^\top$ and $\mathbf{U}$ with respectively $\mathbf{A}$ and $\mathbf{V}$. We have thus proved parts a) and b) of the lemma. Part a) implies that the non-zero eigenvalues of $\mathbf{A}^\top\mathbf{A}$ are equal to those of $\mathbf{A}\mathbf{A}^\top$, which establishes the claim of part c). Part d) follows from part c). □

Lemma 4. *For any $m_1 \leq m_2$ and $\mathbf{A} \in \mathbb{R}^{m_2 \times m_1}$ such that $\mathbf{A}$ has rank m_1, there exists an orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{m_1 \times m_1}$ such that $\mathbf{U}\mathbf{\Lambda}^{1/2} = \mathbf{A}\mathbf{Q}$, where $\mathbf{\Lambda} = \operatorname{diag}\{\lambda_1(\mathbf{A}\mathbf{A}^\top), \cdots, \lambda_{m_1}(\mathbf{A}\mathbf{A}^\top)\}$ and the columns of $\mathbf{U} \in \mathbb{R}^{m_2 \times m_1}$ are orthonormal eigenvectors of $\mathbf{A}\mathbf{A}^\top$ with eigenvalues $\lambda_1(\mathbf{A}\mathbf{A}^\top)$, $\ldots, \lambda_{m_1}(\mathbf{A}\mathbf{A}^\top)$.*

Proof. We have $\mathbf{A}\mathbf{A}^\top = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, hence $\mathbf{U}\mathbf{\Lambda}^{1/2} = \mathbf{A}\mathbf{A}^\top\mathbf{U}\mathbf{\Lambda}^{-1/2}$. Take $\mathbf{Q} \coloneqq \mathbf{A}^\top\mathbf{U}\mathbf{\Lambda}^{-1/2} \in \mathbb{R}^{m_1 \times m_1}$. We then find:

$$\mathbf{Q}^\top\mathbf{Q} = \mathbf{\Lambda}^{-1/2}\mathbf{U}^\top\mathbf{A}\mathbf{A}^\top\mathbf{U}\mathbf{\Lambda}^{-1/2} = \mathbf{\Lambda}^{-1/2}\mathbf{U}^\top\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top\mathbf{U}\mathbf{\Lambda}^{-1/2} = \mathbf{I}_{m_1}$$

and

$$\mathbf{Q}\mathbf{Q}^\top = \mathbf{A}^\top\mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^\top\mathbf{A}. \tag{59}$$

Consider the reduced singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}^\top$ where $\mathbf{V} \in \mathbb{R}^{m_1 \times m_1}$ has orthonormal columns. Substituting into the r.h.s. of (59),

$$\mathbf{Q}\mathbf{Q}^\top = \mathbf{V}\mathbf{\Lambda}^{1/2}\mathbf{U}^\top\mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^\top\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}^\top = \mathbf{V}\mathbf{V}^\top = \mathbf{I}_{m_1}.$$

□

D.5 Some properties of the LMS model

Lemma 5. *On the event that the rank of $\mathbf{Y}^\top\mathbf{Y}$ is at least r, $p^{-1/2}\mathbf{Y}\mathbf{V}_\mathbf{Y} = \mathbf{U}_\mathbf{Y}\mathbf{\Lambda}_\mathbf{Y}^{1/2}$, where the columns of $\mathbf{U}_\mathbf{Y} \in \mathbb{R}^{n \times r}$ are orthonormal eigenvectors of $p^{-1}\mathbf{Y}\mathbf{Y}^\top$ with associated eigenvalues on the diagonal of the diagonal matrix $\mathbf{\Lambda}_\mathbf{Y} \in \mathbb{R}^{r \times r}$.*

Proof. Apply lemma 3, part b). □

Thus by computing the PCA embedding $\zeta_1, \ldots, \zeta_n$ and rescaling by $p^{-1/2}$, we are, in effect, computing the n rows of $\mathbf{U}_\mathbf{Y}\mathbf{\Lambda}_\mathbf{Y}^{1/2}$, where $\mathbf{U}_\mathbf{Y}\mathbf{\Lambda}_\mathbf{Y}^{1/2}(\mathbf{U}_\mathbf{Y}\mathbf{\Lambda}_\mathbf{Y}^{1/2})^\top = \mathbf{U}_\mathbf{Y}\mathbf{\Lambda}_\mathbf{Y}\mathbf{U}_\mathbf{Y}^\top$ is a rank$-r$ approximation to $p^{-1}\mathbf{Y}\mathbf{Y}^\top$.

Lemma 6. *Assume **A1** and **A6**. Then $p^{-1}\mathbb{E}[\mathbf{Y}\mathbf{Y}^\top | Z_1, \ldots, Z_n] = \mathbf{\Phi}\mathbf{\Phi}^\top + \sigma^2\mathbf{I}_n$.*

Proof. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be the matrix with entries $\mathbf{X}_{ij} \coloneqq X_j(Z_i)$. According to the model specification in section 2, $\mathbf{X}$ and $\mathbf{E}$ are independent, and $\mathbb{E}[\mathbf{E}\mathbf{E}^\top] = p\mathbf{I}_n$. Thus:

$$\begin{aligned}\mathbb{E}[\mathbf{Y}\mathbf{Y}^\top | Z_1, \ldots, Z_n] &= \mathbb{E}[\mathbf{X}\mathbf{X}^\top | Z_1, \ldots, Z_n] + \sigma\mathbb{E}[\mathbf{X}\mathbf{E}^\top | Z_1, \ldots, Z_n] \\ &\quad + \sigma\mathbb{E}[\mathbf{E}\mathbf{X}^\top | Z_1, \ldots, Z_n] + \sigma^2\mathbb{E}[\mathbf{E}\mathbf{E}^\top | Z_1, \ldots, Z_n] \\ &= p\mathbf{\Phi}\mathbf{\Phi}^\top + p\sigma^2\mathbf{I}_n.\end{aligned}$$

□

Lemma 7. *Assume **A1**, **A6** and **A9**. Then there exists a constant $c_\lambda^{\max} < \infty$ depending only on the first supremum in **A9**, and a constant $c_\lambda^{\min} > 0$ such that*

$$\sup_{p \geq 1}\left\{\sup_z f(z,z) + \lambda_1^f\right\} \leq c_\lambda^{\max}, \qquad \inf_{p \geq 1}\lambda_r^f \geq c_\lambda^{\min}.$$

Proof. The existence of $c_\lambda^{\min}$ as required is an immediate consequence of **A6**. Using **A9** and Jensen's inequality gives:

$$\sup_z f(z,z) = \sup_z \frac{1}{p}\sum_{j=1}^{p} \mathbb{E}[|X_j(z)|^2] \le \sup_z \frac{1}{p}\sum_{j=1}^{p} \mathbb{E}[|X_j(z)|^{4q}]^{2/4q} < \infty.$$

The existence of $c_\lambda^{\max}$ as required follows from the above inequalities combined with:

$$\lambda_1^f \le \sum_{k=1}^{\infty} \lambda_k^f = \sum_{k=1}^{\infty} \lambda_k^f \mathbb{E}[|u_k^f(Z_1)|^2] = \mathbb{E}[f(Z_1,Z_1)] \le \sup_z f(z,z).$$

□

D.6 Matrix concentration results

The following matrix-valued version of the Bernstein inequality can be found in, e.g., [Tropp, 2015, Thm 1.6.2]

Theorem 5 (Matrix Bernstein inequality)**.** *Let $\mathbf{M}_1, \dots, \mathbf{M}_n$ be independent random matrices with common dimensions $m_1 \times m_2$ satisfying $\mathbb{E}[\mathbf{M}_i] = 0$ and $\|\mathbf{M}_i\|_2 \le L$ for each $1 \le i \le n$ and some constant L. Let $\mathbf{M} := \sum_{i=1}^n \mathbf{M}_i$ and $v(\mathbf{M}) = \max\left\{\|\mathbb{E}[\mathbf{M}\mathbf{M}^\top]\|_2, \|\mathbb{E}[\mathbf{M}^\top\mathbf{M}]\|_2\right\}$. Then for all $t \ge 0$,*

$$\mathbb{P}\left(\|\mathbf{M}\|_2 \ge t\right) \le (m_1+m_2)\exp\left(\frac{-t^2/2}{v(\mathbf{M}) + Lt/3}\right).$$

Lemma 8. *Assume **A1** and **A6**. For any $t \ge 0$,*

$$\mathbb{P}\left(\|n^{-1}\mathbf{\Phi}^\top\mathbf{\Phi} - n^{-1}\mathbb{E}[\mathbf{\Phi}^\top\mathbf{\Phi}]\|_2 \ge t\right) \le 2r\exp\left(\frac{-t^2 n/2}{(c_\lambda^{\max})^2 + c_\lambda^{\max} t/3}\right),$$

where $c_\lambda^{\max}$ is as in lemma 7.

Proof. Apply theorem 5 with $\mathbf{M}_i = \frac{1}{n}\phi(Z_i)\phi(Z_i)^\top - \mathbb{E}[\frac{1}{n}\phi(Z_i)\phi(Z_i)^\top]$,

$$\begin{aligned}
\|\mathbf{M}_i\|_2 &\le \frac{1}{n}\|\phi(Z_i)\phi(Z_i)^\top\|_2 + \frac{1}{n}\|\mathbb{E}[\phi(Z_i)\phi(Z_i)^\top]\|_2 \\
&= \frac{1}{n}\|\phi(Z_i)\|_2^2 + \frac{1}{n}\lambda_1^f \\
&= \frac{1}{n}f(Z_i,Z_i) + \frac{1}{n}\lambda_1^f \\
&\le \frac{1}{n}c_\lambda^{\max} =: L
\end{aligned}$$

and

$$\begin{aligned}
v(\mathbf{M}) &= \left\|\mathbb{E}\left[\left(\sum_i \mathbf{M}_i\right)\left(\sum_i \mathbf{M}_i\right)\right]\right\|_2 \\
&= \left\|\mathbb{E}\left[\sum_i \mathbf{M}_i\mathbf{M}_i\right]\right\|_2 \\
&\le \frac{1}{n}\left\|\mathbb{E}\left[\phi(Z_1)\phi(Z_1)^\top\phi(Z_1)\phi(Z_1)^\top\right]\right\|_2 + \frac{1}{n}\|\mathbb{E}[\phi(Z_1)\phi(Z_1)^\top]^2\|_2 \\
&\le \frac{1}{n}\mathbb{E}\left[\left\|\phi(Z_1)\phi(Z_1)^\top\phi(Z_1)\phi(Z_1)^\top\right\|_2\right] + \frac{1}{n}\|\mathbb{E}[\phi(Z_1)\phi(Z_1)^\top]^2\|_2 \\
&= \frac{1}{n}\mathbb{E}\left[\left\|\phi(Z_1)\phi(Z_1)^\top\right\|_2^2\right] + \frac{1}{n}(\lambda_1^f)^2 \\
&= \frac{1}{n}\mathbb{E}\left[\|\phi(Z_1)\|_2^4\right] + \frac{1}{n}(\lambda_1^f)^2 \le \frac{1}{n}(c_\lambda^{\max})^2.
\end{aligned}$$

□

Lemma 9. *Assume **A1**, **A6** and **A9** with some $q \geq 1$. Then for any $t > 0$,*

$$\mathbb{P}\left(\|p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top - \sigma^2\mathbf{I}_n\|_2 \geq t\right) \leq (16)^q(2q-1)^q \frac{n^{2q}}{t^{2q}}\frac{1}{p^q}\left(c_X(2q)^{1/2q} + \sigma^2 c_E(2q)^{1/2q}\right)^{2q}$$

where

$$c_X(q) \coloneqq \sup_{j\geq 1}\sup_{z\in\mathcal{Z}} \mathbb{E}\left[|X_j(z)|^{2q}\right], \qquad c_E(q) \coloneqq \sup_{j\geq 1}\sup_{i\geq 1}\mathbb{E}\left[|\mathbf{E}_{ij}|^{2q}\right].$$

Proof. Let us write the matrix $\mathbf{Y}$ in terms of its columns $\mathbf{Y} \equiv [Y_1|\cdots|Y_p]$ so that:

$$\mathbf{Y}\mathbf{Y}^\top = \sum_{j=1}^p Y_j Y_j^\top. \tag{60}$$

Observe that under the model of section 2, conditional on $(Z_1,\ldots,Z_n)$ the summands in (60) are independent and as per lemma 6, the conditional expectation of $\mathbf{Y}\mathbf{Y}^\top$ given $Z_1,\ldots,Z_n$ is: $p\mathbf{\Phi}\mathbf{\Phi}^\top + p\sigma^2\mathbf{I}_n$.

The main tool we use from hereon is a direct combination of the matrix Chebyshev inequality [Paulin et al., 2016, Prop. 3.1] and the matrix polynomial Effron-Stein inequality [Paulin et al., 2016, Thm 4.2], applied under the regular conditional distribution of $(Y_1,\ldots,Y_p)$ given $(Z_1,\ldots,Z_n)$. These inequalities taken together tell us that, for any $q \geq 1$, the following holds almost surely:

$$\begin{aligned}
&\mathbb{P}\left(\left\|p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top - \sigma^2\mathbf{I}_n\right\|_2 \geq t \,\middle|\, Z_1,\ldots,Z_n\right)\\
&\leq \frac{1}{t^{2q}}\mathbb{E}\left[\left\|p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top - \sigma^2\mathbf{I}_n\right\|_{S_{2q}}^{2q} \,\middle|\, Z_1,\ldots,Z_n\right]\\
&\leq \frac{2^q(2q-1)^q}{t^{2q}}\mathbb{E}\left[\|\mathbf{\Sigma}\|_{S_q}^q \,\middle|\, Z_1,\ldots,Z_n\right].
\end{aligned}$$

Here $\|\cdot\|_{S_q}$ is the Schatten q-norm and $\mathbf{\Sigma} \in \mathbb{R}^{n\times n}$ is the variance proxy:

$$\mathbf{\Sigma} \coloneqq \frac{1}{2p^2}\sum_{j=1}^p \mathbb{E}\left[\left(Y_jY_j^\top - \tilde{Y}_j\tilde{Y}_j^\top\right)^2 \,\middle|\, Y_j, Z_1,\ldots,Z_n\right], \tag{61}$$

where, conditional on $Z_1,\ldots,Z_n$, $\tilde{Y}_j$ is an independent copy of Y_j. For brevity in the remainder of the proof we shall write $Z \equiv (Z_1,\ldots,Z_n)$, and to avoid repetitive statements of "almost surely", every inequality involving conditional expectations is to be understood as holding in the almost sure sense.

We estimate:

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{\Sigma}\|_{S_q}^q \,\middle|\, Z\right]^{1/q} &= \frac{1}{2p^2}\mathbb{E}\left[\left\|\sum_{j=1}^p \mathbb{E}\left[\left(Y_jY_j^\top - \tilde{Y}_j\tilde{Y}_j^\top\right)^2 \,\middle|\, Y_j, Z\right]\right\|_{S_q}^q \,\middle|\, Z\right]^{1/q}\\
&\leq \frac{1}{2p^2}\sum_{j=1}^p \mathbb{E}\left[\left\|\mathbb{E}\left[\left(Y_jY_j^\top - \tilde{Y}_j\tilde{Y}_j^\top\right)^2 \,\middle|\, Y_j, Z\right]\right\|_{S_q}^q \,\middle|\, Z\right]^{1/q} \qquad (62)\\
&\leq \frac{1}{2p^2}\sum_{j=1}^p \mathbb{E}\left[\left\|\left(Y_jY_j^\top - \tilde{Y}_j\tilde{Y}_j^\top\right)^2\right\|_{S_q}^q \,\middle|\, Z\right]^{1/q} \qquad (63)\\
&= \frac{1}{2p^2}\sum_{j=1}^p \mathbb{E}\left[\left\|Y_jY_j^\top - \tilde{Y}_j\tilde{Y}_j^\top\right\|_{S_{2q}}^{2q} \,\middle|\, Z\right]^{1/q}\\
&\leq \frac{1}{2p^2}\sum_{j=1}^p \left(2\mathbb{E}\left[\|Y_jY_j^\top\|_{S_{2q}}^{2q} \,\middle|\, Z\right]^{1/2q}\right)^2 \qquad (64)\\
&= \frac{2}{p^2}\sum_{j=1}^p \mathbb{E}\left[\|Y_jY_j^\top\|_{S_{2q}}^{2q} \,\middle|\, Z\right]^{1/q}
\end{aligned}$$

Here (62) holds by the second claim of lemma 10; 63 holds by first claim of lemma 10 combined with the fact that $x \mapsto x^q$ is convex for $x \geq 0$ (recall $q \geq 1$); (64) holds by lemma 10 and the fact that $\tilde{Y}_j$ and Y_j are equal in distribution.

By definition of the Schatten-q norm, $\left\|Y_j Y_j^\top\right\|_{S_{2q}}^{2q} = \sum_{k=1}^n \lambda_k^{2q}\left(Y_j Y_j^\top\right)$, where $\lambda_1\left(Y_j Y_j^\top\right) = \|Y_j\|_2^2$ and $\lambda_k\left(Y_1 Y_1^\top\right) = 0$ for $k = 2, \ldots, n$. Thus:

$$\left\|Y_j Y_j^\top\right\|_{S_{2q}}^{2q} = \|Y_j\|_2^{4q} = \left|\sum_{i=1}^n \left(X_j(Z_i) + \sigma \mathbf{E}_{ij}\right)^2\right|^{2q}. \tag{65}$$

By two applications of Minkowski's inequality,

$$\begin{aligned}
\mathbb{E}\left[\left\|Y_j Y_j^\top\right\|_{S_{2q}}^{2q} \middle| Z\right]^{1/2q} &\leq \sum_{i=1}^n \mathbb{E}\left[\left|X_j(Z_i) + \sigma \mathbf{E}_{ij}\right|^{4q} \middle| Z\right]^{1/2q} \\
&\leq 2\sum_{i=1}^n \mathbb{E}\left(\left[\left|X_j(Z_i)\right|^{4q} \middle| Z\right]^{1/2q} + \mathbb{E}\left[\left|\sigma \mathbf{E}_{ij}\right|^{4q} \middle| Z\right]^{1/2q}\right) \\
&\leq 2n\left(\sup_{l\geq 1}\sup_{z\in\mathcal{Z}} \mathbb{E}\left[\left|X_l(z)\right|^{4q}\right]^{1/2q} + \sigma^2 \sup_{i\geq 1, l\geq 1} \mathbb{E}\left[\left|\mathbf{E}_{il}\right|^{4q}\right]^{1/2q}\right),
\end{aligned}$$

where the final inequality uses the facts that X_j, Z and $\mathbf{E}$ are independent.

Combining the above estimates we find:

$$\begin{aligned}
&\mathbb{P}\left(\left\|p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top - \sigma^2 \mathbf{I}_n\right\|_2 \geq t \middle| Z_1, \ldots, Z_n\right) \\
&\leq \frac{2^q(2q-1)^q}{t^{2q}}\left(\frac{2}{p}\right)^q 4^q n^{2q}\left(\sup_{j\geq 1}\sup_{z\in\mathcal{Z}} \mathbb{E}\left[\left|X_j(z)\right|^{4q}\right]^{1/2q} + \sigma^2 \sup_{i\geq 1, j\geq 1} \mathbb{E}\left[\left|\mathbf{E}_{ij}\right|^{4q}\right]^{1/2q}\right)^{2q} \\
&= (16)^q(2q-1)^q \frac{n^{2q}}{t^{2q}}\frac{1}{p^q}\left(\sup_{j\geq 1}\sup_{z\in\mathcal{Z}} \mathbb{E}\left[\left|X_j(z)\right|^{4q}\right]^{1/2q} + \sigma^2 \sup_{i\geq 1, j\geq 1} \mathbb{E}\left[\left|\mathbf{E}_{ij}\right|^{4q}\right]^{1/2q}\right)^{2q},
\end{aligned}$$

from which the result follows by the tower property of conditional expectation. □

Proposition 7. *Assume **A1**, **A6** and **A9** with some $q \geq 1$. For any $\delta, \epsilon \in (0,1)$, if*

$$n \geq \frac{3\sigma^2}{\epsilon c_\lambda^{\min}} \vee \left[\log\left(\frac{1}{\delta}\right) + \log(4r)\right]\frac{1}{\epsilon^2}\frac{2((c_\lambda^{\max})^2 + \epsilon c_\lambda^{\max} c_\lambda^{\min}/9)}{(c_\lambda^{\min})^2/9},$$

and

$$p \geq \frac{1}{\delta^{1/q}\epsilon^2} 2^{1/q} 16(2q-1)\frac{9}{(c_\lambda^{\min})^2}\left(c_X(2q)^{1/2q} + \sigma^2 c_E(2q)^{1/2q}\right)^2$$

where c_X, c_E are as in lemma 9 and $c_\lambda^{\max}$, $c_\lambda^{\min}$ are as in lemma 7, then

$$\mathbb{P}\left(\bigcap_{i=1}^n B_{\mathbf{Y},i}(\epsilon) \cap \bigcap_{i=1}^r B_{\mathbf{\Phi},i}(\epsilon)\right) \geq 1 - \delta.$$

Proof. Throughout the proof we shall adopt the convention $\lambda_i^f := 0$ for all $r+1 \leq i \leq n$ and, in several places, we shall use the fact that $\lambda_i(\mathbf{\Phi}\mathbf{\Phi}^\top) = 0$ for $r+1 \leq i \leq n$ which holds since $\mathbf{\Phi} \in \mathbb{R}^{n\times r}$.

Consider the following decomposition for any $1 \leq i \leq n$:

$$\begin{aligned}
\left|\frac{1}{n}\lambda_i(p^{-1}\mathbf{Y}\mathbf{Y}^\top) - \lambda_i^f\right| &\leq \left|\frac{1}{n}\lambda_i(p^{-1}\mathbf{Y}\mathbf{Y}^\top) - \frac{1}{n}\lambda_i(\mathbf{\Phi}\mathbf{\Phi}^\top + \sigma^2\mathbf{I}_n)\right| \\
&\quad + \left|\frac{1}{n}\lambda_i(\mathbf{\Phi}\mathbf{\Phi}^\top + \sigma^2\mathbf{I}_n) - \frac{1}{n}\lambda_i(\mathbf{\Phi}\mathbf{\Phi}^\top)\right| \\
&\quad + \left|\frac{1}{n}\lambda_i(\mathbf{\Phi}\mathbf{\Phi}^\top) - \lambda_i^f\right|.
\end{aligned}$$

Combining this decomposition with Weyl's inequality; the facts that for $1 \le i \le r$, $\mathbb{E}[\mathbf{\Phi}^\top \mathbf{\Phi}]_{ii} = n\lambda_i^f$ and $\mathbb{E}[\mathbf{\Phi}^\top \mathbf{\Phi}]_{ij} = 0$ for $j \neq i$, hence $\lambda_i^f = \lambda_i(n^{-1}\mathbb{E}[\mathbf{\Phi}^\top \mathbf{\Phi}])$; and by lemma 3, $\lambda_i(\mathbf{\Phi}\mathbf{\Phi}^\top) = \lambda_i(\mathbf{\Phi}^\top \mathbf{\Phi})$; whilst for $i \ge r+1$, $\lambda_i(\mathbf{\Phi}\mathbf{\Phi}^\top) = \lambda_i^f = 0$; we obtain:

$$\begin{aligned}\max_{1\le i\le n}\left|\frac{1}{n}\lambda_i(p^{-1}\mathbf{Y}\mathbf{Y}^\top) - \lambda_i^f\right| \le \frac{1}{n}\|p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top - \sigma^2\mathbf{I}_n\|_2 \\ + \frac{\sigma^2}{n} \\ + \|n^{-1}\mathbf{\Phi}^\top\mathbf{\Phi} - n^{-1}\mathbb{E}[\mathbf{\Phi}^\top\mathbf{\Phi}]\|_2 \end{aligned} \tag{66}$$

and

$$\max_{1\le i\le n}\left|\frac{1}{n}\lambda_i(\mathbf{\Phi}\mathbf{\Phi}^\top) - \lambda_i^f\right| \le \|n^{-1}\mathbf{\Phi}^\top\mathbf{\Phi} - n^{-1}\mathbb{E}[\mathbf{\Phi}^\top\mathbf{\Phi}]\|_2.$$

Now fix any $\epsilon \in (0,1)$. We have

$$\begin{aligned}&\mathbb{P}\left(\bigcap_{i=1}^n B_{\mathbf{Y},i}(\epsilon) \cap \bigcap_{i=1}^r B_{\mathbf{\Phi},i}(\epsilon)\right) \\ &\ge \mathbb{P}\left(\bigcap_{i=1}^n \left\{\left|\frac{1}{n}\lambda_i(p^{-1}\mathbf{Y}\mathbf{Y}^\top) - \lambda_i^f\right| < \epsilon\lambda_r^f\right\} \cap \left\{\left|\frac{1}{n}\lambda_i(\mathbf{\Phi}\mathbf{\Phi}^\top) - \lambda_i^f\right| < \epsilon\lambda_r^f\right\}\right) \\ &\ge 1 - \mathbb{P}\left(\frac{1}{n}\|p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top - \sigma^2\mathbf{I}_n\|_2 \ge \epsilon\lambda_r^f/3\right) - \mathbb{P}\left(\|n^{-1}\mathbf{\Phi}^\top\mathbf{\Phi} - n^{-1}\mathbb{E}[\mathbf{\Phi}^\top\mathbf{\Phi}]\|_2 \ge \epsilon\lambda_r^f/3\right) \\ &\ge 1 - (16)^q(2q-1)^q \frac{1}{(\epsilon c_\lambda^{\min}/3)^{2q}}\frac{1}{p^q}\left(c_X(2q)^{1/2q} + \sigma^2 c_E(2q)^{1/2q}\right)^{2q} - 2r\exp\left(\frac{-(\epsilon/3)^2(c_\lambda^{\min})^2 n/2}{(c_\lambda^{\max})^2 + c_\lambda^{\max}\epsilon c_\lambda^{\min}/9}\right)\end{aligned}$$

where the second inequality holds by using $\lambda_r^f \le \lambda_i^f$ for $i = 1, \dots r$, together with (66) and the condition of the proposition $n \ge 3\sigma^2/(\epsilon\lambda_r^f)$; and the third inequality holds by applying lemma 8 and lemma 9 and using $\lambda_r^f \ge c_\lambda^{\min}$.

The proof is completed by re-arranging each of the two following inequalities:

$$\delta/2 \ge (16)^q(2q-1)^q \frac{1}{(\epsilon c_\lambda^{\min}/3)^{2q}}\frac{1}{p^q}\left(c_X(2q)^{1/2q} + \sigma^2 c_E(2q)^{1/2q}\right)^{2q},$$

$$\frac{\delta}{2} \ge 2r\exp\left(\frac{-(\epsilon/3)^2(c_\lambda^{\min})^2 n/2}{(c_\lambda^{\max})^2 + c_\lambda^{\max}\epsilon c_\lambda^{\min}/9}\right).$$

□

Lemma 10. *For any $m_1, m_2 \ge 1$ and any matrix norm $\|\cdot\|_\star$ on $\mathbb{R}^{m_1\times m_2}$, $\|\cdot\|_\star$ is convex. For any random $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m_1\times m_2}$ and any $1 \le q < \infty$ such that $\mathbb{E}[\|\mathbf{A}\|_\star^q] \vee \mathbb{E}[\|\mathbf{B}\|_\star^q] < \infty$, $\mathbb{E}[\|\mathbf{A}+\mathbf{B}\|_\star^q]^{1/q} \le \mathbb{E}[\|\mathbf{A}\|_\star^q]^{1/q} + \mathbb{E}[\|\mathbf{B}\|_\star^q]^{1/q}$.*

Proof. The convexity holds due to the fact that any norm must be absolutely homogeneous and satisfy the triangle inequality. For the second claim, since $\mathbb{E}[\|\mathbf{A}\|_\star^q] \vee \mathbb{E}[\|\mathbf{B}\|_\star^q] < \infty$ we have the preliminary estimate $\mathbb{E}[\|\mathbf{A}+\mathbf{B}\|_\star^q] \le 2^{q-1}(\mathbb{E}[\|\mathbf{A}\|_\star^q] + \mathbb{E}[\|\mathbf{B}\|_\star^q]) < \infty$. If $\mathbb{E}[\|\mathbf{A}+\mathbf{B}\|_\star^q] = 0$ then the desired inequality is trivial. So suppose $\mathbb{E}[\|\mathbf{A}+\mathbf{B}\|_\star^q] > 0$. Using the triangle inequality for the norm and then Holder's inequality for the expectation,

$$\begin{aligned}\mathbb{E}[\|\mathbf{A}+\mathbf{B}\|_\star^q] &= \mathbb{E}\left[\|\mathbf{A}+\mathbf{B}\|_\star\|\mathbf{A}+\mathbf{B}\|_\star^{q-1}\right] \\ &\le \mathbb{E}\left[(\|\mathbf{A}\|_\star + \|\mathbf{B}\|_\star)\|\mathbf{A}+\mathbf{B}\|_\star^{q-1}\right] \\ &= \mathbb{E}\left[\|\mathbf{A}\|_\star\|\mathbf{A}+\mathbf{B}\|_\star^{q-1}\right] + \mathbb{E}\left[\|\mathbf{B}\|_\star\|\mathbf{A}+\mathbf{B}\|_\star^{q-1}\right] \\ &\le \left(\mathbb{E}[\|\mathbf{A}\|_\star^q]^{1/q} + \mathbb{E}[\|\mathbf{B}\|_\star^q]^{1/q}\right)\mathbb{E}\left[\|\mathbf{A}+\mathbf{B}\|_\star^{(q-1)(\frac{q}{q-1})}\right]^{1-\frac{1}{q}} \\ &= \left(\mathbb{E}[\|\mathbf{A}\|_\star^q]^{1/q} + \mathbb{E}[\|\mathbf{B}\|_\star^q]^{1/q}\right)\frac{\mathbb{E}[\|\mathbf{A}+\mathbf{B}\|_\star^q]}{\mathbb{E}[\|\mathbf{A}+\mathbf{B}\|_\star^q]^{1/q}}.\end{aligned}$$

The proof is completed by multiplying both sides by $\mathbb{E}[\|\mathbf{A}+\mathbf{B}\|_\star^q]^{1/q}/\mathbb{E}[\|\mathbf{A}+\mathbf{B}\|_\star^q]$. □

Lemma 11. *Assume A1, A6, and A9 with some $q \geq 1$. Let U_j denote the jth column of $\mathbf{U}_{\mathbf{\Phi}}$. Then there exists a constant $b(q)$ depending only on q such that for any $t > 0$,*

$$\mathbb{P}\left(\max_{j=1,\ldots,r} \|(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top - \sigma^2\mathbf{I}_n)U_j\|_\infty \leq t\right)$$
$$\geq 1 - \frac{n^{1+q}r}{t^{2q}p^q} b(2q)2^{6q-1}\left(\max_{j=1,\ldots,p}\sup_{z\in\mathcal{Z}} \mathbb{E}\left[|X_j(z)|^{4q}\right] + \sigma^{4q}\max_{i=1,\ldots,n,j=1,\ldots,p}\mathbb{E}[|\mathbf{E}_{ij}|^{4q}]\right).$$

Proof. The ith element of $(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top - \sigma^2\mathbf{I}_n)U_j$ can be written in the form:

$$p^{-1}\sum_{k=1}^{p} \mathbf{\Delta}_{ij}(k)$$

where

$$\mathbf{\Delta}_{ij}(k) \coloneqq Y_k^{(i)}Y_k^\top U_j - \mathbb{E}\left[Y_k^{(i)}Y_k^\top U_j \middle| Z_1,\ldots,Z_n\right]$$

and for any i, j, the random variables $\mathbf{\Delta}_{ij}(k)$, $k = 1,\ldots,p$ are conditionally independent and conditionally mean zero given $Z_1,\ldots,Z_n$.

Applying Markov's inequality, the Marcinkiewicz-Zygmund inequality and Minkowski's inequality, all conditionally on $Z \equiv (Z_1,\ldots,Z_n)$, we have for any $q \geq 1$ the following inequalities hold almost surely,

$$\begin{aligned}
\mathbb{P}\left(\left|p^{-1}\sum_{k=1}^{p}\mathbf{\Delta}_{ij}(k)\right| \geq t \middle| Z\right) &\leq \frac{1}{t^{2q}}\mathbb{E}\left[\left|p^{-1}\sum_{k=1}^{p}\mathbf{\Delta}_{ij}(k)\right|^{2q}\middle| Z\right] \\
&\leq \frac{b(2q)}{t^{2q}p^{2q}}\mathbb{E}\left[\left|\sqrt{\sum_{k=1}^{p}|\mathbf{\Delta}_{ij}(k)|^2}\right|^{2q}\middle| Z\right] \\
&\leq \frac{b(2q)}{t^{2q}p^{2q}}\left(\sum_{k=1}^{p}\mathbb{E}\left[|\mathbf{\Delta}_{ij}(k)|^{2q}\middle| Z\right]^{1/q}\right)^q \\
&= \frac{b(2q)}{t^{2q}p^q}\max_{k=1,\ldots,p}\mathbb{E}\left[|\mathbf{\Delta}_{ij}(k)|^{2q}\middle| Z\right]. \qquad (67)
\end{aligned}$$

Re-arranging the expression for $\mathbf{\Delta}_{ij}(k)$, applying the Cauchy-Schwartz inequality and $\|U_j\|_2 = 1$, we estimate

$$\begin{aligned}
|\mathbf{\Delta}_{ij}(k)| &\leq \left\|Y_k^{(i)}Y_k - \mathbb{E}\left[Y_k^{(i)}Y_k\middle| Z\right]\right\|_2 \|U_j\|_2 \\
&\leq |Y_k^{(i)}|\|Y_k\|_2 + \mathbb{E}\left[|Y_k^{(i)}|\|Y_k\|_2\middle| Z\right]
\end{aligned}$$

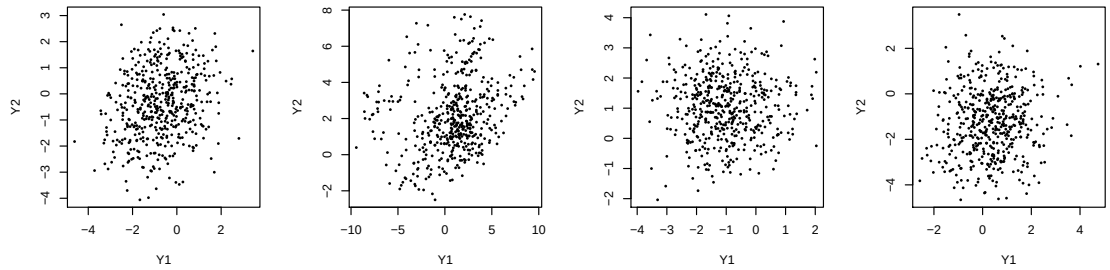

Figure 15: First two coordinates of the data matrices corresponding to figure 6, showing much less structure than the principal components.

and so

$$\begin{aligned}
\mathbb{E}\left[|\mathbf{\Delta}_{ij}(k)|^{2q}\middle| Z\right] &\leq 2^{2q}\mathbb{E}\left[(|Y_k^{(i)}|^2\|Y_k\|_2^2)^q\middle| Z\right] \\
&= 2^{2q}\mathbb{E}\left[\left(\sum_{l=1}^n |Y_k^{(i)}|^2|Y_k^{(l)}|^2\right)^q\middle| Z\right] \\
&\leq 2^{2q}\left(\sum_{l=1}^n \mathbb{E}\left[(|Y_k^{(i)}|^2|Y_k^{(l)}|^2)^q\middle| Z\right]^{1/q}\right)^q \\
&\leq 2^{2q}\left(\sum_{l=1}^n \mathbb{E}\left[|Y_k^{(i)}|^{4q}\middle| Z\right]^{1/(2q)}\mathbb{E}\left[|Y_k^{(l)}|^{4q}\middle| Z\right]^{1/(2q)}\right)^q \\
&= 2^{2q}n^q \max_{l=1,\ldots,n}\mathbb{E}\left[|Y_k^{(l)}|^{4q}\middle| Z\right] \\
&= 2^{2q}n^q \max_{l=1,\ldots,n}\mathbb{E}\left[|X_k(Z_l)+\sigma\mathbf{E}_{kl}|^{4q}\middle| Z\right] \\
&\leq 2^{6q-1}n^q\left(\sup_{l\geq 1}\sup_{z\in\mathcal{Z}}\mathbb{E}\left[|X_l(z)|^{4q}\right]+\sigma^{4q}\sup_{l\geq 1,\tilde{l}\geq 1}\mathbb{E}[|\mathbf{E}_{l\tilde{l}}|^{4q}]\right). \qquad (68)
\end{aligned}$$

Combining the almost sure upper bounds (68) and (67), using the tower property of conditional expectation and then taking a union bound over $i = 1, \ldots, n$ and $j = 1, \ldots, r$, we find:

$$\begin{aligned}
&\mathbb{P}\left(\max_{j=1,\ldots,r}\|(p^{-1}\mathbf{Y}\mathbf{Y}^\top - \mathbf{\Phi}\mathbf{\Phi}^\top - \sigma^2\mathbf{I}_n)U_j\|_\infty \leq t\right) \\
&\geq 1 - \frac{n^{1+q}r}{t^{2q}p^q}b(2q)2^{6q-1}\left(\sup_{j\geq 1}\sup_{z\in\mathcal{Z}}\mathbb{E}\left[|X_j(z)|^{4q}\right]+\sigma^{4q}\sup_{i\geq 1,j\geq 1}\mathbb{E}[|\mathbf{E}_{ij}|^{4q}]\right),
\end{aligned}$$

which completes the proof. □

E Supplementary figures for section 4.2

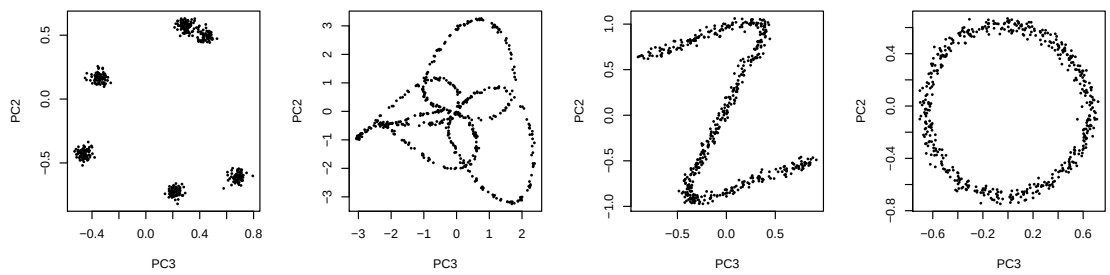

Figure 16: Third and second principal components of the data matrices corresponding to figure 6 (ordered like this to make the resemblance to $\mathcal{Z}$ more obvious).

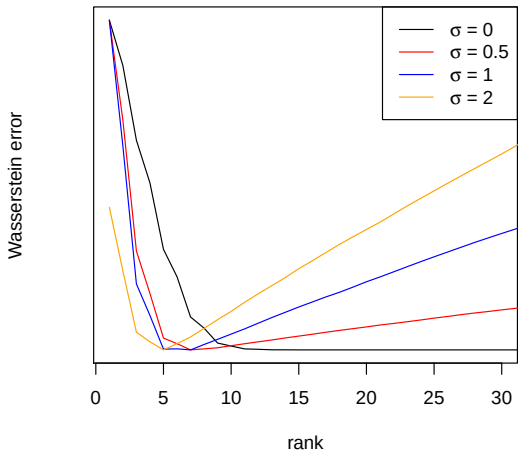

Figure 17: log-Wasserstein error for the fourth configuration in figure 6, for different error variances. As the variance increases, the optimal dimension (point achieving lowest error) decreases. The curves are shifted and rescaled so that their maxima and minima agree.