

Discovering latent topology and geometry in data: a law of large dimension

Nick Whiteley, Annie Gray, and Patrick Rubin-Delanchy

School of Mathematics, University of Bristol

September 2, 2022

Abstract

Complex topological and geometric patterns often appear embedded in high-dimensional data and seem to reflect structure related to the underlying data source, with some distortion. We show that this rich data morphology can be explained by a generic and remarkably simple statistical model, demonstrating that manifold structure in data can emerge from elementary statistical ideas of correlation and latent variables. The Latent Metric Space model consists of a collection of random fields, evaluated at locations specified by latent variables and observed in noise. Driven by high dimensionality, principal component scores associated with data from this model are uniformly concentrated around a topological manifold, homeomorphic to the latent metric space. Under further assumptions this relation may be a diffeomorphism, a Riemannian metric structure appears, and the geometry of the manifold reflects that of the latent metric space. This provides statistical justification for manifold assumptions which underlie methods ranging from clustering and topological data analysis, to nonlinear dimension reduction, regression and classification, and explains the efficacy of Principal Component Analysis as a preprocessing tool for reduction from high to moderate dimension.

1 Introduction

Assumptions that data are concentrated near embedded topological or geometric structures underpin a wide variety of statistical methods, such as clustering and topological data analysis, nonlinear dimension reduction, manifold estimation and advanced classification and regression techniques. In collective terms, such assumptions are sometimes loosely called “the manifold hypothesis” [Bengio et al., 2013, Fefferman et al., 2016], the spirit of which is captured in the following quote from Cayton [2005]:

“...the idea that 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”.

The objective of the present work is to provide a general statistical explanation for this phenomenon.

In some situations, such as image analysis, embedded manifold structure in data can be explained intuitively, albeit heuristically, 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 75 images from [Geusebroek et al., 2005], taken from angles $0, 5, 10, \dots, 355$ 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 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 of the principal component (PC) scores – the coordinates of the data with respect to the eigenvectors associated with the three largest eigenvalues – appear around a loop which is somewhat irregular in shape but resembles the circle of camera positions, subject to deformation by bending and twisting. The PC scores appear roughly equally spaced around the loop, resembling the geometry of the camera positions which are equally spaced at intervals of 5 degrees around a circle.

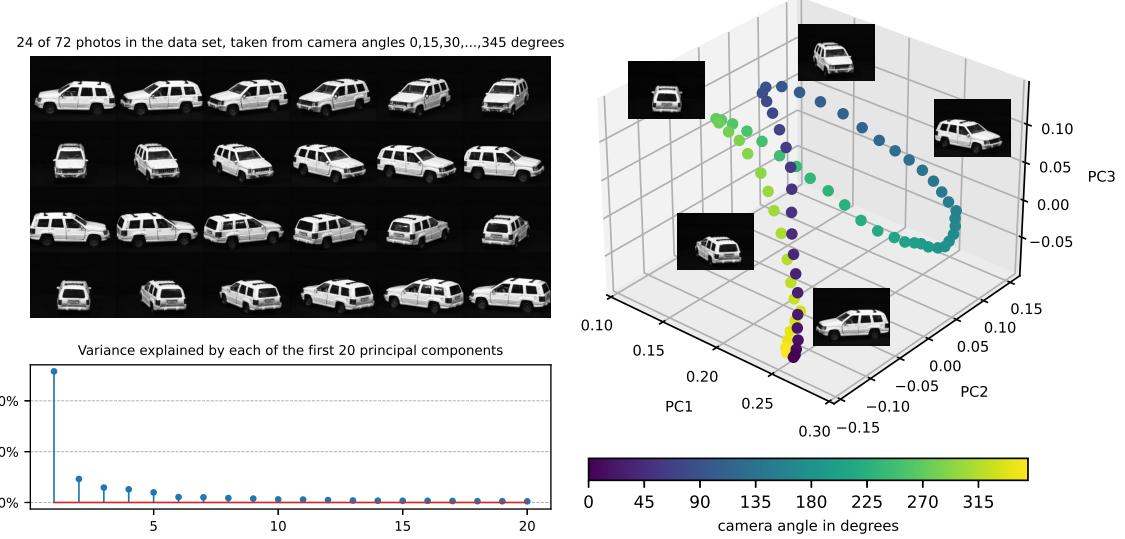


Figure 1: A collection of images reduced in dimension using PCA.

Evidently reducing the dimension of these image data by computing PC scores 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 of the PC scores 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 PC scores 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 5000 randomly chosen documents from the 20 Newsgroups dataset, originating from Lang [1995] and available via `scikit-learn` [Pedregosa et al., 2011] in Python. In preprocessing, each document is summarised by 7.6×10^4 Term Frequency Inverse Document Frequency features, a widely used method for vectorising documents. The left plot in figure 2 shows the result of dimension reduction from 7.6×10^4 to 2 using PCA. The right plot shows the result of first reducing from 7.6×10^4 to 20 dimensions using PCA, followed by reduction to 2 dimensions using t-SNE [Van der Maaten and Hinton, 2008], a nonlinear dimension reduction method which finds a lower dimensional representation of a dataset in a way designed to minimise a particular measure of distortion of pairwise distances. We used the default t-SNE parameter settings in `scikit-learn`. In both plots the points are coloured by the newsgroups from which the documents originate, but this information was not used to help PCA or t-SNE.

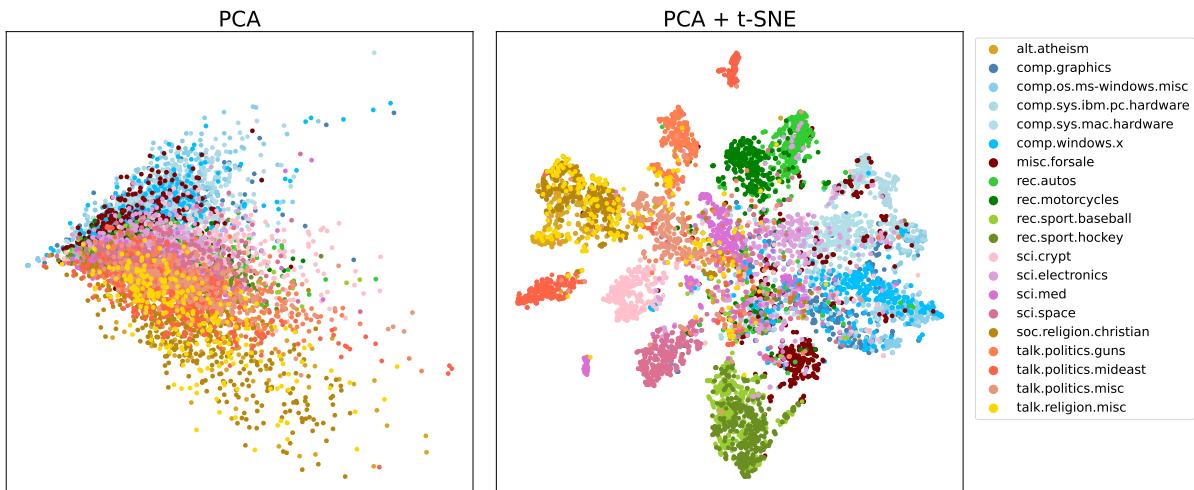


Figure 2: 20 Newsgroups example. Left: first 2 dimensions of the PC scores. Right: representation of the data in 2 dimensions obtained by first reducing to 20 dimensions using PCA, then applying t-SNE.

Similarly to figure 1, it is evident from figure 2 that performing some form of dimension reduction allows us to access structure which underlies the data. However, there are notable differences between these two examples. Firstly, in figure 2 the labels convey something quite abstract about the corpus of documents – newsgroup topics – whereas in figure 1, the camera positions have a direct physical interpretation. Secondly, the newsgroup topics are discrete and hierarchical, but do not have an inherent geometry, whereas in figure 1 the camera positions are points in Euclidean space. Thirdly in figure 2 using only PCA and reducing to 2 dimensions, documents associated with distinct newsgroups are not clearly separated, whereas PCA down to 20 dimensions followed by t-SNE is much more effective in revealing newsgroup clusters. By contrast, in figure 1 using PCA to reduce to 3 dimensions was enough to make visible the geometry of the data generating mechanism.

These differences illustrate just some of the ways in which different kinds of underlying structure can manifest themselves 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].

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 objective is not to devise very precise or accurate models for particular datasets, but rather to demonstrate that rich topological and geometric structure can emerge from generic and simple statistical assumptions. This provides an essentially statistical grounding for the ideas described by the manifold hypothesis.

We shall dispense with the default assumption underlying many statistical analyses and much of high-dimensional statistical theory that multivariate data points are independent and identically distributed. Instead we consider a model in which, modulo noise, data points are dependent and exchangeable, and there is conditional independence but not exchangeability across dimensions for each data point. This dependence structure arises from modelling each data point as the evaluation of a collection of random fields at an abstract location given by a latent variable valued in a metric space. Under this model, we show that PC scores concentrate around a manifold which may be topologically or geometrically related to the latent metric space. We emphasise that the presence of this manifold is not an axiomatic starting point, but a consequence of a combination of fairly plausible statistical assumptions. High dimensionality is a key factor driving this concentration and we call this a “law of large dimension”.

2 The Latent Metric Space model

2.1 Model definition and assumptions

The Latent Metric space model is constructed from three independent sources of randomness.

Latent Variables. Z_1, \dots, Z_n are random elements of a compact metric space $(\mathcal{Z}, d_{\mathcal{Z}})$, independent and identically distributed according to a Borel probability measure μ supported on \mathcal{Z} .

Random Fields. X_1, \dots, X_p are independent but not necessarily identically distributed, \mathbb{R} -valued and square-integrable random fields, each with index set \mathcal{Z} . That is, for each $z \in \mathcal{Z}$ and $j = 1, \dots, p$, $X_j(z)$ is an \mathbb{R} -valued random variable such that $\mathbb{E}[|X_j(z)|^2] < \infty$.

Noise. $\mathbf{E} \in \mathbb{R}^{p \times n}$ 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:

$$\mathbf{Y}_{ij} := X_j(Z_i) + \sigma \mathbf{E}_{ij} \quad (1)$$

for some $\sigma \geq 0$.

We shall later refer to the following assumptions:

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

A2. For some $q \geq 1$, $\max_{j=1,\dots,p} \sup_{z \in \mathcal{Z}} \mathbb{E}[|X_j(z)|^{4q}] < \infty$ and $\max_{j=1,\dots,p, i=1,\dots,n} \mathbb{E}[|\mathbf{E}_{ij}|^{4q}] < \infty$.

Consider the *mean correlation kernel*:

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

By a generalisation of Mercer's theorem [Mercer, 1909] which applies to the case of interest here where $(\mathcal{Z}, d_{\mathcal{Z}})$ is a compact metric space and μ is a Borel probability measure on \mathcal{Z} (see theorem 7 in section A, appendix, for details), when **A1** holds, there exists a countable collection of non-negative real numbers $(\lambda_k^f)_{k \geq 1}$, $\lambda_1^f \geq \lambda_2^f \geq \dots$, and a sequence of functions $(u_k^f)_{k \geq 1}$ which are orthonormal in $L_2(\mu)$ such that, with

$$\phi(z) := [(\lambda_1^f)^{1/2} u_1^f(z) (\lambda_2^f)^{1/2} u_2^f(z) \dots]^{\top}. \quad (2)$$

the kernel f has the representation:

$$f(z, z') = \langle \phi(z), \phi(z') \rangle_2 = \sum_{k=1}^{\infty} \lambda_k^f u_k^f(z) u_k^f(z'), \quad (3)$$

where the convergence is absolute and uniform. The dependence of $(\lambda_k^f, u_k^f)_{k \geq 1}$ on μ is not shown in the notation.

In some but not all of our analysis we shall consider the following assumption, which says that the kernel f has finite rank.

A3. For some $r \geq 1$, $\lambda_k^f > 0$ for all $k \leq r$ and $\lambda_k^f = 0$ for all $k > r$.

When **A3** holds we shall abuse notation slightly and let $\phi(z) \equiv [(\lambda_1^f)^{1/2} u_1^f(z) \dots (\lambda_r^f)^{1/2} u_r^f(z)]^{\top}$.

2.2 Principal Component Scores

Given data $\mathbf{Y} \in \mathbb{R}^{n \times p}$ and $r \leq \min\{p, n\}$, let the columns of $\mathbf{V}_Y \in \mathbb{R}^{p \times r}$ be orthonormal eigenvectors associated with the r largest eigenvalues of $\mathbf{Y}^{\top} \mathbf{Y} \in \mathbb{R}^{p \times p}$. The dimension- r PC scores, ζ_1, \dots, ζ_n , each valued in \mathbb{R}^r , are:

$$[\zeta_1 | \dots | \zeta_n]^{\top} := \mathbf{Y} \mathbf{V}_Y.$$

Calculation of PC scores is often performed for purposes of dimension reduction, this reduction is from dimension p to r for each of the n rows of the data matrix \mathbf{Y} .

When conducting PCA the data are usually first centered using the sample mean, and eigenvectors then computed from a sample covariance matrix. This leads to the interpretation of eigenvalues as the sample variance explained in the subspaces corresponding to eigenvectors. This variance-explained interpretation is not essential for the developments in the present work; note that in the Latent Metric Space model we have not made any assumptions about the expected values of the random variables $X_j(z)$, $j = 1, \dots, p$, $z \in \mathcal{Z}$, so we have not assumed that $\mathbb{E}[\mathbf{Y}_{ij}] = 0$, but at the same time there is nothing in our setup which rules out $\mathbb{E}[\mathbf{Y}_{ij}] = 0$. Nevertheless the PC scores we shall study, as defined above, are derived from the eigenvectors of the matrix $\mathbf{Y}^{\top} \mathbf{Y}$, i.e., without sample-centering. We do this primarily for ease of exposition in our theoretical results, the consequences of working with sample-centered data are discussed in section 5.9.

3 Uniform concentration of PC scores

Theorem 1. Assume **A1-A3** hold for some $q \geq 1$ and $r \geq 1$. Let \mathbf{Y} follow the model of section 2 and let ζ_1, \dots, ζ_n be the dimension- r PC scores defined in section 2.2. Then there exists a random orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{r \times r}$ 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

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

with probability at least $1 - \delta$. Here c_1 depends only on $f, \lambda_1^f, \lambda_r^f$; $c_2(q)$ depends only on $q, \lambda_1^f, \lambda_r^f, \max_{j=1,\dots,p} \sup_{z \in \mathcal{Z}} \mathbb{E}[|X_j(z)|^{4q}], \max_{i=1,\dots,n, j=1,\dots,p} \mathbb{E}[|\mathbf{E}_{ij}|^{4q}]$ and σ^2 ; and $\|\cdot\|_2$ is the Euclidean norm.

The proof of theorem 1 is in section B, appendix.

The roles of p and n

To help interpret theorem 1, suppose for purposes of exposition that the fields X_1, \dots, X_p are identically distributed. In this situation we have

$$f(z, z') = \mathbb{E}[X_1(z)X_1(z')],$$

so f is fixed irrespective of p , and $\max_{j=1,\dots,p} \sup_{z \in \mathcal{Z}} \mathbb{E}[|X_j(z)|^{4q}]$ does not depend on p . If also $\max_{i=1,\dots,n, j=1,\dots,p} \mathbb{E}[|\mathbf{E}_{ij}|^{4q}]$ and σ^2 do not depend on p , then it is a corollary of theorem 1 that:

$$\max_{i=1,\dots,n} \left\| p^{-1/2} \mathbf{Q} \zeta_i - \phi(Z_i) \right\|_2 \rightarrow 0 \text{ in probability if both } n \rightarrow \infty \text{ and } \frac{p}{n} \rightarrow \infty. \quad (4)$$

In the proof of theorem 1, the logarithmic dependence $\log(1/\delta)$ can be traced to sub-Gaussian concentration of certain random matrices associated with the latent variables Z_1, \dots, Z_n , whilst the polynomial dependence $1/\delta^q$ can be traced to concentration of certain random matrices associated with the random fields X_1, \dots, X_p and the p columns of \mathbf{E} , obtained under assumption A2.

The convergence (4) tells us that in order for $\max_{i=1,\dots,n} \left\| p^{-1/2} \mathbf{Q} \zeta_i - \phi(Z_i) \right\|_2$ to be small with high probability, it is sufficient that both n and p/n are large. What it really means to be “large” here depends on the best possible constants which could, in principle, be taken for c_1 and $c_2(q)$ in theorem 1. The authors believe their theoretical analysis is not sharp enough to be very revealing here. However, in section 5.2 we use simulations to examine how $\max_{i=1,\dots,n} \left\| p^{-1/2} \mathbf{Q} \zeta_i - \phi(Z_i) \right\|_2$ is influenced by p and n , in particular addressing the case $p \leq n$.

The dimension r

It is important to notice that in theorem 1 the dimension r of the PC scores coincides with the rank of the kernel f , which is finite by assumption A3. This is a simplification in two respects. Firstly, it is a mathematical convenience to assume that f is finite rank, we make this assumption only because it is what we can accommodate using our proof techniques. In order to generalise our results, one might consider looser assumptions under which r can be infinite, perhaps subject to conditions on the rate of decay of the eigenvalues. Secondly, in practice one usually adopts some data-driven method to choose the dimension of the PC scores. Both of these points are further discussed in section 5.8; in particular, we show by simulation that applying PCA into moderate dimension can be beneficial, achieving a favourable bias/variance trade-off, even under infinite rank.

The random orthogonal matrix \mathbf{Q}

Transformation of a set of vectors by multiplication with an orthogonal matrix, by definition, preserves inner-products and hence distances. Theorem 1 therefore implies that the pairwise distance $p^{-1/2} \|\zeta_i - \zeta_j\|_2$ is concentrated around $\|\phi(Z_i) - \phi(Z_j)\|_2$, uniformly in i, j . To see this consider the fact that if $\|p^{-1/2} \mathbf{Q} \zeta_i - \phi(Z_i)\|_2 \leq \epsilon$, then

$$\begin{aligned} p^{-1/2} \|\zeta_i - \zeta_j\|_2 &\leq \|p^{-1/2} \zeta_i - \mathbf{Q}^{-1} \phi(Z_i)\|_2 + \|\mathbf{Q}^{-1} \phi(Z_i) - \mathbf{Q}^{-1} \phi(Z_j)\|_2 + \|\mathbf{Q}^{-1} \phi(Z_j) - p^{-1/2} \zeta_j\|_2 \\ &= \|p^{-1/2} \mathbf{Q} \zeta_i - \phi(Z_i)\|_2 + \|\phi(Z_i) - \phi(Z_j)\|_2 + \|p^{-1/2} \mathbf{Q} \zeta_j - \phi(Z_j)\|_2 \\ &\leq 2\epsilon + \|\phi(Z_i) - \phi(Z_j)\|_2 \end{aligned}$$

and similarly

$$\|\phi(Z_i) - \phi(Z_j)\|_2 \leq 2\epsilon + p^{-1/2} \|\zeta_i - \zeta_j\|_2,$$

hence

$$\max_{i,j=1,\dots,n} \left| p^{-1/2} \|\zeta_i - \zeta_j\|_2 - \|\phi(Z_i) - \phi(Z_j)\|_2 \right| \leq 2\epsilon. \quad (5)$$

Many methods which may be applied downstream of computing PC scores, e.g. from Topological Data Analysis, nonlinear dimension reduction, nearest-neighbour classification or regression and certain types of clustering, take as inputs pairwise distances. Hence for such techniques, via (5) theorem 1 quantifies how far away these pairwise distances are from the “ideal” pairwise distances $\|\phi(Z_i) - \phi(Z_j)\|_2$. Readers interested in the mathematical reason why \mathbf{Q} appears in theorem 1 are directed to the start of the proof in section B.2, where this orthogonal matrix is defined constructively.

4 Topological and geometric relationships between \mathcal{M} and \mathcal{Z}

Theorem 1 tells us that, up to an orthogonal transformation, the scaled PC scores $\{p^{-1/2}\zeta_i; i = 1, \dots, n\}$ are uniformly concentrated around the corresponding points $\{\phi(Z_i); i = 1, \dots, n\}$. These latter points are each valued in $\mathcal{M} := \phi(\mathcal{Z})$, the image of \mathcal{Z} by ϕ . In section 4 we explore how \mathcal{M} is topologically and geometrically related to \mathcal{Z} , allowing us to understand how the PC scores convey the structure of \mathcal{Z} . Part of the study of kernel feature maps we develop in section 4 refines and extends previous work of the authors [Rubin-Delanchy, 2020, Whiteley et al., 2021]. This relationship is detailed at the end of section 4.3.

4.1 A brief introduction to manifold terminology

A *homeomorphism* is a mapping between two topological spaces which is continuous, invertible and has a continuous inverse. When a homeomorphism exists between two such spaces they are said to be homeomorphic, or *topologically equivalent*.

A sufficient condition for topological equivalence of two subsets of Euclidean space, loosely stated, is that each can be continuously deformed into the other by bending, pulling or stretching, but without breaking or tearing [Bing, 1960]. A canonical example of a pair of sets whose topological equivalence can be understood in these terms is a tea cup and a ring doughnut. However, transformations by such deformations constitute only one narrow aspect of topological equivalence. For example, all pairs in figure 3 are homeomorphic, and yet the doughnut could not be continuously deformed into either of the knots without somehow breaking and then re-connecting.

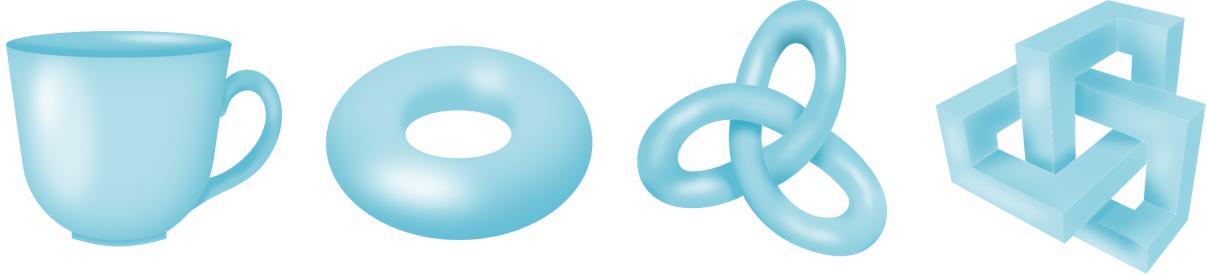


Figure 3: Examples of homeomorphic and diffeomorphic subsets of \mathbb{R}^3 . All pairs are homeomorphic. The left-most two are not diffeomorphic, because the tea cup has sharp edges but the doughnut does not; the middle two are diffeomorphic, the right-most two are not diffeomorphic, because the knot on the far right has sharp edges but the other knot is smooth.

Whilst the pictures in figure 3 convey some feel for topological equivalence of shapes in Euclidean space, the general concept of topological equivalence is by no means limited to that context. For an illustration in a discrete setting, consider an abstract set with three distinct elements $\mathcal{U} = \{a, b, c\}$. To turn \mathcal{U} into a topological space we can endow it with the discrete metric:

$$d_{\mathcal{U}}(u, u') := \begin{cases} 0, & u = u', \\ 1, & u \neq u'. \end{cases}$$

Under this metric all subsets of \mathcal{U} are open. Now let $\mathcal{V} = \{v_a, v_b, v_c\}$ be a set of three distinct points in \mathbb{R}^d , endowed with the Euclidean distance (similarly all subsets of \mathcal{V} are open). Then the mapping $\psi(u) := v_u$, $u \in \mathcal{U}$, is a homeomorphism. This example might seem trivial, but it turns out to be central to our later discussion of mixture models and clustering.

A *diffeomorphism* takes the definition of a homeomorphism further by requiring that the mapping in question and its inverse are not only continuous but also in some sense differentiable. The specific

concept of derivative involved may depend on the context. For example, denoting any pair of the sets in figure 3 by $\mathcal{U}, \mathcal{V} \subset \mathbb{R}^3$ and $\psi : \mathcal{U} \rightarrow \mathcal{V}$ a homeomorphism, neither \mathcal{U} nor \mathcal{V} is a linear subspace of \mathbb{R}^3 , so if $u \in \mathcal{U}$, $w \in \mathbb{R}^3$ and $h \in \mathbb{R}$, in general $u + hw \notin \mathcal{U}$. Therefore neither the difference quotient $[\psi(u + hw) - \psi(u)]/h$ nor the associated notion of directional derivative is well defined. Instead one can appeal to the following notion of differentiation with respect to a curve: $\psi : \mathcal{U} \rightarrow \mathcal{V}$ can be said to be a diffeomorphism if, for any curve $t \mapsto \eta_t$ in \mathcal{U} which is differentiable in the sense of elementary calculus, the mapping $t \mapsto \psi(\eta_t)$ is a differentiable curve in \mathcal{V} ; and also for any differentiable curve $t \mapsto \gamma_t$ in \mathcal{V} , the mapping $t \mapsto \psi^{-1}(\gamma_t)$ is a differentiable curve in \mathcal{U} . Derivatives with respect to curves in fact underlie the concept of a Lie derivative, see e.g., [Petersen, 2006, Ch. 2], but we shan't need details of that here.

We thus see that such a diffeomorphism ψ cannot involve deforming a smooth set in a way which creates non-differentiably sharp edges or rough surfaces, since that would result in some differentiable curves being transformed by ψ into non-differentiable curves. For this reason neither the left-most pair nor the right-most pair of sets in figure 3 are diffeomorphic. Also note that in the discrete example of $\mathcal{U} = \{a, b, c\}$ and $\mathcal{V} = \{v_a, v_b, v_b\}$ discussed above, mappings between \mathcal{U} and \mathcal{V} come along with no automatically present nor conventionally defined notion of differentiation, so in that situation the concept of a diffeomorphism does not have meaning.

With the concepts of homeomorphism and diffeomorphism introduced we can now speak of manifolds: a *topological manifold* is a topological space which is locally homeomorphic to an underlying topological space [Lee, 2010]. A *differentiable manifold* is locally diffeomorphic to an underlying domain, and a *d-dimensional differentiable manifold* is locally diffeomorphic to \mathbb{R}^d . The interested reader is referred to section C.1 for precise details of this latter definition, based on the presentation of Guillemin and Pollack [1974].

4.2 Topological equivalence of \mathcal{M} and \mathcal{Z}

In sections 4.2 and 4.3 we shall explore \mathcal{M} as a topological and then differentiable manifold with the underlying domain being the latent metric space $(\mathcal{Z}, d_{\mathcal{Z}})$. We need some further notation to accommodate this discussion. By definition, the set $\mathcal{M} := \phi(\mathcal{Z})$ is a subset of $\mathbb{R}^{\mathbb{N}} = \{[x_1 x_2 \dots]^{\top}; x_k \in \mathbb{R}, k \in \mathbb{N}\}$. Consider the inner-product and norm, $\langle x, x' \rangle_2 := \sum_{k=1}^{\infty} x_k x'_k$, $\|x\|_2 := \langle x, x \rangle_2^{1/2}$, where $x, x' \in \mathbb{R}^{\mathbb{N}}$, and let ℓ_2 be the real Hilbert space: $\{x \in \mathbb{R}^{\mathbb{N}} : \|x\|_2 < \infty\}$ equipped with $\langle \cdot, \cdot \rangle_2$. Note that re-writing (3), we have $f(z, z') \equiv \langle \phi(z), \phi(z') \rangle_2$, and since \mathcal{Z} is compact, under the continuity assumption A1 we have $\sup_{z \in \mathcal{Z}} \|\phi(z)\|_2 = \sup_{z \in \mathcal{Z}} f(z, z) < \infty$, hence \mathcal{M} is a subset of ℓ_2 . With $d_{\mathcal{M}}(x, x') := \|x - x'\|_2$, the pair $(\mathcal{M}, d_{\mathcal{M}})$ is a metric space.

We alert the reader to the fact that throughout section 4 we do *not* invoke assumption A3. However, we will need an assumption which was not present in the setting of theorem 1.

A4. *For each $z, z' \in \mathcal{Z}$ such that $z \neq z'$, there exists $\xi \in \mathcal{Z}$ such that $f(z, \xi) \neq f(z', \xi)$.*

Assumption A4 can be interpreted as a “distinguishability-by-correlations” condition. Indeed suppose that we ask if we can distinguish between two points $z, z' \in \mathcal{M}$ by comparing $f(z, \xi)$ to $f(z', \xi)$ for some “query” point ξ . A4 says that whenever $z \neq z'$ there exists a query point, possibly depending on z, z' , such that it is possible to distinguish z from z' in this sense.

Lemma 2. *Assume A1 and A4. Then ϕ is a homeomorphism between $(\mathcal{Z}, d_{\mathcal{Z}})$ and $(\mathcal{M}, d_{\mathcal{M}})$.*

Proof. We must show that ϕ is continuous, injective, and its inverse on \mathcal{M} is also continuous. The continuity of ϕ follows from the continuity of f under A1 combined with:

$$\begin{aligned} \|\phi(z) - \phi(z')\|_2^2 &= \|\phi(z)\|_2^2 + \|\phi(z')\|_2^2 - 2 \langle \phi(z), \phi(z') \rangle_2 \\ &= f(z, z) + f(z', z') - 2f(z, z'). \end{aligned}$$

We claim that A4 implies ϕ is injective. We prove the contrapositive to this claim. Suppose that ϕ is not injective. 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} , $f(z, \xi) = \langle \phi(z), \phi(\xi) \rangle_2 = \langle \phi(z'), \phi(\xi) \rangle_2 = f(z', \xi)$, which is the converse of A4. Finally, since $(\mathcal{Z}, d_{\mathcal{Z}})$ is a metric space, \mathcal{Z} is compact, $(\mathcal{M}, d_{\mathcal{M}})$ is a metric space and ϕ is continuous, the inverse of ϕ on \mathcal{M} must be continuous by a general result from the theory of metric spaces [Sutherland, 2009, Prop. 13.26]. \square

4.3 Geometric relationships between \mathcal{M} and \mathcal{Z}

Our next objective is to establish conditions under which the homeomorphic relationship between $(\mathcal{M}, d_{\mathcal{M}})$ and $(\mathcal{Z}, d_{\mathcal{Z}})$ from lemma 2 can be strengthened to diffeomorphism. This is important because it enables us to explain how the geometric shape of \mathcal{M} is related to that of \mathcal{Z} . We shall consider the following assumption, which is taken to hold throughout the remainder of section 4.3.

A5. For $d \leq \tilde{d}$, $\mathcal{Z} \subset \mathbb{R}^{\tilde{d}}$ is a d -dimensional differentiable manifold. There exists an open set $\tilde{\mathcal{Z}} \subset \mathbb{R}^{\tilde{d}}$ such that: $\mathcal{Z} \subset \tilde{\mathcal{Z}}$; the definition of f can be extended from $\mathcal{Z} \times \mathcal{Z}$ to $\tilde{\mathcal{Z}} \times \tilde{\mathcal{Z}}$; for all $1 \leq i, j \leq \tilde{d}$, the mixed partial derivative $\partial^2 f / \partial z_i \partial z'_j$ exists and is a symmetric, continuous function on $\tilde{\mathcal{Z}} \times \tilde{\mathcal{Z}}$; and for all $\xi \in \mathcal{Z}$, the matrix $\mathbf{H}_{\xi} \in \mathbb{R}^{\tilde{d} \times \tilde{d}}$ with elements:

$$(\mathbf{H}_{\xi})_{ij} := \left. \frac{\partial^2 f}{\partial z_i \partial z'_j} \right|_{(\xi, \xi)}, \quad \xi \in \mathcal{Z},$$

is positive-definite.

Also in the remainder of section 4.3, the Euclidean inner-product and norm on $\mathbb{R}^{\tilde{d}}$ will be denoted $\langle \cdot, \cdot \rangle_{\mathbb{R}^{\tilde{d}}}$ and $\|\cdot\|_{\mathbb{R}^{\tilde{d}}}$ in order to clarify the distinction between them and the inner-product and norm $\langle \cdot, \cdot \rangle_2$ and $\|\cdot\|_2$ on ℓ_2 .

To approach the question of diffeomorphism between \mathcal{M} and \mathcal{Z} we need to specify a notion of derivative for mappings between \mathcal{Z} and \mathcal{M} . We encounter the same issue described in section 4.1: even when A5 holds, \mathcal{Z} and \mathcal{M} are not, in general, linear subspaces of $\mathbb{R}^{\tilde{d}}$ and ℓ_2 respectively, so instead of directional derivatives we work with derivatives along curves.

For any two points $x, x' \in \mathcal{M}$, we call a function $\gamma : [0, 1] \rightarrow \mathcal{M}$ with $\gamma_0 = x$ and $\gamma_1 = x'$ a curve in \mathcal{M} with end-points x, x' : if: $t \mapsto \gamma_t$ is continuous in ℓ_2 , for each $t \in [0, 1]$ there exists $\dot{\gamma}_t \in \ell_2$ such that

$$\lim_{h \rightarrow 0} \left\| \frac{\gamma_{t+h} - \gamma_t}{h} - \dot{\gamma}_t \right\|_2 = 0,$$

and $t \mapsto \dot{\gamma}_t$ is continuous in ℓ_2 . Similarly for $z, z' \in \mathcal{Z}$ we call a function $\eta : [0, 1] \rightarrow \mathcal{Z}$ such that $\eta_0 = z$ and $\eta_1 = z'$ a curve in \mathcal{Z} with end-points z, z' if it continuously differentiable as a $\mathbb{R}^{\tilde{d}}$ -valued function in the sense of elementary Euclidean calculus.

Proposition 3. Assume A1, A4 and A5.

- i) If η is a curve in \mathcal{Z} , then $\gamma : [0, 1] \rightarrow \mathcal{M}$ defined by $t \mapsto \gamma_t := \phi(\eta_t)$ is a curve in \mathcal{M} .
- ii) If γ is a curve in \mathcal{M} , then $\eta : [0, 1] \rightarrow \mathcal{Z}$ defined by $t \mapsto \eta_t := \phi^{-1}(\gamma_t)$ is a curve in \mathcal{Z} .
- iii) If either η is a curve in \mathcal{Z} and $\gamma_t := \phi(\eta_t)$, or γ is a curve in \mathcal{M} and $\eta_t := \phi^{-1}(\gamma_t)$, then

$$\int_0^1 \|\dot{\gamma}_t\|_2 dt = \int_0^1 \langle \dot{\eta}_t, \mathbf{H}_{\eta_t} \dot{\eta}_t \rangle_{\mathbb{R}^{\tilde{d}}}^{1/2} dt. \quad (6)$$

The proof is in section C. Points i) and ii) of proposition 3 together tell us that ϕ is differentiable along any curve in \mathcal{Z} , and ϕ^{-1} is differentiable along any curve in \mathcal{M} . In that sense we can say that \mathcal{M} and \mathcal{Z} are diffeomorphic.

Part iii) establishes a geometric relationship between \mathcal{M} and \mathcal{Z} . The quantity on the l.h.s. of (6) is the length of the curve γ , see figure 4 for interpretation. The family of inner products $\{\langle \cdot, \mathbf{H}_{\xi} \cdot \rangle_{\mathbb{R}^{\tilde{d}}} ; \xi \in \mathcal{Z}\}$ constitute a Riemannian metric – a key concept in Riemannian geometry – see e.g. [Petersen, 2006, Ch.1 and p.121] for background. $\langle \cdot, \mathbf{H}_{\xi} \cdot \rangle_{\mathbb{R}^{\tilde{d}}}$ tells us how to compute inner-products locally at ξ , and thus \mathbf{H}_{ξ} can be understood as defining a local geometry. The identity in (6) tells us that the length of the curve γ in \mathcal{M} measured with respect to the $\langle \cdot, \cdot \rangle_2$ inner product, i.e., the l.h.s. of (6), is equal to the length of the

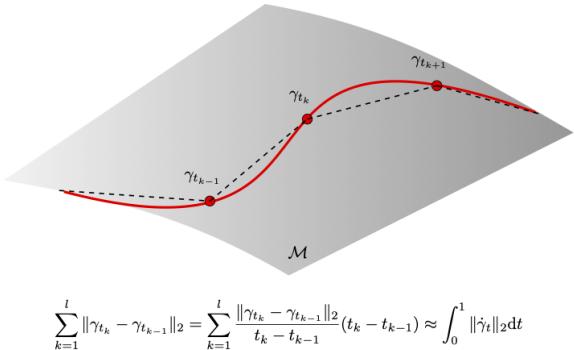


Figure 4: The length of a curve $\gamma : [0, 1] \rightarrow \mathcal{M}$ approximated by a sum of straight-line distances.

corresponding curve η in \mathcal{Z} measured with respect to the aforementioned Riemannian metric, rather than the usual Euclidean inner-product $\langle \cdot, \cdot \rangle_{\mathbb{R}^{\tilde{d}}}$.

If a particular functional form for f is assumed, then it may be possible to obtain a simplified expression for \mathbf{H}_ξ , check [A5](#) and use (6) to make even more precise statements about geometric relationships between \mathcal{M} and \mathcal{Z} . Here are some examples.

Translation invariant kernels. If $f(z, z') = g(z - z')$ for z, z' in an open neighbourhood of the diagonal $\mathcal{D} := \{(z, z') \in \mathcal{Z} \times \mathcal{Z} : z = z'\}$, where $g : \mathbb{R}^{\tilde{d}} \rightarrow \mathbb{R}$ is twice continuously differentiable, then \mathbf{H}_ξ is constant in ξ and given by the negative Hessian of g evaluated at the origin. The requirement in [A5](#) that \mathbf{H}_ξ is positive-definite for all $\xi \in \mathcal{Z}$ is then equivalent to requiring that g has a local maximum at the origin. If we denote by \mathbf{WSW}^\top the eigendecomposition of this negative Hessian, with eigenvalues on the diagonal of \mathbf{S} and orthonormal eigenvectors as the columns of \mathbf{W} , then (6) reduces to:

$$\int_0^1 \|\dot{\gamma}_t\|_2 dt = \int_0^1 \|\mathbf{S}^{1/2} \mathbf{W}^\top \dot{\eta}_t\|_{\mathbb{R}^{\tilde{d}}} dt.$$

In this situation we thus conclude that the length of the curve γ in \mathcal{M} is equal to the Euclidean length of the curve $t \mapsto \mathbf{S}^{1/2} \mathbf{W}^\top \dot{\eta}_t$ in $\mathbb{R}^{\tilde{d}}$, that is the curve η re-scaled by the square-root eigenvalues of the negative Hessian of g at the origin, in directions given by its eigenvectors.

Squared Euclidean distance kernels. If $f(z, z') = g(\|z - z'\|_{\mathbb{R}^{\tilde{d}}}^2)$ in an open neighbourhood of \mathcal{D} where $g : [0, \infty) \rightarrow \mathbb{R}$ is twice continuously differentiable, then $\mathbf{H}_\xi = -g'(0)\mathbf{I}_{\tilde{d}}$ for all $\xi \in \mathcal{Z}$. The positive-definiteness of \mathbf{H}_ξ is then equivalent to $g'(0) < 0$ and (6) reduces to:

$$\int_0^1 \|\dot{\gamma}_t\|_2 dt = \sqrt{-2g'(0)} \int_0^1 \|\dot{\eta}_t\|_{\mathbb{R}^{\tilde{d}}} dt.$$

In this situation, up to re-scaling by $\sqrt{-2g'(0)}$, we thus find isometry between \mathcal{M} and \mathcal{Z} .

Inner-product kernels on the sphere. If $f(z, z') = g(\langle z, z' \rangle_{\mathbb{R}^{\tilde{d}}})$ in an open neighbourhood of \mathcal{D} where $g : \mathbb{R} \rightarrow \mathbb{R}$, then $\mathbf{H}_\xi = g'(\|\xi\|_{\mathbb{R}^{\tilde{d}}}^2) \mathbf{I}_{\tilde{d}} + g''(\|\xi\|_{\mathbb{R}^{\tilde{d}}}^2) \xi \xi^\top$. If one specialises to the case of the sphere $\mathcal{Z} = \{z \in \mathbb{R}^{\tilde{d}}; \|z\|_{\mathbb{R}^{\tilde{d}}} = 1\}$, a result of [Kar and Karnick \[2012\]](#) states that any kernel of the form $f(z, z') = g(\langle z, z' \rangle_{\mathbb{R}^{\tilde{d}}})$ on this sphere is positive definite if and only if $g(x) = \sum_{n=0}^{\infty} a_n x^n$ for some nonnegative $(a_n)_{n \geq 0}$, hence we must have $g''(1) \geq 0$, and \mathbf{H}_ξ being positive definite for all $\xi \in \mathcal{Z}$ is equivalent to $g'(1) > 0$. Moreover for any curve η on the sphere we have $\eta_t^\top \dot{\eta}_t = 0$, and (6) reduces to:

$$\int_0^1 \|\dot{\gamma}_t\|_2 dt = g'(1)^{1/2} \int_0^1 \|\dot{\eta}_t\|_{\mathbb{R}^{\tilde{d}}} dt,$$

so there is isometry between \mathcal{M} and \mathcal{Z} , up to re-scaling by $g'(1)^{1/2}$.

Polynomial kernels. If $f(z, z') = (\langle z, z' \rangle + a)^b$ in an open neighbourhood of \mathcal{D} , where b is a positive integer and $a \geq 0$, and $\mathcal{Z} \subset \{z \in \mathbb{R}^{\tilde{d}} : |z| < a\}$, then \mathbf{H}_ξ is positive definite for any $\xi \in \mathcal{Z}$. In the particular case of $b = 2$, (6) becomes:

$$\int_0^1 \|\dot{\gamma}_t\|_2 dt = \sqrt{2} \int_0^1 \left[\|\eta_t\|_{\mathbb{R}^{\tilde{d}}}^2 + \frac{|\langle \eta_t, \dot{\eta}_t \rangle_{\mathbb{R}^{\tilde{d}}}|^2}{\|\dot{\eta}_t\|_{\mathbb{R}^{\tilde{d}}}^2} + a \right]^{1/2} \|\dot{\eta}_t\|_{\mathbb{R}^{\tilde{d}}} dt.$$

In this case there is evidently not isometry between \mathcal{M} and \mathcal{Z} ; the term $\|\eta_t\|_{\mathbb{R}^{\tilde{d}}}$ modulates the curve length depending on the size of the radial component of η , where as the term $|\langle \eta_t, \dot{\eta}_t \rangle|$ modulates the curve length depending on how much η deviates from being a circular arc around the origin, since for such an arc $\langle \eta_t, \dot{\eta}_t \rangle_{\mathbb{R}^{\tilde{d}}} = 0$.

Relation to the literature. The results of sections 4.2 and 4.3 are connected to some earlier observations. In the context of latent position models of random graphs, one of the authors of the present work, [Rubin-Delanchy \[2020\]](#), observed that when a kernel feature map with domain \mathbb{R}^d is Hölder continuous with exponent α , the image of \mathbb{R}^d through this feature map has Hausdorff dimension at most

d/α . Whiteley et al. [2021] further investigated this image set and proved that for two curves γ in \mathcal{M} and η in $\mathcal{Z} \subset \mathbb{R}^d$, if both are assumed continuously differentiable, then the identity (6) holds. The crucial difference between that result and proposition 3 above is that in the latter it is proved that continuous differentiability of η implies continuous differentiability of γ , and vice-versa. It is these two implications which together establish \mathcal{M} and \mathcal{Z} are diffeomorphic: that is absent in [Whiteley et al., 2021].

Other differences include that Whiteley et al. [2021, Props. 1-3.] considered translation invariant and squared Euclidean distance kernels in the case where \mathcal{Z} is convex – that requirement is not present here, we are working under the generality of assumption A5; Whiteley et al. [2021] addressed the case where the functional forms of the kernels in question hold for all $z, z' \in \mathcal{Z}$ – this is unnecessary, all that matters is the behaviour of f in a neighbourhood of the diagonal \mathcal{Z} ; Whiteley et al. [2021] did not consider the special case of polynomial kernels.

5 Discussion

5.1 Dependence structure of the model and outline of proof of theorem 1

Under the Latent Metric Space model, the n rows of the noise-free data matrix $\mathbf{Y} - \sigma\mathbf{E}$ are exchangeable, with dependence across rows reflected in f . On the other hand, the independence of the fields X_1, \dots, X_p implies the p columns of $\mathbf{Y} - \sigma\mathbf{E}$ are conditionally independent given Z_1, \dots, Z_n , and moreover these columns are not exchangeable in general. This structure is in contrast to the standard assumptions underlying much of high-dimensional statistical theory in which n data points are independent and identically distributed, with dependence across dimensions for each data point. For example, the asymptotic behaviour of PC scores relative to their population counterparts for independent and identically distributed data has been studied under spiked covariance models in various asymptotic regimes [Lee et al., 2010, Yata and Aoshima, 2009, 2012, Shen et al., 2012, 2013, Hellton and Thoresen, 2017]. Hellton and Thoresen [2017] give an insightful discussion of how, for these models, asymptotic behaviour of PC scores is connected to the behaviour of sample eigenvalues and eigenvectors in PCA, as investigated in the seminal works Paul [2007], Johnstone and Lu [2009], Jung and Marron [2009]. However there is no latent structure nor emergent manifold structure in these spiked covariance models. We also note that theorem 1 establishes uniform concentration across the set of PC scores ζ_1, \dots, ζ_n . The results about PC scores in Lee et al. [2010], Yata and Aoshima [2009, 2012], Shen et al. [2012, 2013], Hellton and Thoresen [2017] are not uniform in nature.

Why does the dependence structure of the Latent Metric Space model lead to concentration of the PC scores? To help explain it is useful to consider an elementary fact from linear algebra, in lemma 4. Recall from section 2.2 that the PC scores are defined by $[\zeta_1 | \dots | \zeta_n]^\top := \mathbf{Y}\mathbf{V}_\mathbf{Y}$, where the columns of $\mathbf{V}_\mathbf{Y} \in \mathbb{R}^{p \times r}$ are orthonormal eigenvectors associated with the r largest eigenvalues of $\mathbf{Y}^\top \mathbf{Y}$.

Lemma 4. *On the event that the rank of $\mathbf{Y}^\top \mathbf{Y}$ is at least r , $p^{-1/2}\mathbf{Y}\mathbf{V}_\mathbf{Y} = \mathbf{U}_\mathbf{Y}\Lambda_\mathbf{Y}^{1/2}$, where $\Lambda_\mathbf{Y} \in \mathbb{R}^{r \times r}$ is the diagonal matrix whose diagonal elements are the r largest eigenvalues of $p^{-1}\mathbf{Y}\mathbf{Y}^\top$, and the columns of $\mathbf{U}_\mathbf{Y} \in \mathbb{R}^{n \times r}$ are orthonormal eigenvectors associated with these eigenvalues.*

Proof. Apply lemma 10. □

Thus by computing the PC scores ζ_1, \dots, ζ_n and rescaling by $p^{-1/2}$, we are, in effect, computing the n rows of $\mathbf{U}_\mathbf{Y}\Lambda_\mathbf{Y}^{1/2}$, where $\mathbf{U}_\mathbf{Y}\Lambda_\mathbf{Y}^{1/2}(\mathbf{U}_\mathbf{Y}\Lambda_\mathbf{Y}^{1/2})^\top = \mathbf{U}_\mathbf{Y}\Lambda_\mathbf{Y}\mathbf{U}_\mathbf{Y}^\top$ is a rank- r approximation to $p^{-1}\mathbf{Y}\mathbf{Y}^\top$. The connection between $p^{-1}\mathbf{Y}\mathbf{Y}^\top$ and ϕ is given by:

Lemma 5. *Assume A1 and A3. Then $p^{-1}\mathbb{E}[\mathbf{Y}\mathbf{Y}^\top | Z_1, \dots, Z_n] = \Phi\Phi^\top + \sigma^2\mathbf{I}_n$, where $\Phi := [\phi(Z_1) | \dots | \phi(Z_n)]^\top \in \mathbb{R}^{n \times r}$.*

Proof. Under A1 and A3, ϕ is well-defined as a length- r vector. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be the matrix with entries $\mathbf{X}_{ij} := X_j(Z_i)$. 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:

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

□

In fact lemma 4 and 5 together form the starting point for the proof of theorem 1. In the Latent Metric Space model the random fields X_1, \dots, X_p are independent, although not necessarily identically distributed. As a consequence of this independence, together with the independence across the columns of \mathbf{E} , $p^{-1}\mathbf{Y}\mathbf{Y}^\top$ can be written as an average of p conditionally independent rank-1 matrices. This fact is exploited in the proof of theorem 1 to show that $p^{-1}\mathbf{Y}\mathbf{Y}^\top$ is in a suitable sense concentrated about its conditional expectation $\Phi\Phi^\top + \sigma^2\mathbf{I}_n$. If one were to relax the independence of X_1, \dots, X_p in a way which still allows this concentration to be established, then the overall strategy of the proof of theorem 1 would still be applicable. This might be particularly relevant in situations where the n rows of \mathbf{Y} are, for example, time series of length p , or where each row of \mathbf{Y} corresponds to an image or some form of geo-spatially indexed measurement, where dependence may model smoothness.

To pass from concentration of $p^{-1}\mathbf{Y}\mathbf{Y}^\top$ about $\Phi\Phi^\top + \sigma^2\mathbf{I}_n$ to concentration of the PC scores, we rely heavily on certain matrix decomposition techniques used by Lyzinski et al. [2016] in the study of spectral embedding of random graphs under a random dot product model. These techniques are often used to justify spectral clustering under the stochastic block model [Abbe, 2017], showing that exact recovery is possible in sufficiently dense graphs. The uniform 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 $2 \rightarrow \infty$ 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.

5.2 The case of discrete \mathcal{Z} , mixture models and clustering

Our next objective is to begin exploring instances of the Latent Metric Space model and demonstrate different forms of embedded structure which they give rise to. Consider the case where \mathcal{Z} has finitely many elements, say $\mathcal{Z} = \{1, \dots, m\}$. For the following discussion it is not important that we take these elements to be the numbers $1, \dots, m$, any m distinct abstract elements will do. In this situation the Latent Metric Space 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 associated with $z \in \mathcal{Z}$, and the latent variable Z_i indicates which mixture component the i th row of the data matrix \mathbf{Y} is drawn from. The simple form of the noise in the Latent Metric Space 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. We leave more general assumptions on \mathbf{E} for future investigation.

To make \mathcal{Z} into a metric space we may consider the discrete metric as was discussed in section 4.1, that is $d_{\mathcal{Z}}(z, z') := 0$ for $z = z'$, otherwise $d_{\mathcal{Z}}(z, z') := 1$. The kernel f is specified by the matrix $\mathbf{F} \in \mathbb{R}^{m \times m}$ with entries

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

In this situation **A1** and **A3** 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 $\{\phi(1), \dots, \phi(m)\}$, each associated with exactly one element of \mathcal{Z} . If such topological equivalence were to hold then theorem 1 would tell us that the PC scores will be clustered around the m distinct points $\{\mathbf{Q}^{-1}\phi(1), \dots, \mathbf{Q}^{-1}\phi(m)\}$, with specifically $p^{-1/2}\zeta_i$ being close to $\mathbf{Q}^{-1}\phi(Z_i)$.

To verify topological equivalence it remains to check **A4** 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 $k, l \in \{1, \dots, m\}$ such that $k \neq l$, there must exist some $\xi \in \{1, \dots, m\}$ such that $f(k, \xi) = \mathbf{F}_{k\xi} \neq \mathbf{F}_{l\xi} = f(l, \xi)$. Thus assumption **A4** 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 PC scores and then apply some clustering technique to those PC scores. 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 PC scores 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].

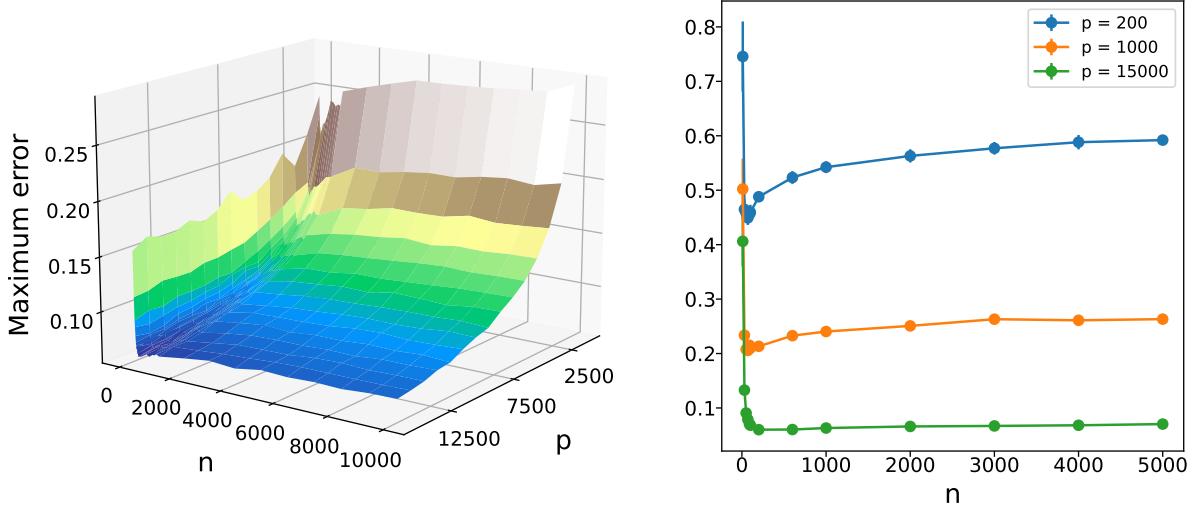


Figure 5: Mixture model example. Left: maximum error $\max_{i \neq j} |p^{-1/2} \|\zeta_i - \zeta_j\|_2 - \|\phi(Z_i) - \phi(Z_j)\|_2|$, averaged over 50 independent realisations from the model, as a function of n and p . Right: the same error for $p = 200, 1000, 15000$, as a function of n .

To illustrate the behaviour of the Latent Metric Space model and PC scores 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 $j = 1, \dots, p$, $[X_j(1) \ X_j(2) \ X_j(3)]^\top \sim \mathcal{N}(\mathbf{0}, \Sigma)$ where Σ is full-rank; and the elements of \mathbf{E} are independent and identically distributed $\mathcal{N}(0, 1)$ with $\sigma = 1$. Figure 5 shows the error on the left hand side of the inequality (5), 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 the theorem 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 5. We see that for fixed p , as n increases the error initially quickly decreases, but then slowly increases and appears to stabilise about some value, even when $n \gg p$. We also see that this stable value decreases as p increases.

Figure 6 illustrates how this error performance relates to the clustering of the PC scores. When n is fixed, we see that as p increases the PC scores are increasingly tightly clustered around $\phi(1), \phi(2), \phi(3)$, in keeping with the concentration in theorem 1. When p is fixed, we see that three clusters of PC scores 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 PC scores still convey the topological or geometric structure of \mathcal{M} and hence \mathcal{Z} .

5.3 Gaussian Process Latent Variable models, Bayesian inference and MCMC

We now turn to another instance of the Latent Metric Space model. Lawrence [2003] and Lawrence and Hyvärinen [2005] proposed a likelihood-based nonlinear dimension reduction technique using a Gaussian Process Latent Variable model (GPLVM). Under a GPLVM, a data matrix $\mathbf{Y} = [Y_1 | \dots | Y_p] \in \mathbb{R}^{n \times p}$ follows an additive model:

$$\mathbf{Y} = \tilde{\mathbf{X}}\mathbf{W} + \sigma\mathbf{E}$$

where $\mathbf{W} \equiv [W_1 | \dots | W_p] \in \mathbb{R}^{r \times p}$, $\tilde{\mathbf{X}} \equiv [\tilde{X}_1 | \dots | \tilde{X}_n]^\top \in \mathbb{R}^{n \times r}$ (the tilde here distinguishes these \tilde{X}_i 's from the X_j 's in the Latent Metric Space model) and the elements of $\mathbf{E} \in \mathbb{R}^{n \times p}$ are independent and identically distributed $\mathcal{N}(0, 1)$. Lawrence and Hyvärinen [2005] integrate out the matrix \mathbf{W} under a Gaussian prior $p(\mathbf{W}) := \prod_{j=1}^p p(w_j)$, $p(w_j) := \mathcal{N}(\mathbf{0}_r, \mathbf{I}_r)$, yielding a Gaussian density which factorises

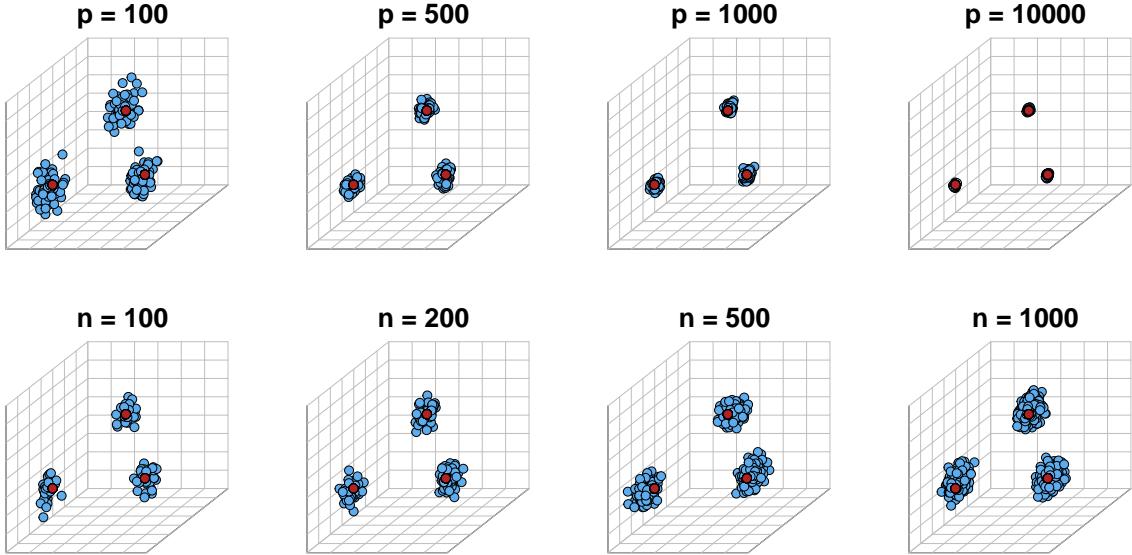


Figure 6: Mixture model example. PC scores $\{p^{-1/2}\zeta_1, \dots, p^{-1/2}\zeta_n\}$ (blue dots) and $\phi(1), \phi(2), \phi(3)$ (red dots). Top row: n fixed to 200 and p varying. Bottom row p fixed to 200 and n varying.

over the p columns of \mathbf{Y} :

$$p(\mathbf{Y}|\tilde{\mathbf{X}}, \sigma^2) = \prod_{j=1}^p p(y_j|\tilde{\mathbf{X}}, \sigma^2), \quad p(y_j|\tilde{\mathbf{X}}, \sigma^2) = \mathcal{N}(\mathbf{0}_n, \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top + \sigma^2\mathbf{I}_n).$$

The authors then take the matrix $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top$ to be the Gram matrix for some kernel κ evaluated at latent positions $Z_1, \dots, Z_n \in \mathbb{R}^d$, i.e., $(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top)_{ij} = \kappa(Z_i, Z_j)$.

To see that this is an instance of the Latent Metric Space model we take $\mathcal{Z} \subseteq \mathbb{R}^d$ and take X_1, \dots, X_p to be independent and identically distributed, zero-mean Gaussian processes with common covariance function κ . Then f and $[X_j(Z_1) \cdots X_j(Z_n)]^\top$ in the Latent Metric Space model are respectively identical to κ and the j th column of $\tilde{\mathbf{X}}\mathbf{W}$ in the GPLVM, and the matrix product $\tilde{\mathbf{X}}\mathbf{W}$ in the GPVLM can be interpreted as a Karhunen–Loëve expansion of each of the Gaussian processes X_1, \dots, X_p , evaluated at Z_1, \dots, Z_n .

Under the assumption that the kernel κ belongs to a given parametric family, e.g., a radial basis function kernel, Lawrence and Hyvärinen [2005] proposed maximum a-posteriori estimation of Z_1, \dots, Z_n by using a gradient method to jointly optimise the likelihood combined with a prior on $\tilde{\mathbf{X}}$, with respect to the variables Z_1, \dots, Z_n , parameters of the kernel and σ^2 . This leads to an algorithm with undesirable $O(n^3)$ complexity, which the authors circumvent using an active-set technique.

Given that the likelihood function of GPLVM can be evaluated given Z_1, \dots, Z_n , σ^2 and any kernel parameters, it is natural to approach model assessment and inference for σ^2 and kernel parameters by trying to integrate out Z_1, \dots, Z_n under a prior, and evaluate the resulting marginal likelihood. This integration is analytically intractable in general. Titsias and Lawrence [2010] proposed approximate inference using variational methods, again under the assumption the kernel belongs to a parametric family, leading to a computable lower bound on the marginal likelihood. Over the last fifteen or so years a significant amount of research has been devoted to developing Markov chain Monte Carlo (MCMC) methods to facilitate exact Bayesian inference in models with latent variables and intractable likelihoods, notably Pseudo-marginal MCMC [Beaumont, 2003, Andrieu and Roberts, 2009]. Designing such methods for GPLVM's has been considered only quite recently, by Gadd et al. [2021], who points out the opportunities for further computational speed-up which might potentially arise from using Gaussian process approximations [Drovandi et al., 2018] and parallelised implementation.

We also note that Lawrence [2012] derived a Gaussian Markov random field model related to a GPLVM through which Locally Linear Embedding [Roweis and Saul, 2000] has a statistical interpretation. In this model data dimensions are independent and identically distributed. The associated “blessing of dimensionality” which Lawrence [2012] discusses — essentially averaging over dimensions — is similar

in spirit to the concentration behaviour of PC scores we prove, although in our context the dependence structure of the model is more complex and, crucially, we explain how this leads to emergent manifold structure. The utility of this explanation is that, in contrast to [Lawrence, 2003, Lawrence and Hyvärinen, 2005, Titsias and Lawrence, 2010], it enables us to approach some inference tasks in the Latent Metric Space model without having to assume a parametric form for the kernel, by exploiting what we know about topological and geometric relationships between \mathcal{M} and \mathcal{Z} . In the next sections we explore these inference tasks, specifically topological data analysis, nonlinear dimension reduction and manifold estimation.

5.4 Topological Data Analysis

Persistent homology techniques [Edelsbrunner et al., 2008, Carlsson, 2009, Chazal and Michel, 2021] are designed to estimate topological features of a set, such as number of connected components, numbers of holes, cavities, etc. on the basis of a cloud of points sampled from a distribution supported on that set. Theorem 1 combined with lemma 2 indicates that by applying such persistent homology techniques to the point cloud $\{p^{-1/2}\zeta_i; i = 1, \dots, n\}$ we can estimate topological features of the latent domain \mathcal{Z} . Indeed, theorem 1 tells us that up to an orthogonal transformation (which is an isometry, hence preserves topological features) $\{p^{-1/2}\zeta_i; i = 1, \dots, n\}$ are uniformly concentrated around $\{\phi(Z_i); i = 1, \dots, n\}$; the points $\{\phi(Z_i); i = 1, \dots, n\}$ are independent and identically distributed, with common distribution whose support is \mathcal{M} ; and by lemma 2 $(\mathcal{M}, d_{\mathcal{M}})$ is topologically equivalent to $(\mathcal{Z}, d_{\mathcal{Z}})$. Thus the Latent Metric Space model is a use-case for persistent homology techniques.

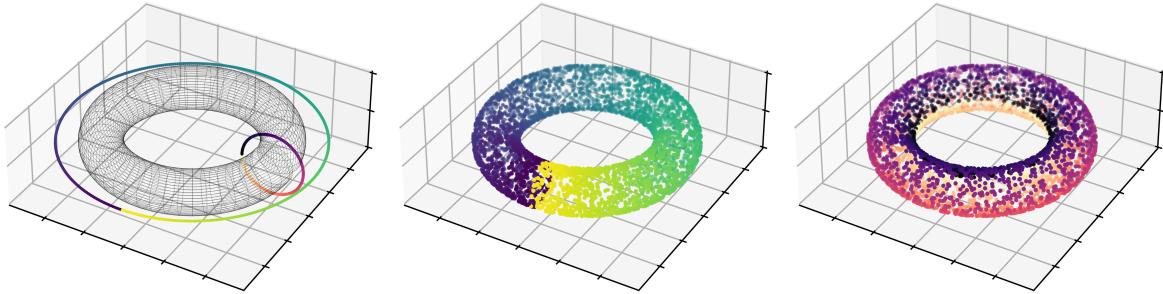


Figure 7: Left: grey wireframe of \mathcal{Z} , a torus, with colour bars indicating coordinates with respect to two circles. Both the middle and right plots show the same $n = 4000$ points, which are sampled uniformly on the torus, coloured by their coordinates with respect to each of the two circles.

As an example we consider the case \mathcal{Z} is a torus embedded in \mathbb{R}^3 and μ is the uniform distribution on this torus. Latent variables Z_1, \dots, Z_{4000} simulated from μ are shown in figure 7. 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: the angle around the circle in the horizontal plane is called the *azimuth*, the other is called the *elevation* angle. We assume all the fields X_j are equal in distribution, with $\mathbb{E}[X_j(z)] = 0$ for all $z \in \mathcal{Z}$, so that $f(z, z') = \mathbb{E}[X_j(z)X_j(z')]$ for $j = 1, \dots, p$. Furthermore we take the fields to be Gaussian processes with a radial basis function kernel as their common covariance kernel, that is $f(z, z') = \exp(-\|z - z'\|_{\mathbb{R}^3}^2)$, which satisfies A1 and A4, hence by lemma 2, \mathcal{M} is topologically equivalent to \mathcal{Z} . A3 does not hold.

Figure 8 shows the first 9 dimensions of the PC scores, obtained with $p = 500$ and $\sigma^2 = 0$; we choose this latter setting in order for figure 8 to be visually clear, noisier settings are considered in section 5.7. The only difference between the two rows of plots in figure 8 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 7, similarly the colouring in the bottom row matches the plot on the right of figure 7.

The first 6 dimensions of the PC scores noticeably reflect the azimuth angle, whilst the 7th to 9th dimensions noticeably reflect the elevation angle. To investigate this from a topological point of view, we applied the `ripser.py` package [Tralie et al., 2018] in Python to compute persistence diagrams associated with the set of latent points Z_1, \dots, Z_n and with the PC scores, see figure 9. Recall that a point (x, y) in a persistence diagram signifies a topological feature born at scale x and persisting until scale y . Figure 9 shows such points for connected components (Dim 0) and 1-dimensional holes (Dim 1). For \mathcal{Z} the true number of connected components is 1, and the true number of 1-dimensional holes is 2, corresponding to the two circles highlighted on the left of figure 7. This is reflected in the left-most persistence diagram

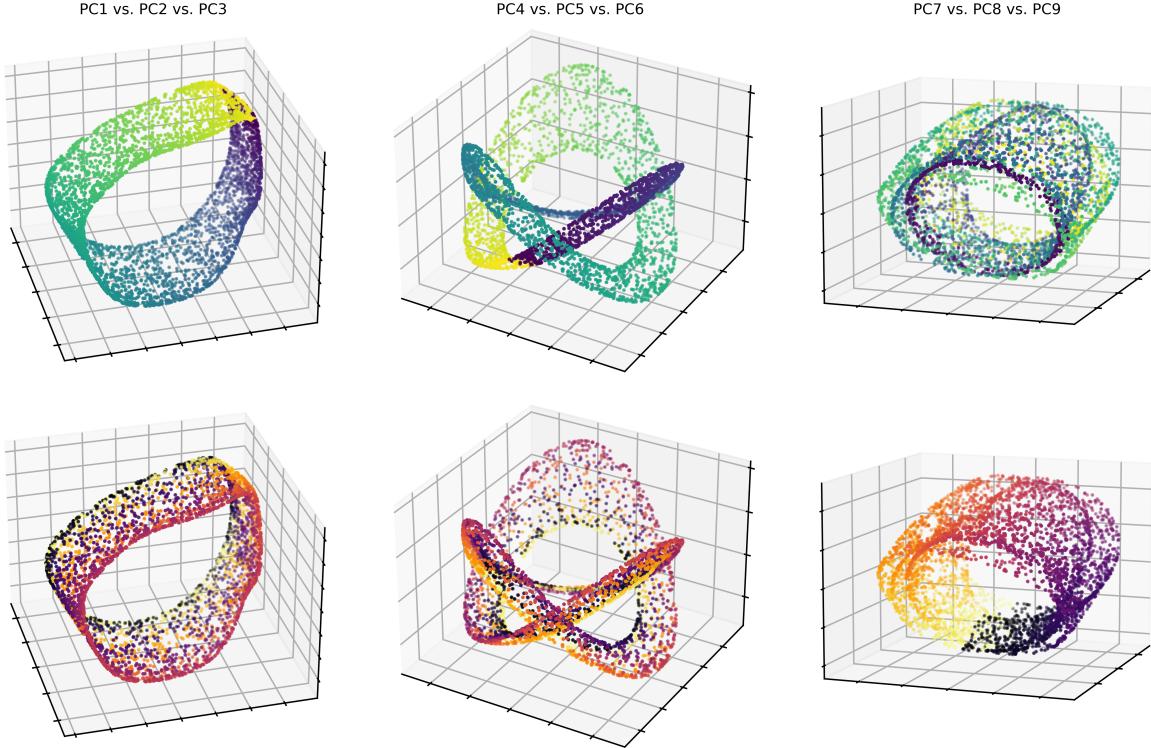


Figure 8: Torus example. Both the top and bottom rows show the first 9 dimensions of the PC scores $\{p^{-1/2}\zeta_1, \dots, p^{-1/2}\zeta_n\}$. In each row, points are coloured according to the coordinates of the underlying points $\{Z_1, \dots, Z_n\}$ with respect to the two circles shown in figure 7. Numerical scales are omitted to de-clutter the plots.

in figure 9. Observe the resemblance between the left-most and right-most persistence diagrams, that is the dimension-20 PC scores faithfully convey both of the two 1-dimensional holes that are present in \mathcal{Z} ; whilst the middle persistence diagram shows that the dimension-3 PC scores erroneously convey only 1 rather than 2 holes. We shall return to topic of how to choose the dimension of the PC scores in section 5.8.

We also note that some TDA techniques explicitly assume data are distributed on a manifold, including the homology estimation methods [Niyogi et al., 2008, Balakrishnan et al., 2012] and persistence-based clustering methods [Chazal et al., 2013]. If PC scores are taken as input to these methods then theorem 1 combined with the results of section 1 could be used to verify a relaxed version of this assumption: being uniformly ϵ -close to the manifold \mathcal{M} , with high probability.

5.5 Linear then nonlinear dimension reduction

Our next objective is to demonstrate how proposition 3 allows us to relate the geometry of \mathcal{M} to that of \mathcal{Z} . Continuing with the example from section 5.4 where \mathcal{Z} is a torus and $f(z, z') = \exp(-\|z - z'\|_{\mathbb{R}^3}^2)$, A5 holds with $d = 2$ and $\tilde{d} = 3$, and $\mathbf{H}_\xi = \sqrt{2}\mathbf{I}_3$ for all ξ , hence part iii) of proposition 3 tells us that for any curve η in \mathcal{Z} and $\gamma : [0, 1] \rightarrow \mathcal{M}$ defined by $\gamma_t := \phi(\eta_t)$,

$$\int_0^1 \|\dot{\gamma}_t\| dt = \sqrt{2} \int_0^1 \|\dot{\eta}_t\| dt. \quad (7)$$

Thus, up to a factor of $\sqrt{2}$, there is isometry between \mathcal{M} and \mathcal{Z} . If we fix the end-points of the curve η , say $a = \eta_0$, $b = \eta_1$, and then take the infimum in (7) over all curves in \mathcal{Z} with these end-points, we find that the geodesic distance in \mathcal{Z} between a and b , i.e. this infimum, is equal to geodesic distance in \mathcal{M} between $\phi(a)$ and $\phi(b)$, up to a factor of $\sqrt{2}$.

This theoretical relationship between geodesic distances in \mathcal{Z} and in \mathcal{M} is shown by the red line in the left plot of figure 10. In the same plot the blue points are estimated geodesic distances, pairwise amongst

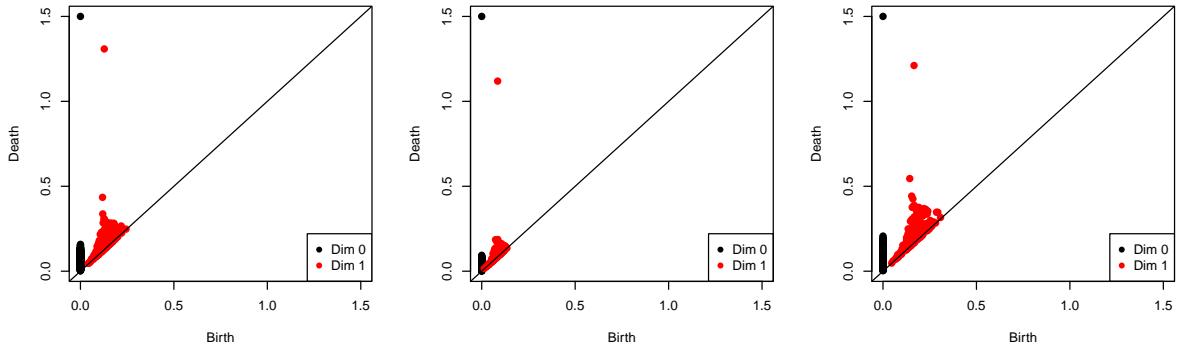


Figure 9: Topological data analysis for the example in which \mathcal{Z} is a torus. Persistence diagrams for connected components (Dim 0) and 1-dimensional holes (Dim 1) computed from: (left) $\{Z_1, \dots, Z_n\}$, (middle) the dimension-3 PC scores, (right) the dimension-20 PC scores. In all three analyses the maximum death scale considered was 1.5, and in each plot the black dot at coordinates $(0.0, 1.5)$ corresponds to a connected component which persisted to this maximum death scale.

the simulated points $\{Z_1, \dots, Z_n\}$ in \mathcal{Z} , plotted against the corresponding estimated geodesic distances pairwise amongst the dimension-20 PC scores $\{p^{-1/2}\zeta_1, \dots, p^{-1/2}\zeta_n\}$. These geodesic distances were estimated using part of the functionality of `sklearn.manifold.Isomap` in Python: for each of the two point clouds $\{Z_1, \dots, Z_n\}$ and $\{p^{-1/2}\zeta_1, \dots, p^{-1/2}\zeta_n\}$, this procedure constructs a k -nearest neighbour graph with edges weighted by Euclidean distances (in respectively \mathbb{R}^3 and \mathbb{R}^{20}), and then returns graph distances in these weighted graphs as estimates of geodesic distances. We took the default parameter setting in `sklearn.manifold.Isomap` [Pedregosa et al., 2011] of $k = 5$. The left plot of figure 10 shows a good fit of these estimated distances with the theoretical relationship.

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 24 thousand 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.

Pre-processing data by computing PC scores, before applying nonlinear dimension reduction, has been advocated in the literature. For example, 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 made by Van Der Maaten [2014] and Saul [2020]. Up until now however, there has been no detailed or rigorous statistical justification for this pre-processing. Theorem 1 along with the contents of section 4 fill that gap. 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, *linear then nonlinear dimensionality reduction*, may be particularly effective.

The two plots in the middle and on the right of figure 10 show the output of t-SNE applied to the dimension-20 PC scores $\{p^{-1/2}\zeta_1, \dots, p^{-1/2}\zeta_n\}$, using `scikit-learn` in Python, with perplexity 200 and 500 iterations, reducing the dimension to 3. Evidently t-SNE is able to reconstruct the torus structure of \mathcal{Z} .

Theoretical guarantees for nonlinear dimension reduction is a topic of active research. Building from Bernstein et al. [2000], Trosset and Buyukbas [2020] proved that Isomap asymptotically recovers Rie-

Algorithm 1 Linear then nonlinear dimensionality reduction.

Input: data matrix $\mathbf{Y} \in \mathbb{R}^{n \times p}$.

- 1: Compute the r -dimensional PC scores $p^{-1/2}\zeta_1, \dots, p^{-1/2}\zeta_n$, to reduce data from high dimension p to moderate dimension r .
- 2: Apply a nonlinear dimension reduction technique to $p^{-1/2}\zeta_1, \dots, p^{-1/2}\zeta_n$, yielding a low-dimensional representation $\hat{Z}_1, \dots, \hat{Z}_n$.

Output: $\hat{Z}_1, \dots, \hat{Z}_n$.

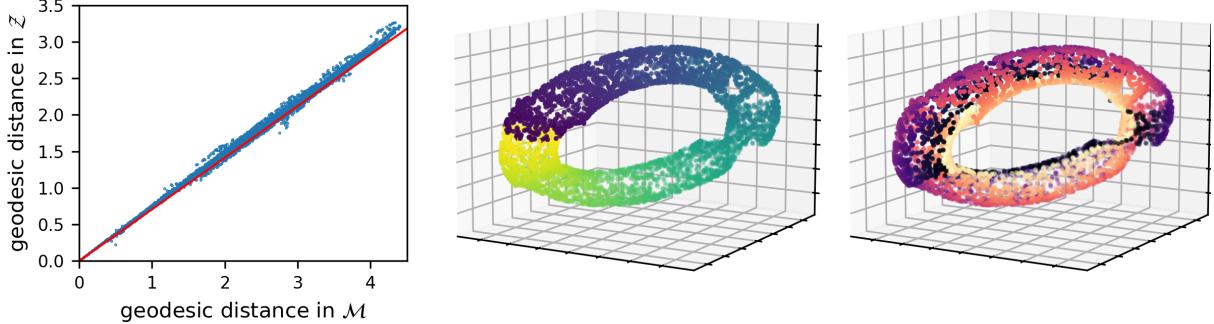


Figure 10: Dimensionality reduction for the torus example. Left: theoretical relationship between geodesic distance in \mathcal{Z} and \mathcal{M} (red line) and estimated geodesic distances in these two domains, obtained from $\{Z_1, \dots, Z_n\}$ and $\{p^{-1/2}\zeta_1, \dots, p^{-1/2}\zeta_n\}$ (blue points). Middle and right: output from t -SNE applied to the dimension-20 PC scores $\{p^{-1/2}\zeta_1, \dots, p^{-1/2}\zeta_n\}$, coloured according to the azimuth and elevation angles of the corresponding points $\{Z_1, \dots, Z_n\}$ as in figure 7.

mannian distances when input data are supported on a compact, connected Riemannian manifold, whilst [Belkin and Niyogi \[2006\]](#) analyzed Laplacian Eigenmaps assuming input data are uniformly distributed on a Riemannian sub-manifold. As such these methods seem naturally suited to processing of the PC scores in the setting of [A5](#) and the results of the present work could be used to validate the assumptions made by the aforementioned authors.

Theoretical understanding of t -SNE is far from complete [[Cai and Ma, 2021](#)]. Analyses such as those of [Arora et al. \[2018\]](#), [Linderman and Steinerberger \[2019\]](#) assume data are partitioned into some number of latent clusters, which could be interpreted as a topological assumption about the support of the data distribution, but there appear to be no existing theoretical results which could give a detailed explanation of the performance of t -SNE evident in (10). Thus rigorous explanation of how t -SNE can reconstruct latent differentiable manifolds, as can arise from the Latent Metric Space model, is an open problem.

5.6 Manifold estimation

Instead of exploiting manifold structure implicitly for dimension reduction, we might seek an explicit estimate of \mathcal{M} . There are many existing theoretical and methodological perspectives on this problem, see [[Wasserman, 2018](#)] for a survey. For example, assuming data are noisy observations of a Riemannian manifold, the minimax rate for manifold estimation was studied by [Genovese et al. \[2012b\]](#) and Hausdorff distance risk bounds were obtained by [Genovese et al. \[2012a\]](#). Hypothesis testing for the presence of a manifold was addressed by [Fefferman et al. \[2016\]](#).

These ‘worst-case’ analyses are sometimes discouraging: [Wasserman \[2018\]](#) concludes from [Genovese et al. \[2012a\]](#) that ‘estimating [the manifold] is hopeless’. However, the papers assume that the noise distribution is fixed, whereas in theorem 1, applying PCA under the latent metric space model with $p/n \rightarrow \infty$ results in a point cloud with *vanishing* off-manifold noise. We might hope for a much better rate in this regime, e.g., approaching the noiseless case [[Genovese et al., 2012a, Theorem 2](#)].

Even before estimating \mathcal{M} , we may simply be interested in estimating its dimension. This problem is known as intrinsic dimension estimation, and a wide variety of methods are available, for instance, [[Kégl, 2002](#), [Costa and Hero, 2004](#), [Hein and Audibert, 2005](#), [Levina and Bickel, 2004](#), [Little, 2011](#), [Lombardi et al., 2011](#)], many implemented in the R package ‘intrinsicDimension’. Again in the setting of [A5](#), the Latent Metric Space model and theorem 1 could be used to validate many of the assumptions involved in these works.

5.7 Regression and classification

We now consider how the Latent Metric Space model can fit in to supervised learning tasks. Suppose that in addition to \mathbf{Y} , one observes response variables modelled as $m(Z_1) + W_1, \dots, m(Z_{n-1}) + W_{n-1}$ where m is some unknown function, the W_i are noise disturbances, and one seeks to predict $m(Z_n)$. Consider the procedure in algorithm 2.

Algorithm 2 Regression on \mathcal{M}

Input: data matrix $\mathbf{Y} \in \mathbb{R}^{n \times p}$ and responses $m(Z_1) + W_1, \dots, m(Z_{n-1}) + W_{n-1}$

- 1: Compute the r -dimensional PC scores ζ_1, \dots, ζ_n from \mathbf{Y} , to reduce data from high-dimension p to dimension r .
- 2: Fit a regression model with responses $m(Z_1) + W_1, \dots, m(Z_{n-1}) + W_{n-1}$, and the PC scores $\zeta_1, \dots, \zeta_{n-1}$ as predictors.
- 3: Evaluate the regression model at ζ_n to yield a prediction $\widehat{m(Z_n)}$ of $m(Z_n)$.

Output: Prediction $\widehat{m(Z_n)}$.

To see the motivation for this procedure consider the trivial identity:

$$m(Z_i) = m \circ \phi^{-1} \circ \phi(Z_i).$$

Thus $m(Z_i)$ can be viewed as an evaluation of the function $m \circ \phi^{-1}$ at the point $\phi(Z_i)$ which lies in \mathcal{M} . By theorem 1, up to an orthogonal transformation, the points $\phi(Z_1), \dots, \phi(Z_n)$ are uniformly well approximated by the PC scores $p^{-1/2}\zeta_1, \dots, p^{-1/2}\zeta_n$. Hence we may regard algorithm 2 as solving a regression problem in which the unknown function to be approximated is $m \circ \phi^{-1}$, with $p^{-1/2}\zeta_1, \dots, p^{-1/2}\zeta_n$ approximating $\phi(Z_1), \dots, \phi(Z_n)$ as predictors.

What specific kind of regression model or technique should one use in this situation? Under the assumption that predictors lie on an unknown manifold, various forms of regression and classification techniques have been proposed and studied [Lafferty and Wasserman, 2007, Bickel and Li, 2007, Cheng and Wu, 2013, Yang and Dunson, 2016, Moscovich et al., 2017, Lin et al., 2019, Niu et al., 2019]. The statistical performance of such methods often depends on the dimension of the manifold, rather than the dimension of the ambient space in which the predictors are valued. For example, Yang and Dunson [2016] showed that for a Gaussian process regression technique, the posterior contraction rate is largely driven by the dimension of the manifold. Moscovich et al. [2017] showed that k -nearest neighbour regression with neighbourhoods defined by estimated geodesic distance can achieve the optimal finite-sample minimax bound on the mean squared error, as if the manifold were known, with rate depending only the dimension of the manifold. This kind of result contrasts with a common informal misconception that nearest neighbour methods perform poorly when predictors are high-dimensional because in that regime neighbours tend to be far apart, e.g. [Hastie et al., 2009, Sec. 3.5]. However, that tendency is far from automatic (Beyer et al. [1999] provides rigorous sufficient conditions) and indeed seems to be ruled out by the assumptions of Lafferty and Wasserman [2007], Bickel and Li [2007], Cheng and Wu [2013], Yang and Dunson [2016], Moscovich et al. [2017], Lin et al. [2019], Niu et al. [2019] that predictors lie on a manifold.

In order to implement algorithm 2 one needs to choose the dimension of the PC scores, r . We performed numerical experiments to explore the impact of r on regression and classification performance in six scenarios. Our objective here is to give illustrative results for a simple and generic choice of regression/classification technique with default parameter settings, rather than more sophisticated alternatives. Pursuant to this objective we used simple k -nearest neighbour regression/classification with default settings in scikit-learn [Pedregosa et al., 2011]: $k = 5$, Euclidean distance function and uniform weighting of neighbours. We note that this is a much less sophisticated approach than e.g. Moscovich et al. [2017], we do not estimate geodesic distances, and it is reasonable to expect that improved regression/classification performance could be obtained using approaches which are specifically designed to exploit manifold structure.

Regression examples

- *Torus model.* \mathbf{Y} simulated from the model in section 5.4, with $n = 1000$, $p = 500$, $\sigma = 1$. The response variables are bivariate and taken to be the azimuth and elevation angles for each Z_i , cf. figure 7.

- *Wi-Fi localisation.* Predictors are received signal strengths from $p = 520$ Wi-Fi routers, recorded at $n = 2000$ locations across three buildings at Universitat Jaume I, Spain. This is a randomly chosen subset of data for 19,937 locations published by [Torres-Sospedra et al. \[2014\]](#). The response is bivariate, consisting of the longitude and latitude of the location where each recording was made. The aim is to predict location on the basis of Wi-Fi signal strength. [Torres-Sospedra et al. \[2014\]](#) applied k -nearest neighbour regression to the same problem, but using the full 520 dimensions of the predictors, rather than first reducing dimension with PCA.
- *CT scans.* Predictors are $p = 384$ bone structure and air inclusion features derived for each of $n = 5000$ Computed Tomography (CT) scan images of human bodies, a randomly chosen subset of 53,500 CT images from [Graf et al. \[2011\]](#). The response is the univariate location at which the scan was taken along the axis of the patient’s body, ranging from the top of the head to the end of the coccyx. [Graf et al. \[2011\]](#) applied PCA to first reduce dimension from $p = 384$ to $r = 50$, then predicted the same response with a two-level nearest neighbour technique. [Cheng and Wu \[2013\]](#) considered the same data set and regression problem, but used a more sophisticated manifold adaptive local linear estimator.

Classification examples

- *Mixture model.* \mathbf{Y} simulated from the model in section 5.2 with $n = 250$, $p = 10^4$, $\sigma = 10$ and 10 mixture components. The class labels were taken to be the Z_i , i.e. the indices of the mixture components which the data points are associated with in the simulation.
- *MNIST handwritten digits.* This is a well-known data set of grayscale images of written characters 0-9, each image consisting of $p = 784 = 28 \times 28$ pixels. The full MNIST database [[LeCun et al., 1998](#)] consists of over 60,000 images, we took a random subset of size $n = 2000$. This is a 10-class problem, the class labels being the numbers 0-9 which the images depict.
- *DOROTHEA drug discovery.* Predictors are $p = 10000$ structural features for $n = 800$ chemical compounds. This is a binary classification problem, in which compounds are labelled as thrombin-binding active, or inactive [[Guyon et al., 2004](#)].

Figures 11 and 12 show performance of k -nearest neighbour regression/classification in the six scenarios, across 200 randomly chosen 70/30 train/test splits of each data set. Noting that the coefficient of determination “ R^2 ” shown in 11 is associated with the test data, hence it is not impossible that $1 - R^2 \geq 1$, it is visible that using a low r , in the range roughly 1 – 10, has a deleterious effect on predicting elevation for the torus example. This can be connected with figure 8: there it is apparent that the first six dimensions of the PC scores seem to more strongly convey azimuth than elevation. In all three regression scenarios, it is noticeable that as r increases, $1 - R^2$ first decreases, then increases and plateaus. For the classification examples in figure 12 we find similar results. For the mixture model example, it is unsurprising that there is a minimum in the misclassification rate when r is around the true number of mixture components, 10.

Overall these numerical results illustrate that using PCA as a pre-processing step, to reduce predictors from high to moderate dimension, can have a substantial benefit in terms of statistical performance of nearest neighbour regression/classification. We also see that further reducing dimension can have a strong deleterious effect. These observations are consistent with existing theory showing that nearest neighbour regression/classification is automatically well adapted to situations where the predictors lie on a manifold of low dimension [[Kpotufe, 2011](#)]. Formally establishing whether the same adaptivity occurs here, in particular, under vanishing rather than zero off-manifold error, is left as an open question.

It’s important to note that for a long time statisticians have warned against the use of PCA to reduce the dimension of predictors in regression. [Jolliffe \[1982\]](#) discussed examples of regression problems in which the particular linear combinations of predictors which are most useful for prediction are also those which happen to have the lowest variance, hence discarding them by performing dimensionality reduction using PCA cripples predictive performance. Further investigation may reveal more detailed, problem-specific and potentially data-driven guidance, to help practitioners decide what approach to take given a specific problem at hand.

5.8 Choosing the PCA dimension

Having covered a diversity of downstream applications of PCA — e.g. clustering, TDA, dimension reduction, manifold estimation, regression — we now approach the question of choosing the dimension

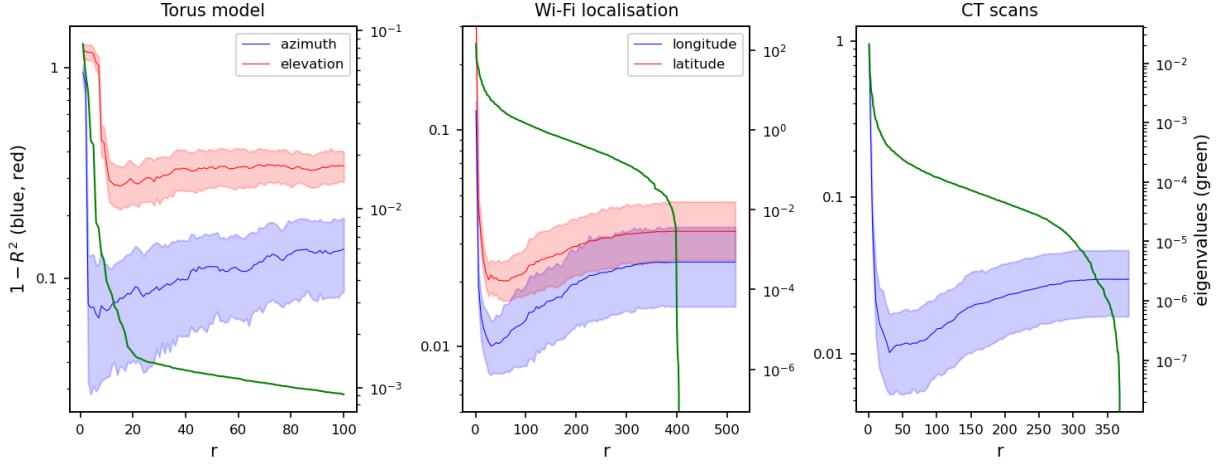


Figure 11: Regression examples. For 200 random 70/30 splits of the data into training/test, the blue and red lines show mean $1 - R^2$ for the test data, with shaded bands indicating 5% – 95% percentiles. The green lines show the ordered eigenvalues of $(np)^{-1}\mathbf{Y}^\top \mathbf{Y}$.

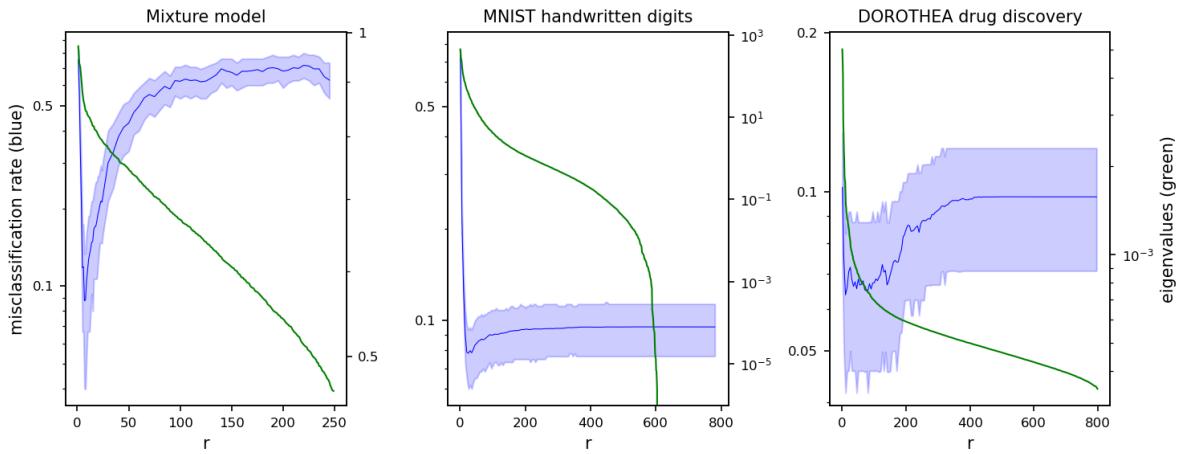


Figure 12: Classification examples. For 200 random 70/30 splits of the data into training/test, the blue lines show mean misclassification rate for the test data, with shaded bands indicating 5% – 95% percentiles. The green lines show the ordered eigenvalues of $(np)^{-1}\mathbf{Y}^\top \mathbf{Y}$.

r in practice.

First, a caution against any claim of theoretical optimality: Under a latent metric space model with known and finite rank r , despite our theory, we should not universally expect that applying PCA into r dimensions should improve downstream results, even asymptotically. For example, in the context of clustering under the isotropic Gaussian mixture model, we would always prefer the theoretical solution to the p -dimensional k -means objective to the solution obtained by PCA into r dimensions followed by k -means. The advantage of the latter is that it is computationally attainable, and has *almost* the same accuracy [Löffler et al., 2021]. The statistical advantages of applying PCA become much more substantial when we compare the performance of *actual* algorithms, rather than their theoretical objectives.

In this paper, the principal benefit of applying PCA is to remove noise, and this must usually come at the cost of losing at least some of the signal. In choosing a dimension one should be mindful of this bias/variance tradeoff, and how it will affect downstream inference; ideally, with a specific application and algorithm in view. Even if the rank is known to be r , the bias/variance tradeoff may favour a different choice — typically, a lower one if the variance is high. By extension, even if the rank is infinite, the bias/variance tradeoff might nonetheless suggest applying PCA into a moderately low dimension. If the problem allows, such as in regression and classification, we would recommend picking the dimension by cross-validation, as is common practice. However, sometimes one must pick a dimension, simply based on the matrix \mathbf{Y} , for an open-ended goal, which is the situation we now confront.

Dimension selection for PCA is an old and well-studied problem, reviewed for example in Zhu and Ghodsi [2006] and Jolliffe [2002]. Despite this, we have found existing methods unsatisfactory for several reasons, and prefer to suggest a new technique. We do not claim objective superiority — just that our method is automatic, moderately scalable, sound in practice, and provides a useful visualisation of the bias/variance trade-off at play in the given data. In particular, we have found the ladle method by [Luo and Li, 2016] to be comparable in all metrics, except computational scalability (as implemented), which has made analysing some of the larger datasets in this paper by this method impossible (e.g. $\min(n, p)$ much larger than 1000).

Our proposed method is described precisely in Algorithm 3. In words, we split the data into two, and for each r we project the first half onto the r principal eigenvectors — the points remain p -dimensional, just constrained to an r -dimensional subspace. Next, we measure how much this projection step has brought the first half closer to the second, using the first Wasserstein distance, as implemented in the R package ‘transport’. The r achieving the lowest distance is selected.

We now explore the performance of Algorithm 3 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 13a)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, $f(x, y) = \exp(-\|x - y\|_{\mathbb{R}^2}^2/2)$, which has infinite rank.

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

By eye, it is impossible to distinguish any of the structure of \mathcal{Z} by visualising selected pairs or triples of data coordinates from \mathbf{Y} (figure 14, appendix), but the first two principal components of the data, shown in figures 13b)1-4, draw out some structure. For example, in figure 13b)1, we can distinguish 6 clusters. In the remaining configurations, this two-dimensional view is insufficient and obscures obvious topological and geometric features of \mathcal{Z} , for example, the hole in configuration 4, or the ‘Z’ shape in configuration 3. A plot of the second and third principal components (figure 15, appendix) gives a better view of \mathcal{Z} in configurations 2-4 but in isolation could tell the wrong story in configuration 1, suggesting 5 rather than 6 clusters.

In figures 13c)1-4 we show the Wasserstein error (log-scale), i.e., the distance computed in Algorithm 3, for different choices of dimension. 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 16, 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>:

Algorithm 3 PCA dimension selection

Input: data matrix $\mathbf{Y} \in \mathbb{R}^{n \times p}$.

- 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** $r \in \{1, \dots, \min(n, p)\}$ **do**
- 3: Let $\mathbf{V}^{(1)} \in \mathbb{R}^{p \times r}$ denote the matrix of orthogonal eigenvectors associated with the r largest eigenvalues of $\mathbf{Y}^{(1)^\top} \mathbf{Y}^{(1)}$
- 4: Project $\mathbf{Y}^{(1)}$ onto the linear span of $\mathbf{V}^{(1)}$, $\hat{\mathbf{X}}^{(1)} := \mathbf{Y}^{(1)} \mathbf{V}^{(1)} \mathbf{V}^{(1)^\top}$
- 5: Compute Wasserstein distance d_r between $\hat{\mathbf{X}}^{(1)}$ and $\mathbf{Y}^{(2)}$ (as point sets in \mathbb{R}^p)
- 6: **end for**

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

//github.com/WenlanZ) 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 rank selection under the Latent Metric Space 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 geodesic distances between Z_1, \dots, Z_n , along \mathcal{Z} , as estimated geodesic distances between $p^{-1/2}\zeta_1, \dots, p^{-1/2}\zeta_n$, along \mathcal{M} . The method in Section 5.5 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 13d)3-4 shows the entrywise mean square error between the estimated geodesic distance matrices of Z_1, \dots, Z_n and $p^{-1/2}\zeta_1, \dots, p^{-1/2}\zeta_n$, 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, although even asymptotically we should not expect an exact match. The red line in figures 13d)3-4 shows the bottleneck distance between the persistence diagrams of the Rips filtrations of Z_1, \dots, Z_n and $p^{-1/2}\zeta_1, \dots, p^{-1/2}\zeta_n$, 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 13d) we have recentered and rescaled the curves to make their maxima and minima agree.

In figures 13e)1-4 we show the persistence diagrams of the Rips filtrations of $p^{-1/2}\zeta_1, \dots, p^{-1/2}\zeta_n$ computed on the basis of the rank selected by the Wasserstein method (Algorithm 3), 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).

5.9 Sample-centered data and a connection to Kernel PCA

We now examine the impact of sample-centering data. Let $\bar{\mathbf{Y}} \in \mathbb{R}^{n \times p}$ be the matrix whose rows are all equal to the sample average $\frac{1}{n}(\mathbf{1}_n^\top \mathbf{Y})$. When calculating PC scores using the sample-centered data $\mathbf{Y} - \bar{\mathbf{Y}}$, one projects the data onto eigenvectors of $(\mathbf{Y} - \bar{\mathbf{Y}})^\top (\mathbf{Y} - \bar{\mathbf{Y}})$, which up to scaling is a sample covariance matrix. Theorem 1 does not immediately apply to this scenario because the dependence structure of $\mathbf{Y} - \bar{\mathbf{Y}}$ is different to that of \mathbf{Y} . However, if one were to pursue a concentration result analogous to theorem 1 for the PC scores associated with $\mathbf{Y} - \bar{\mathbf{Y}}$, a natural first step analogous to lemma 5 would be to compute the conditional expectation of $(\mathbf{Y} - \bar{\mathbf{Y}})(\mathbf{Y} - \bar{\mathbf{Y}})^\top$ given Z_1, \dots, Z_n .

Lemma 6. *Assume A1 and A3. Then*

$$p^{-1} \mathbb{E}[(\mathbf{Y} - \bar{\mathbf{Y}})(\mathbf{Y} - \bar{\mathbf{Y}})^\top | Z_1, \dots, Z_n] = \mathbf{C}_n (\Phi \Phi^\top + \sigma^2 \mathbf{I}_n) \mathbf{C}_n, \quad (8)$$

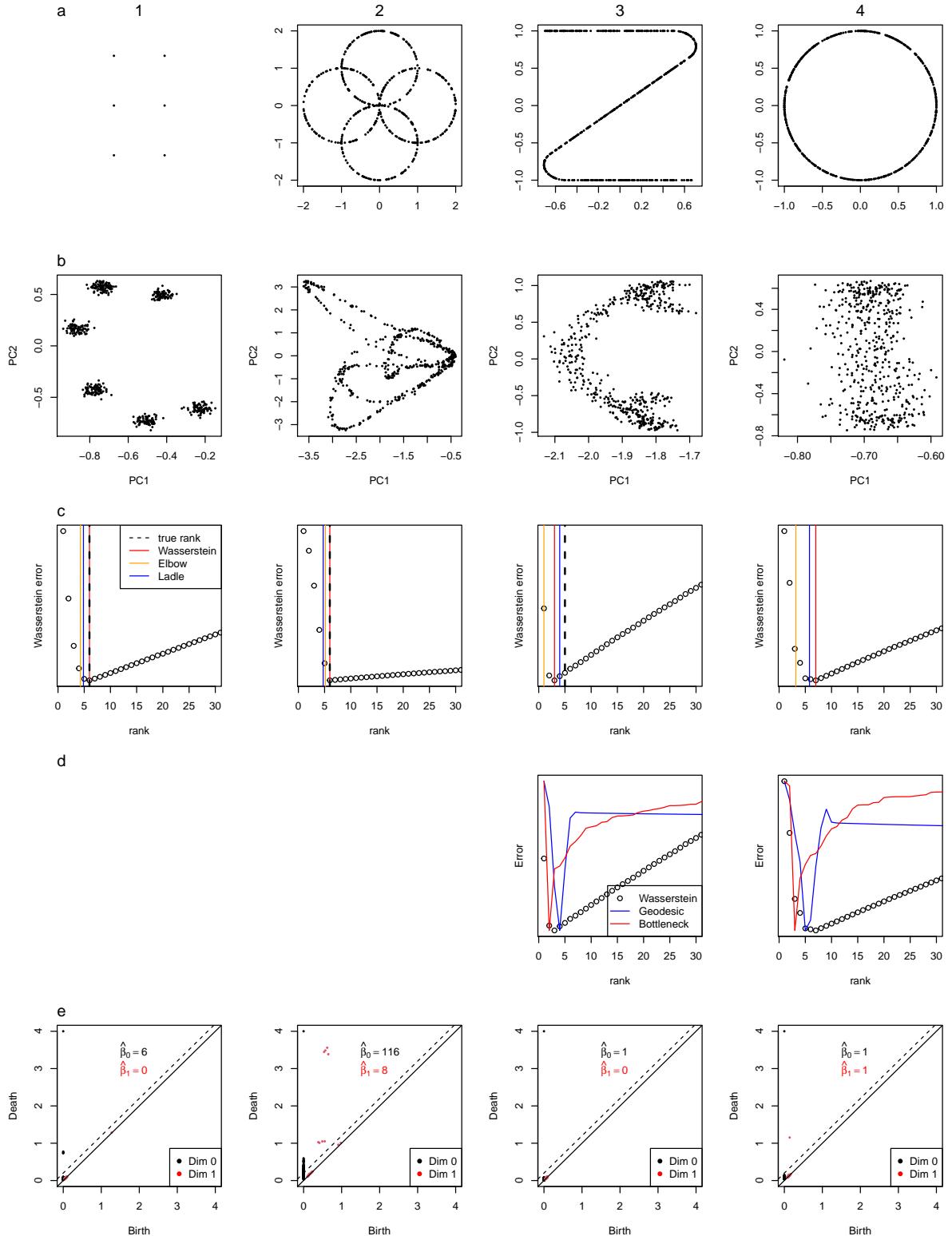


Figure 13: PCA dimension selection. Columns 1-4: different latent space/kernel configurations. 1-3 are finite rank, 4 infinite rank; configurations 3 and 4 are isometric. Row a: sampled positions ($n = 500$); b: first two principal components ($p = 1000$); c: the dimension selected by different methods, 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.

where $\mathbf{C}_n := \mathbf{I}_n - n^{-1}\mathbf{J}_n$ and \mathbf{J}_n is the $n \times n$ matrix of 1's. Moreover

$$(\mathbf{C}_n \Phi \Phi^\top \mathbf{C}_n)_{ij} = \tilde{f}(Z_i, Z_j; Z_1, \dots, Z_n) \quad (9)$$

where for $z, z' \in \mathcal{Z}$,

$$\tilde{f}(z, z'; Z_1, \dots, Z_n) := \mathbb{E} \left[\frac{1}{p} \sum_{j=1}^p \left(X_j(z) - \frac{1}{n} \sum_{i=1}^n X_j(Z_i) \right) \left(X_j(z') - \frac{1}{n} \sum_{i=1}^n X_j(Z_i) \right) \middle| Z_1, \dots, Z_n \right] \quad (10)$$

$$= f(z, z') - \frac{1}{n} \sum_{i=1}^n f(Z_i, z) - \frac{1}{n} \sum_{i=1}^n f(Z_i, z') + \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n f(Z_i, Z_j). \quad (11)$$

Proof. The first equality follows from the identity $(\mathbf{Y} - \bar{\mathbf{Y}})(\mathbf{Y} - \bar{\mathbf{Y}})^\top = \mathbf{C}_n \mathbf{Y} \mathbf{Y}^\top \mathbf{C}_n$ combined with lemma 5. The other equalities are derived from the following two identities: $(\Phi \Phi^\top)_{ij} = f(Z_i, Z_j) = p^{-1} \sum_{j=1}^p \mathbb{E}[X_j(Z_j) X_j(Z_i) | Z_1, \dots, Z_n]$ and the linearity of conditional expectation. \square

If we define

$$\tilde{\phi}(z; Z_1, \dots, Z_n) := \phi(z) - \frac{1}{n} \sum_{i=1}^n \phi(Z_i)$$

then using (11) and $f(z, z') = \langle \phi(z), \phi(z') \rangle_2$ we find

$$\tilde{f}(z, z'; Z_1, \dots, Z_n) = \left\langle \tilde{\phi}(z; Z_1, \dots, Z_n), \tilde{\phi}(z'; Z_1, \dots, Z_n) \right\rangle_2,$$

i.e., $\tilde{\phi}(\cdot; Z_1, \dots, Z_n)$ is a feature map for $\tilde{f}(\cdot, \cdot; Z_1, \dots, Z_n)$. We thus see that a consequence of working with the sample-centered data $\mathbf{Y} - \bar{\mathbf{Y}}$ rather than \mathbf{Y} is that it leads us to the centered feature map $\tilde{\phi}(\cdot; Z_1, \dots, Z_n)$. It is straightforward to check that injectivity and continuity of ϕ is equivalent to that of $\tilde{\phi}(\cdot; Z_1, \dots, Z_n)$. Recalling the proof of lemma 2, this tells us that the set $\tilde{\mathcal{M}}(Z_1, \dots, Z_n) := \tilde{\phi}(\mathcal{Z}; Z_1, \dots, Z_n)$ is topologically equivalent to \mathcal{Z} whenever \mathcal{M} is. Moreover, it is clear from (11) that the second-order mixed derivatives of f are equal to those of $\tilde{f}(\cdot, \cdot; Z_1, \dots, Z_n)$, so the matrix \mathbf{H}_ξ which defines the Riemannian metric in proposition 3 would be unchanged if f was replaced by $\tilde{f}(\cdot, \cdot; Z_1, \dots, Z_n)$. In these senses we can say that the centered feature map $\tilde{\phi}(\cdot; Z_1, \dots, Z_n)$ conveys the topology and geometry of \mathcal{Z} in the same way that ϕ does.

Consideration of \tilde{f} and $\tilde{\phi}$ also highlight a connection to kernel PCA [Schölkopf et al., 1998]: the double-centered Gram matrix $\mathbf{C}_n \Phi \Phi^\top \mathbf{C}_n$, or equivalently (11) evaluated over $(z, z') \in \{Z_1, \dots, Z_n\}^2$, is exactly the matrix which one would evaluate if applying kernel PCA with kernel f to Z_1, \dots, Z_n as input data. We thus see that computing the PC scores associated with $\mathbf{Y} - \bar{\mathbf{Y}}$ can be viewed as a “noisy” (owing to the randomness in the errors $p^{-1/2} \mathbf{Q} \zeta_i - \phi(Z_i)$) version of kernel PCA with kernel f applied to Z_1, \dots, Z_n .

Taking this thought further, we note that Kernel PCA and kernel methods more generally are often motivated by the idea that clusters in data which are not linearly separable in their low dimensional native space may become linearly separable when mapped into a high-dimensional feature space. It is tantalising to observe that in terms of such linear separability, computing the PC scores $p^{-1/2} \zeta_1, \dots, p^{-1/2} \zeta_n$ with n and p large enough to achieve suitable concentration around $\phi(Z_1), \dots, \phi(Z_n)$, potentially could be preferable to having direct access to the unobserved variables Z_1, \dots, Z_n . Of course in the practice of kernel PCA the user chooses a kernel, whereas in our setting of the Latent Metric Space model the kernel f arises implicitly once the fields X_1, \dots, X_p are defined.

Concerning kernel methods more generally, we note the results of section 4 apply to any kernel defined on a compact metric space, not just the mean correlation kernel f . It may be of interest to investigate whether the results of section 4 shed light on the theoretical performance of kernel PCA or other kernel methods.

References

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

- 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.
- Christophe Andrieu and Gareth O Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.
- Sanjeev Arora, Wei Hu, and Pravesh K Kothari. An analysis of the t-SNE algorithm for data visualization. In *Conference On Learning Theory*, pages 1455–1462. PMLR, 2018.
- Sivaraman Balakrishnan, Alessandro Rinaldo, Don Sheehy, Aarti Singh, and Larry Wasserman. Minimax rates for homology inference. In *Artificial Intelligence and Statistics*, pages 64–72. PMLR, 2012.
- Mark A Beaumont. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160, 2003.
- Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.
- Mikhail Belkin and Partha Niyogi. Convergence of laplacian eigenmaps. *Advances in neural information processing systems*, 19, 2006.
- Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.
- 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.
- Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is “nearest neighbor” meaningful? In *International conference on database theory*, pages 217–235. Springer, 1999.
- 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>.
- 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.
- T Tony Cai and Rong Ma. Theoretical foundations of t-SNE for visualizing high-dimensional clustered data. *arXiv preprint arXiv:2105.07536*, 2021.
- 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.
- 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.
- 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.

- Jose A Costa and Alfred O Hero. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Transactions on Signal Processing*, 52(8):2210–2221, 2004.
- 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.
- Christopher C Drovandi, Matthew T Moores, and Richard J Boys. Accelerating pseudo-marginal MCMC using Gaussian processes. *Computational Statistics & Data Analysis*, 118:1–17, 2018.
- Herbert Edelsbrunner, John Harer, et al. Persistent homology-a survey. *Contemporary mathematics*, 453:257–282, 2008.
- 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.
- 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*, pages 1–6, 2022.
- 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.
- Franz Graf, Hans-Peter Kriegel, Matthias Schubert, Sebastian Pölsterl, and Alexander Cavallaro. 2D image registration in CT images using radial image descriptors. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 607–614. Springer, 2011.
- Victor Guillemin and Alan Pollack. *Differential topology*. Prentice-Hall, 1974.
- Isabelle Guyon, Steve Gunn, Asa Ben-Hur, and Gideon Dror. Result analysis of the Nips 2003 feature selection challenge. *Advances in neural information processing systems*, 17, 2004.
- Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.
- Matthias Hein and Jean-Yves Audibert. Intrinsic dimensionality estimation of submanifolds in rd. In *Proceedings of the 22nd international conference on Machine learning*, 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 *NIPS*, volume 15, pages 833–840. Citeseer, 2002.
- 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.

- Ian T Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(3):300–303, 1982.
- Ian T Jolliffe. *Principal component analysis*. Springer, 2002.
- Sungkyu Jung and J Stephen Marron. PCA consistency in high dimension, low sample size context. *The Annals of Statistics*, 37(6B):4104–4130, 2009.
- Purushottam Kar and Harish Karnick. Random feature maps for dot product kernels. In *Artificial intelligence and statistics*, pages 583–591. PMLR, 2012.
- Balázs Kégl. Intrinsic dimension estimation using packing numbers. *Advances in neural information processing systems*, 15, 2002.
- Samory Kpotufe. k-NN regression adapts to local intrinsic dimension. *Advances in neural information processing systems*, 24, 2011.
- John Lafferty and Larry Wasserman. Statistical analysis of semi-supervised regression. *Advances in Neural Information Processing Systems*, 20, 2007.
- Ken Lang. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, pages 331–339. Elsevier, 1995.
- Oscar Lao, Timothy T Lu, Michael Nothnagel, Olaf Junge, Sandra Freitag-Wolf, Amke Caliebe, Miroslava Balascakova, Jaume Bertranpetti, Laurence A Bindoff, David Comas, et al. Correlation between genetic and geographic structure in europe. *Current Biology*, 18(16):1241–1248, 2008.
- Neil Lawrence and Aapo Hyvärinen. 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 *Nips*, volume 2, page 5. Citeseer, 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.
- Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- 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.
- John Lee. *Introduction to topological manifolds*, volume 202. Springer Science & Business Media, 2010.
- 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*, 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.
- George C Linderman and Stefan Steinerberger. Clustering with t-SNE, provably. *SIAM Journal on Mathematics of Data Science*, 1(2):313–332, 2019.
- Anna V Little. *Estimating the intrinsic dimension of high-dimensional data sets: a multiscale, geometric approach*. PhD thesis, Duke University, 2011.
- 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.
- Gabriele Lombardi, Alessandro Rozza, Claudio Ceruti, Elena Casiraghi, and Paola Campadelli. Minimum neighbor distance estimators of intrinsic dimension. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 374–389. Springer, 2011.

- 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.
- 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.
- 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.
- Amit Moscovich, Ariel Jaffe, and Nadler Boaz. Minimax-optimal semi-supervised regression on unknown manifolds. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 933–942. PMLR, 20–22 Apr 2017. URL <https://proceedings.mlr.press/v54/moscovich17a.html>.
- 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.
- 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.
- Peter Petersen. *Riemannian geometry*, volume 171 of *Graduate Texts in Mathematics*. Springer, 2006.
- Robert Pless and Richard Souvenir. A survey of manifold learning for images. *IPSJ Transactions on Computer Vision and Applications*, 1:83–94, 2009.
- 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*, 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.
- Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

- 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.
- 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.
- 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.
- Joaquín Torres-Sospedra, Raúl Montoliu, Adolfo Martínez-Usó, Joan P Avariento, Tomás J Arnau, Mauri Benedito-Bordonau, and Joaquín Huerta. Ujiindoorloc: A new multi-building and multi-floor database for WLAN fingerprint-based indoor localization problems. In *2014 international conference on indoor positioning and indoor navigation (IPIN)*, pages 261–270. IEEE, 2014.
- 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.
- Laurens Van Der Maaten. Accelerating t-SNE using tree-based algorithms. *The 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.
- 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*, 34, 2021.
- 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 7 (Mercer's theorem). *Let \mathcal{Z} be a compact metric space and let $f : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$, 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 $(\lambda_k^f)_{k \geq 1}$, $\lambda_1^f \geq \lambda_2^f \geq \dots$ 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 Proof and supporting results for theorem 1

B.1 Definitions and preliminaries

Throughout section B the probability measure μ is considered fixed and $(\lambda_k^f, u_k^f)_{k \geq 1}$ are as in section 2.1.

B.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 $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \dots$. 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 $\|\cdot\|_F$.

B.1.2 Some matrices of interest

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

$$\Phi := [\phi(Z_1) | \dots | \phi(Z_n)]^\top,$$

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

B.1.3 Some events of interest

With U_j denoting the j th column of \mathbf{U}_Φ , define:

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

B.2 Proof of the concentration theorem

Proof of theorem 1. Let $\mathbf{F}_1 \Sigma \mathbf{F}_2^\top$ be the full singular value decomposition of $\mathbf{U}_\Phi^\top \mathbf{U}_Y$ and define the random orthogonal matrix $\mathbf{F}_* := \mathbf{F}_1 \mathbf{F}_2^\top$. On the event A_{rank} we have $\mathbf{U}_\Phi \Lambda_\Phi \mathbf{U}_\Phi^\top = \Phi \Phi^\top$, and applying lemma 11 we find there exists a random orthogonal matrix $\widehat{\mathbf{Q}}$ such that $\mathbf{U}_\Phi \Lambda_\Phi^{1/2} = \Phi \widehat{\mathbf{Q}}$, hence $[\mathbf{U}_\Phi \Lambda_\Phi^{1/2} \mathbf{F}_*]_i = \phi(Z_i)^\top \mathbf{Q}$ for all $i = 1, \dots, n$, where $\mathbf{Q} := \widehat{\mathbf{Q}} \mathbf{F}_*$ is orthogonal and $[\cdot]_i$ denotes the i th row of a matrix. Lemma 4 shows that $[\mathbf{U}_Y \Lambda_Y^{1/2}]_i = p^{-1/2} \zeta_i$. Combining these observations we have shown that on the event A_{rank} ,

$$\|p^{-1/2} \mathbf{Q} \zeta_i - \phi(Z_i)\|_2 = \|[\mathbf{U}_Y \Lambda_Y^{1/2} - \mathbf{U}_\Phi \Lambda_\Phi^{1/2} \mathbf{F}_*]_i\|_2, \quad i = 1, \dots, n. \quad (12)$$

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 1 we shall establish various identities and inequalities involving random variables, random matrices, etc; all such identities 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 “ $X \leq Y$ ” as shorthand for “ $X(\omega) \leq Y(\omega)$ 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 $\Lambda_Y^{-1/2}$ and $\Lambda_\Phi^{-1/2}$ are well-defined, let us introduce:

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

We now claim that:

$$\mathbf{U}_Y \Lambda_Y^{1/2} - \mathbf{U}_\Phi \Lambda_\Phi^{1/2} \mathbf{F}_* = (p^{-1} \mathbf{Y} \mathbf{Y}^\top - \Phi \Phi^\top) \mathbf{U}_\Phi \Lambda_\Phi^{-1/2} \mathbf{F}_* + \sum_{i=1}^5 \mathbf{D}_i, \quad (13)$$

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 (13), observe:

$$\begin{aligned}
\mathbf{U}_Y \Lambda_Y^{1/2} - \mathbf{U}_\Phi \Lambda_\Phi^{1/2} \mathbf{F}_* &= \mathbf{U}_Y \Lambda_Y^{1/2} - \mathbf{U}_\Phi \mathbf{F}_* \Lambda_Y^{1/2} \\
&\quad + \mathbf{U}_\Phi \mathbf{C}_1 \\
&= (\mathbf{I}_n - \mathbf{U}_\Phi \mathbf{U}_\Phi^\top) \mathbf{U}_Y \Lambda_Y^{1/2} \\
&\quad + \mathbf{U}_\Phi \mathbf{C}_2 \\
&\quad + \mathbf{U}_\Phi \mathbf{C}_1 \\
&= (\mathbf{I}_n - \mathbf{U}_\Phi \mathbf{U}_\Phi^\top) (p^{-1} \mathbf{Y} \mathbf{Y}^\top - \Phi \Phi^\top) \mathbf{U}_Y \Lambda_Y^{-1/2} \\
&\quad + \mathbf{U}_\Phi \mathbf{C}_2 \\
&\quad + \mathbf{U}_\Phi \mathbf{C}_1 \\
&= (p^{-1} \mathbf{Y} \mathbf{Y}^\top - \Phi^\top \Phi) \mathbf{U}_\Phi \mathbf{F}_* \Lambda_Y^{-1/2} \\
&\quad - \mathbf{U}_\Phi \mathbf{U}_\Phi^\top (p^{-1} \mathbf{Y} \mathbf{Y}^\top - \Phi \Phi^\top) \mathbf{U}_\Phi \mathbf{F}_* \Lambda_Y^{-1/2} \\
&\quad + (\mathbf{I} - \mathbf{U}_\Phi \mathbf{U}_\Phi^\top) (p^{-1} \mathbf{Y} \mathbf{Y}^\top - \Phi \Phi^\top) \mathbf{C}_3 \Lambda_Y^{-1/2} \\
&\quad + \mathbf{U}_\Phi \mathbf{C}_2 \\
&\quad + \mathbf{U}_\Phi \mathbf{C}_1 \\
&= (p^{-1} \mathbf{Y} \mathbf{Y}^\top - \Phi \Phi^\top) \mathbf{U}_\Phi \Lambda_\Phi^{-1/2} \mathbf{F}_* \\
&\quad + (p^{-1} \mathbf{Y} \mathbf{Y}^\top - \Phi \Phi^\top) \mathbf{U}_\Phi (\mathbf{F}_* \Lambda_Y^{-1/2} - \Lambda_\Phi^{-1/2} \mathbf{F}_*) \\
&\quad - \mathbf{U}_\Phi \mathbf{U}_\Phi^\top (p^{-1} \mathbf{Y} \mathbf{Y}^\top - \Phi \Phi^\top) \mathbf{U}_\Phi \mathbf{F}_* \Lambda_Y^{-1/2} \\
&\quad + (\mathbf{I} - \mathbf{U}_\Phi \mathbf{U}_\Phi^\top) (p^{-1} \mathbf{Y} \mathbf{Y}^\top - \Phi \Phi^\top) \mathbf{C}_3 \Lambda_Y^{-1/2} \\
&\quad + \mathbf{U}_\Phi \mathbf{C}_2 \\
&\quad + \mathbf{U}_\Phi \mathbf{C}_1 \\
&= (p^{-1} \mathbf{Y} \mathbf{Y}^\top - \Phi \Phi^\top) \mathbf{U}_\Phi \Lambda_\Phi^{-1/2} \mathbf{F}_* + \mathbf{D}_5 + \mathbf{D}_4 + \mathbf{D}_3 + \mathbf{D}_2 + \mathbf{D}_1 \tag{14}
\end{aligned}$$

where (14) holds because $\mathbf{U}_Y \Lambda_Y^{1/2} = p^{-1} \mathbf{Y}^\top \mathbf{Y} \mathbf{U}_Y \Lambda_Y^{-1/2}$ and $\mathbf{U}_\Phi \mathbf{U}_\Phi^\top \Phi \Phi^\top = \Phi \Phi^\top$.

The proof proceeds by bounding the Frobenius norm of each matrix \mathbf{D}_i , $i = 1, \dots, 5$. Using lemma 9,

$$\begin{aligned}
\|\mathbf{D}_1\|_F &= \|\mathbf{C}_1\|_F \\
&\leq \frac{r^{1/2}}{2n^{1/2}(1-\epsilon_2)^{1/2}(\lambda_r^f)^{1/2}} \left[n \frac{(\epsilon_1 + n^{-1}\sigma^2)^2}{\lambda_r^f(1-2\epsilon_2)} \left(1 + 2 \frac{\lambda_1^f}{\lambda_r^f} \left(\frac{1+\epsilon_2}{1-2\epsilon_2} \right) \right) + n\epsilon_1 + \sigma^2 \right] \\
&= \frac{r^{1/2}n^{1/2}(\epsilon_1 + n^{-1}\sigma^2)}{2(1-\epsilon_2)^{1/2}(\lambda_r^f)^{1/2}} \left[\frac{(\epsilon_1 + n^{-1}\sigma^2)}{\lambda_r^f(1-2\epsilon_2)} \left(1 + 2 \frac{\lambda_1^f}{\lambda_r^f} \left(\frac{1+\epsilon_2}{1-2\epsilon_2} \right) \right) + 1 \right]. \tag{16}
\end{aligned}$$

Using lemma 8,

$$\begin{aligned}
\|\mathbf{D}_2\|_F &\leq r^{1/2} \|\mathbf{C}_2\|_2 \\
&= r^{1/2} n^{1/2} [\lambda_1^f(1+\epsilon_2)]^{1/2} \left[\frac{\epsilon_1 + n^{-1}\sigma^2}{\lambda_r^f(1-2\epsilon_2)} \right]^2. \tag{17}
\end{aligned}$$

Again using lemma 8 and the fact that $\mathbf{U}_Y - \mathbf{U}_\Phi \mathbf{U}_\Phi^\top \mathbf{U}_Y = (\mathbf{U}_Y \mathbf{U}_Y^\top - \mathbf{U}_\Phi \mathbf{U}_\Phi^\top) \mathbf{U}_Y$,

$$\begin{aligned}
\|\mathbf{D}_3\|_F &\leq 2r^{1/2} \|p^{-1} \mathbf{Y} \mathbf{Y}^\top - \Phi \Phi^\top\|_2 \|\mathbf{C}_3\|_2 \|\Lambda_Y^{-1/2}\|_2 \\
&\leq 2r^{1/2} \frac{(\epsilon_1 n + \sigma^2)}{n^{1/2} [\lambda_r^f(1-\epsilon_2)]^{1/2}} (\|\mathbf{U}_Y \mathbf{U}_Y^\top - \mathbf{U}_\Phi \mathbf{U}_\Phi^\top\|_2 + \|\mathbf{U}_\Phi^\top \mathbf{U}_Y - \mathbf{F}_*\|_2) \\
&\leq 2r^{1/2} n^{1/2} \frac{(\epsilon_1 + n^{-1}\sigma^2)^2}{[\lambda_r^f(1-\epsilon_2)]^{3/2}} \left(1 + \frac{\epsilon_1 + n^{-1}\sigma^2}{\lambda_r^f(1-\epsilon_2)} \right) \tag{18}
\end{aligned}$$

Directly:

$$\begin{aligned}
\|\mathbf{D}_4\|_F &\leq r^{1/2} \|\mathbf{D}_4\|_2 \\
&\leq r^{1/2} \|p^{-1} \mathbf{Y} \mathbf{Y}^\top - \Phi \Phi^\top\|_2 \|\Lambda_{\mathbf{Y}}^{-1/2}\|_2 \\
&\leq r^{1/2} \frac{(\epsilon_1 n + \sigma^2)}{n^{1/2} [\lambda_r^f (1 - \epsilon_2)]^{1/2}} \\
&= r^{1/2} n^{1/2} \frac{(\epsilon_1 + n^{-1} \sigma^2)}{[\lambda_r^f (1 - \epsilon_2)]^{1/2}}
\end{aligned} \tag{19}$$

Using lemma 9,

$$\begin{aligned}
\|\mathbf{D}_5\|_F &= \|(p^{-1} \mathbf{Y} \mathbf{Y}^\top - \Phi \Phi^\top) \mathbf{U}_\Phi (\mathbf{F}_* \Lambda_{\mathbf{Y}}^{-1/2} - \Lambda_\Phi^{-1/2} \mathbf{F}_*)\|_F \\
&\leq r^{1/2} \|p^{-1} \mathbf{Y} \mathbf{Y}^\top - \Phi \Phi^\top\|_2 \|\mathbf{F}_* \Lambda_{\mathbf{Y}}^{-1/2} - \Lambda_\Phi^{-1/2} \mathbf{F}_*\|_F \\
&\leq r^{1/2} (\epsilon_1 n + \sigma^2) \frac{\|\mathbf{F}_* \Lambda_{\mathbf{Y}} - \Lambda_\Phi \mathbf{F}_*\|_F}{2n^{3/2} (\lambda_r^f)^{3/2} (1 - \epsilon_2)^{3/2}} \\
&\leq \frac{rn^2 (\epsilon_1 + n^{-1} \sigma^2)}{2n^{3/2} (\lambda_r^f)^{3/2} (1 - \epsilon_2)^{3/2}} \left[\frac{(\epsilon_1 + n^{-1} \sigma^2)^2}{\lambda_r^f (1 - 2\epsilon_2)} \left(1 + 2 \frac{\lambda_1^f}{\lambda_r^f} \left(\frac{1 + \epsilon_2}{1 - 2\epsilon_2} \right) \right) + \epsilon_1 + \frac{\sigma^2}{n} \right] \\
&= \frac{rn^{1/2} (\epsilon_1 + n^{-1} \sigma^2)^2}{2(\lambda_r^f)^{3/2} (1 - \epsilon_2)^{3/2}} \left[\frac{(\epsilon_1 + n^{-1} \sigma^2)}{\lambda_r^f (1 - 2\epsilon_2)} \left(1 + 2 \frac{\lambda_1^f}{\lambda_r^f} \left(\frac{1 + \epsilon_2}{1 - 2\epsilon_2} \right) \right) + 1 \right]
\end{aligned} \tag{20}$$

Having obtained the above bounds on $\|\mathbf{D}_i\|_F$, for $i = 1, \dots, 5$, we turn to the first term on the r.h.s. of (13). Writing $[\cdot]_i$ to indicate the i th row of a matrix,

$$\begin{aligned}
\max_{i=1, \dots, n} \|[(p^{-1} \mathbf{Y} \mathbf{Y}^\top - \Phi \Phi^\top) \mathbf{U}_\Phi \Lambda_\Phi^{-1/2} \mathbf{F}_*]_i\|_2 &= \max_i \|[(p^{-1} \mathbf{Y} \mathbf{Y}^\top - \Phi \Phi^\top) \mathbf{U}_\Phi \Lambda_\Phi^{-1/2}]_i\|_2 \\
&\leq \frac{1}{n^{1/2} (\lambda_r^f)^{1/2} (1 - \epsilon_2)^{1/2}} \max_{i=1, \dots, n} \|[(p^{-1} \mathbf{Y} \mathbf{Y}^\top - \Phi \Phi^\top) \mathbf{U}_\Phi]_i\|_2 \\
&\leq \frac{r^{1/2}}{n^{1/2} (\lambda_r^f)^{1/2} (1 - \epsilon_2)^{1/2}} \max_{j=1, \dots, r} \|(p^{-1} \mathbf{Y} \mathbf{Y}^\top - \Phi \Phi^\top) U_j\|_\infty \\
&\leq \frac{r^{1/2} \epsilon_3}{(\lambda_r^f)^{1/2} (1 - \epsilon_2)^{1/2}}.
\end{aligned} \tag{21}$$

where U_j is the j th column of \mathbf{U}_Φ .

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 (16)-(20) then show that there exists \tilde{c}_0 depending only on λ_1^f, λ_r^f such that

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

Now assuming

$$n \geq 2\sigma^2 r^{1/2} \tag{22}$$

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

$$\epsilon_1 r^{1/2} \leq 1/2 \quad (23)$$

we have

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

Applying this inequality in the above bound on $\|\mathbf{D}_5\|_F$ and allowing \tilde{c}_0 to increase where necessary we obtain:

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

Combining this estimate with (21) and again allowing \tilde{c}_0 to increase as needed,

$$\begin{aligned} \max_{i=1,\dots,n} \|[\mathbf{U}_Y \Lambda_Y^{1/2} - \mathbf{U}_\Phi \Lambda_\Phi^{1/2} \mathbf{F}_\star]_i\|_2 &\leq \max_{i=1,\dots,n} \|[(p^{-1} \mathbf{Y} \mathbf{Y}^\top - \Phi \Phi^\top) \mathbf{U}_\Phi \Lambda_\Phi^{-1/2} \mathbf{F}_\star]_i\|_2 + \sum_{i=1}^5 \|\mathbf{D}_i\|_F \\ &\leq r^{1/2} \tilde{c}_0 n^{1/2} \left(\epsilon_1 + \frac{\sigma^2}{n} \right) + r^{1/2} \tilde{c}_0 \epsilon_3. \end{aligned} \quad (24)$$

Now fix any $\epsilon \in (0, 1]$ and let us strengthen (22) to

$$n \geq \left(2\sigma^2 r^{1/2} \right) \vee \left(\frac{9}{\epsilon^2} \tilde{c}_0^2 r \sigma^4 \right) \quad (25)$$

so that $r^{1/2} \tilde{c}_0 n^{-1/2} \sigma^2 \leq \epsilon/3$. Then setting $\epsilon_1 := \epsilon/(3n^{1/2}r^{1/2}\tilde{c}_0)$ (which satisfies (23) since $\tilde{c}_0 \geq 1$), $\epsilon_3 := \epsilon/(3r^{1/2}\tilde{c}_0)$ and recalling that we have already chosen $\epsilon_2 := 1/4$ we have as a consequence of (24),

$$\mathbb{P} \left(\max_{i=1,\dots,n} \|[\mathbf{U}_Y \Lambda_Y^{1/2} - \mathbf{U}_\Phi \Lambda_\Phi^{1/2} \mathbf{F}_\star]_i\|_2 \leq \epsilon \right) \geq 1 - \mathbb{P}(A_1(\epsilon/[3n^{1/2}r^{1/2}\tilde{c}_0])^c) - \mathbb{P}(A_2(1/4)^c) - \mathbb{P}(A_3(\epsilon/[3r^{1/2}\tilde{c}_0])^c).$$

Now fix any $\delta \in (0, 1)$. By lemma 14, proposition 15 and lemma 17, there exists constants $\tilde{c}_1(q)$, \tilde{c}_2 and $\tilde{c}_3(q)$ (expressions for which could be deduced using the statements of the aforementioned results should the reader be so inclined) such that

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

Combining these conditions with (25) and appropriately defining c_1 and c_2 gives the conditions in the statement of the theorem. Recalling (12), the proof is complete. \square

B.3 Matrix estimates

Lemma 8. *Assume A1 and A3. 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

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

and

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

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:

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

and with σ_i denoting the i th singular value of $U_\Phi^\top U_Y$ and $\sigma_i = \cos(\theta_i)$, the Davis-Kahan $\sin(\theta)$ theorem gives:

$$\begin{aligned}\|U_\Phi^\top U_Y - U_\Phi^\top U_\Phi\|_2 &= \max_i |\sin(\theta_i)| \leq \frac{\|p^{-1}YY^\top - \Phi\Phi^\top\|_2}{|\lambda_r(\Phi\Phi^\top) - \lambda_{r+1}(p^{-1}YY^\top)|} \\ &\leq \frac{\epsilon_1 + n^{-1}\sigma^2}{\lambda_r^f(1 - 2\epsilon_2)}.\end{aligned}\quad (26)$$

Therefore

$$\begin{aligned}\|U_\Phi^\top U_Y - F_*\|_2 &= \|F_1\Sigma F_2^\top - F_1 F_2^\top\|_2 \\ &= \|F_1(\Sigma - I_r)F_2^\top\|_2 \\ &= \|\Sigma - I_r\|_2 \\ &= \max_{i=1,\dots,r} |1 - \sigma_i| \\ &\leq \max_{i=1,\dots,r} |1 - \sigma_i^2| = \max_{i=1,\dots,r} |\sin(\theta_i)|^2 \\ &\leq \left[\frac{\epsilon_1 + n^{-1}\sigma^2}{\lambda_r^f(1 - 2\epsilon_2)} \right]^2\end{aligned}$$

where for the first inequality uses $\|U_\Phi^\top U_Y\|_2 \leq 1$ and the second inequality is from (26). \square

Lemma 9. Assume **A1** and **A3**. 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

$$\begin{aligned}\|F_*\Lambda_Y - \Lambda_\Phi F_*\|_F &\leq r^{1/2} \left[n \frac{(\epsilon_1 + n^{-1}\sigma^2)^2}{\lambda_r^f(1 - 2\epsilon_2)} \left(1 + 2 \frac{\lambda_1^f}{\lambda_r^f} \left(\frac{1 + \epsilon_2}{1 - 2\epsilon_2} \right) \right) + n\epsilon_1 + \sigma^2 \right], \\ \|F_*\Lambda_Y^{1/2} - \Lambda_\Phi^{1/2} F_*\|_F &\leq \frac{\|F_*\Lambda_Y - \Lambda_\Phi F_*\|_F}{2n^{1/2}(1 - \epsilon_2)^{1/2}(\lambda_r^f)^{1/2}}, \\ \|F_*\Lambda_Y^{-1/2} - \Lambda_\Phi^{-1/2} F_*\|_F &\leq \frac{\|F_*\Lambda_Y^{1/2} - \Lambda_\Phi^{1/2} F_*\|_F}{n(1 - \epsilon_2)\lambda_r^f}.\end{aligned}$$

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

$$R := U_Y - U_\Phi U_\Phi^\top U_Y,$$

we have

$$\begin{aligned}F_*\Lambda_Y - \Lambda_\Phi F_* &= (F_* - U_\Phi^\top U_Y)\Lambda_Y + U_\Phi^\top(p^{-1}YY^\top - \Phi\Phi^\top)R \\ &\quad + U_\Phi^\top(p^{-1}YY^\top - \Phi\Phi^\top)U_\Phi U_\Phi^\top U_Y \\ &\quad + \Lambda_\Phi(U_\Phi^\top U_Y - F_*)\end{aligned}$$

hence

$$\|F_*\Lambda_Y - \Lambda_\Phi F_*\|_2 \leq \|U_\Phi^\top U_Y - F_*\|_2 (\|\Lambda_Y\|_2 + \|\Lambda_\Phi\|_2) \quad (27)$$

$$+ \|U_\Phi^\top(p^{-1}YY^\top - \Phi\Phi^\top)R\|_2 \quad (28)$$

$$+ \|U_\Phi^\top(p^{-1}YY^\top - \Phi\Phi^\top)U_\Phi U_\Phi^\top U_Y\|_2 \quad (29)$$

For the term on the r.h.s. of (27), on the event in the statement of the present lemma and using lemma 8 we have:

$$\|U_\Phi^\top U_Y - F_*\|_2 (\|\Lambda_Y\|_2 + \|\Lambda_\Phi\|_2) \leq \left[\frac{\epsilon_1 + n^{-1}\sigma^2}{\lambda_r^f(1 - 2\epsilon_2)} \right]^2 2n\lambda_1^f(1 + \epsilon_2).$$

For the term in (28), using $\mathbf{R} = (\mathbf{U}_Y \mathbf{U}_Y^\top - \mathbf{U}_\Phi^\top \mathbf{U}_\Phi) \mathbf{U}_Y$, we have again on the event in the statement of the present lemma and using lemma 8,

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

For the term in (29),

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

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

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

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

hence

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

and so

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

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

B.4 Some linear algebra

Lemma 10. *For any $m_1, m_2 \geq 1$, $\mathbf{A} \in \mathbb{R}^{m_1 \times m_2}$, $q \leq \min\{m_1, m_2\}$ and strictly positive real numbers $\lambda_1, \dots, \lambda_q$,*

- a) *there exists $\mathbf{U} \in \mathbb{R}^{m_2 \times q}$ such that $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_q$ and $\mathbf{A} \mathbf{A}^\top \mathbf{U} = \mathbf{U} \Lambda$, 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 $\mathbf{A}^\top \mathbf{A} \mathbf{V} = \mathbf{V} \Lambda$, where $\Lambda := \text{diag}(\lambda_1, \dots, \lambda_q)$;*
- 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 $\mathbf{U} = \mathbf{A} \mathbf{V} \Lambda^{-1/2}$;*
- c) *$\lambda_i(\mathbf{A}^\top \mathbf{A}) = \lambda_i(\mathbf{A} \mathbf{A}^\top)$, for $i = 1, \dots, \min\{m_1, m_2\}$.*
- 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 $\mathbf{U} := \mathbf{A} \mathbf{V} \Lambda^{-1/2}$ we have

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

and

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

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). \square

Lemma 11. For any $m_1 \leq m_2$ and $\mathbf{A} \in \mathbb{R}^{m_1 \times m_2}$ such that $\mathbf{A}\mathbf{A}^\top$ has rank m_1 , there exists an orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{m_1 \times m_1}$ such that $\mathbf{U}\Lambda^{1/2} = \mathbf{A}\mathbf{Q}$, where $\Lambda = \text{diag}\{\lambda_1(\mathbf{A}\mathbf{A}^\top), \dots, \lambda_{m_1}(\mathbf{A}\mathbf{A}^\top)\}$ and the columns of $\mathbf{U} \in \mathbb{R}^{m_2 \times m_1}$ are orthonormal eigenvectors of $\mathbf{A}\mathbf{A}^\top$ corresponding to $\lambda_1(\mathbf{A}\mathbf{A}^\top), \dots, \lambda_{m_1}(\mathbf{A}\mathbf{A}^\top)$.

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

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

and

$$\mathbf{Q}\mathbf{Q}^\top = \mathbf{A}^\top\mathbf{U}\Lambda^{-1}\mathbf{U}^\top\mathbf{A}. \quad (30)$$

Consider the reduced singular value decomposition $\mathbf{A} = \mathbf{U}\Lambda^{1/2}\mathbf{V}^\top$ where $\mathbf{V} \in \mathbb{R}^{m_1 \times m_1}$ has orthonormal columns. Substituting into the r.h.s. of (30),

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

□

B.5 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 12 (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 $\|\mathbf{M}_i\|_2 \leq L$ for each $1 \leq i \leq n$ and some constant L . Let $\mathbf{M} := \sum_{i=1}^n \mathbf{M}_i$ and $v(\mathbf{M}) = \max\{\|\mathbb{E}[\mathbf{M}\mathbf{M}^\top]\|_2, \|\mathbb{E}[\mathbf{M}^\top\mathbf{M}]\|_2\}$. Then for all $t \geq 0$,

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

Lemma 13. Assume **A1** and **A3**. For any $t \geq 0$,

$$\mathbb{P}\left(\|n^{-1}\Phi^\top\Phi - n^{-1}\mathbb{E}[\Phi^\top\Phi]\|_2 \geq t\right) \leq 2r \exp\left(\frac{-t^2n/2}{c_f^2 + c_f t/3}\right),$$

where $c_f := \sup_{z \in \mathcal{Z}} f(z, z) + \lambda_1^f$.

Proof. Apply theorem 12 with $\mathbf{M}_i = \frac{1}{n}\phi(Z_i)\phi(Z_i)^\top - \mathbb{E}[\frac{1}{n}\phi(Z_i)\phi(Z_i)^\top]$,

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

and

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

□

Lemma 14. Assume **A1**, **A3** and **A2** with some $q \geq 1$. Then for any $t > 0$,

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

where

$$c_X(q) := \max_{j=1,\dots,p} \sup_{z \in \mathcal{Z}} \mathbb{E} \left[|X_j(z)|^{2q} \right], \quad c_E(q) := \max_{j=1,\dots,p, i=1,\dots,n} \mathbb{E} \left[|\mathbf{E}_{ij}|^{2q} \right]$$

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

$$\mathbf{Y}\mathbf{Y}^\top = \sum_{j=1}^p Y_j Y_j^\top. \quad (31)$$

Observe that under the model of section 2, conditional on (Z_1, \dots, Z_n) the summands in (31) are independent and as per lemma 5, the conditional expectation of $\mathbf{Y}\mathbf{Y}^\top$ given Z_1, \dots, Z_n is: $p\Phi\Phi^\top + p\sigma^2\mathbf{I}_n$.

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, \dots, Y_p) given (Z_1, \dots, Z_n) . These inequalities taken together tell us that, for any $q \geq 1$, the following hold almost surely.

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

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

$$\Sigma := \frac{1}{2p^2} \sum_{j=1}^p \mathbb{E} \left[\left(Y_j Y_j^\top - \tilde{Y}_j \tilde{Y}_j^\top \right)^2 \mid Y_j, Z_1, \dots, Z_n \right], \quad (32)$$

where, conditional on Z_1, \dots, Z_n , \tilde{Y}_j is an independent copy of Y_j . For brevity in the remainder of the proof we shall write $Z \equiv (Z_1, \dots, Z_n)$, 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:

$$\begin{aligned} \mathbb{E} \left[\|\Sigma\|_{S_q}^q \mid Z \right]^{1/q} &= \frac{1}{2p^2} \mathbb{E} \left[\left\| \sum_{j=1}^p \mathbb{E} \left[\left(Y_j Y_j^\top - \tilde{Y}_j \tilde{Y}_j^\top \right)^2 \mid Y_j, Z \right] \right\|_{S_q}^q \mid Z \right]^{1/q} \\ &\leq \frac{1}{2p^2} \sum_{j=1}^p \mathbb{E} \left[\left\| \mathbb{E} \left[\left(Y_j Y_j^\top - \tilde{Y}_j \tilde{Y}_j^\top \right)^2 \mid Y_j, Z \right] \right\|_{S_q}^q \mid Z \right]^{1/q} \end{aligned} \quad (33)$$

$$\leq \frac{1}{2p^2} \sum_{j=1}^p \mathbb{E} \left[\left\| \left(Y_j Y_j^\top - \tilde{Y}_j \tilde{Y}_j^\top \right)^2 \right\|_{S_q}^q \mid Z \right]^{1/q} \quad (34)$$

$$= \frac{1}{2p^2} \sum_{j=1}^p \mathbb{E} \left[\left\| Y_j Y_j^\top - \tilde{Y}_j \tilde{Y}_j^\top \right\|_{S_{2q}}^{2q} \mid Z \right]^{1/q} \quad (35)$$

$$\leq \frac{1}{2p^2} \sum_{j=1}^p \left(2 \mathbb{E} \left[\|Y_j Y_j^\top\|_{S_{2q}}^{2q} \mid Z \right]^{1/2q} \right)^2$$

$$= \frac{2}{p^2} \sum_{j=1}^p \mathbb{E} \left[\|Y_j Y_j^\top\|_{S_{2q}}^{2q} \mid Z \right]^{1/q}$$

Here (33) holds by the second claim of lemma 16; 34 holds by first claim of lemma 16 combined with the fact that $x \mapsto x^q$ is convex for $x \geq 0$ (recall $q \geq 1$); (35) holds by lemma 16 and the fact that \tilde{Y}_j and Y_j are equal in distribution.

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

$$\|Y_j Y_j^\top\|_{S_{2q}}^{2q} = \|Y_j\|_2^{4q} = \left| \sum_{i=1}^n (X_j(Z_i) + \sigma \mathbf{E}_{ij})^2 \right|^{2q}. \quad (36)$$

By two applications of Minkowski's inequality,

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

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

Combining the above estimates we find:

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

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

Proposition 15. Assume **A1**, **A3** and **A2** with some $q \geq 1$. For any $\delta, \epsilon \in (0, 1)$, if

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

and

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

where c_X and c_E are as in lemma 13, then

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

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 $\lambda_i(\Phi \Phi^\top) = 0$ for $r+1 \leq i \leq n$ which holds since $\Phi \in \mathbb{R}^{n \times r}$.

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

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

Combining this decomposition with Weyl's inequality; the facts that for $1 \leq i \leq r$, $\mathbb{E}[\Phi^\top \Phi]_{ii} = n \lambda_i^f$ and $\mathbb{E}[\Phi^\top \Phi]_{ij} = 0$ for $j \neq i$, hence $\lambda_i^f = \lambda_i(n^{-1} \mathbb{E}[\Phi^\top \Phi])$; and by lemma 10, $\lambda_i(\Phi \Phi^\top) = \lambda_i(\Phi^\top \Phi)$; whilst for $i \geq r+1$, $\lambda_i(\Phi \Phi^\top) = \lambda_i^f = 0$; we obtain:

$$\begin{aligned} \max_{1 \leq i \leq n} \left| \frac{1}{n} \lambda_i(p^{-1} \mathbf{Y} \mathbf{Y}^\top) - \lambda_i^f \right| &\leq \frac{1}{n} \|p^{-1} \mathbf{Y} \mathbf{Y}^\top - \Phi \Phi^\top - \sigma^2 \mathbf{I}_n\|_2 \\ &\quad + \frac{\sigma^2}{n} \\ &\quad + \|n^{-1} \Phi^\top \Phi - n^{-1} \mathbb{E}[\Phi^\top \Phi]\|_2 \end{aligned} \quad (37)$$

and

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

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

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

where the second inequality holds by using $\lambda_r^f \leq \lambda_i^f$ for $i = 1, \dots, r$, together with (37) and the condition of the proposition $n \geq 3\sigma^2/(\epsilon \lambda_r^f)$; and the third inequality holds by applying lemma 13 and lemma 14.

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

$$\delta/2 \geq (16)^q (2q-1)^q \frac{1}{(\epsilon \lambda_r^f / 3)^{2q}} \frac{1}{p^q} \left(c_X (2q)^{1/2q} + \sigma^2 c_E (2q)^{1/2q} \right)^{2q},$$

$$\frac{\delta}{2} \geq 2r \exp \left(\frac{-(\epsilon/3)^2 (\lambda_r^f)^2 n/2}{c_f^2 + c_f \epsilon \lambda_r^f / 9} \right).$$

□

Lemma 16. For any $m_1, m_2 \geq 1$ 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 $1 \leq q < \infty$ such that $\mathbb{E}[\|\mathbf{A}\|_\star^q] \vee \mathbb{E}[\|\mathbf{B}\|_\star^q] < \infty$, $\mathbb{E}[\|\mathbf{A} + \mathbf{B}\|_\star^{q/1}] \leq \mathbb{E}[\|\mathbf{A}\|_\star^{q/1}] + \mathbb{E}[\|\mathbf{B}\|_\star^{q/1}]$.

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 $\mathbb{E}[\|\mathbf{A} + \mathbf{B}\|_\star^q] \leq 2^{q-1} (\mathbb{E}[\|\mathbf{A}\|_\star^q] + \mathbb{E}[\|\mathbf{B}\|_\star^q]) < \infty$. 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,

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

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

Lemma 17. Assume **A1**, **A3**, and **A2** with some $q \geq 1$. Let U_j denote the j th column of \mathbf{U}_Φ . Then there exists a constant $b(q)$ depending only on q such that for any $t > 0$,

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

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

$$p^{-1} \sum_{k=1}^p \Delta_{ij}(k)$$

where

$$\Delta_{ij}(k) := Y_k^{(i)} Y_k^\top U_j - \mathbb{E} \left[Y_k^{(i)} Y_k^\top U_j \mid Z_1, \dots, Z_n \right]$$

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

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

$$\begin{aligned} \mathbb{P} \left(\left| p^{-1} \sum_{k=1}^p \Delta_{ij}(k) \right| \geq t \mid Z \right) &\leq \frac{1}{t^{2q}} \mathbb{E} \left[\left| p^{-1} \sum_{k=1}^p \Delta_{ij}(k) \right|^{2q} \mid Z \right] \\ &\leq \frac{b(2q)}{t^{2q} p^{2q}} \mathbb{E} \left[\left| \sqrt{\sum_{k=1}^p |\Delta_{ij}(k)|^2} \right|^{2q} \mid Z \right] \\ &\leq \frac{b(2q)}{t^{2q} p^{2q}} \left(\sum_{k=1}^p \mathbb{E} [|\Delta_{ij}(k)|^{2q} \mid Z]^{1/q} \right)^q \\ &= \frac{b(2q)}{t^{2q} p^q} \max_{k=1, \dots, p} \mathbb{E} [|\Delta_{ij}(k)|^{2q} \mid Z]. \end{aligned} \quad (38)$$

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

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

and so

$$\begin{aligned} \mathbb{E} [|\Delta_{ij}(k)|^{2q} \mid Z] &\leq 2^{2q} \mathbb{E} \left[(|Y_k^{(i)}|^2 \|Y_k\|_2^2)^q \mid Z \right] \\ &= 2^{2q} \mathbb{E} \left[\left(\sum_{l=1}^n |Y_k^{(i)}|^2 |Y_k^{(l)}|^2 \right)^q \mid Z \right] \\ &\leq 2^{2q} \left(\sum_{l=1}^n \mathbb{E} [(|Y_k^{(i)}|^2 |Y_k^{(l)}|^2)^q \mid Z]^{1/q} \right)^q \\ &\leq 2^{2q} \left(\sum_{l=1}^n \mathbb{E} [|Y_k^{(i)}|^{4q} \mid Z]^{1/(2q)} \mathbb{E} [|Y_k^{(l)}|^{4q} \mid Z]^{1/(2q)} \right)^q \\ &= 2^{2q} n^q \max_{l=1, \dots, n} \mathbb{E} [|Y_k^{(l)}|^{4q} \mid Z] \\ &= 2^{2q} n^q \max_{l=1, \dots, n} \mathbb{E} [|X_k(Z_l) + \sigma \mathbf{E}_{kl}|^{4q} \mid Z] \\ &\leq 2^{6q-1} n^q \left(\max_{l=1, \dots, p} \sup_{z \in \mathcal{Z}} \mathbb{E} [|X_l(z)|^{4q}] + \sigma^{4q} \max_{l=1, \dots, p, \tilde{l}=1, \dots, n} \mathbb{E} [|\mathbf{E}_{l\tilde{l}}|^{4q}] \right). \end{aligned} \quad (39)$$

Combining the almost sure upper bounds (39) and (38), using the tower property of conditional expectation and then taking a union bound over $i = 1, \dots, n$ and $j = 1, \dots, r$, we find:

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

which completes the proof. \square

C Supporting material for section 4

C.1 d -dimensional differentiable manifolds

Definition 18 extends the elementary calculus notion of continuous differentiability to functions on real domains which may not be open sets. The important point here is that if a function, say $\psi : U \rightarrow \mathbb{R}$ has a domain U which is an arbitrary set in \mathbb{R}^d , and ξ is point in U , then the difference quotient used to define the usual notion of directional derivative $[\psi(\xi + hv) - \psi(\xi)]/h$ may not be well defined, since there is no guarantee that $\xi + hv \in U$. Definition 18 deals with this issue and is then incorporated in definition 19, in turn leading to definition 20. This presentation very closely follows [Guillemen and Pollack \[1974\]](#), with some slight differences in terminology to give just what we need for the present work.

Definition 18 (C^1 functions on arbitrary domains). For any $d_1, d_2 \in \mathbb{N}$ and an arbitrary set $U \subset \mathbb{R}^{d_1}$, a function $\psi : U \rightarrow \mathbb{R}^{d_2}$ is said to be C^1 if for each $\xi \in U$ there exists an open set $\tilde{U} \subset \mathbb{R}^{d_1}$ and a function $\tilde{\psi} : \tilde{U} \rightarrow \mathbb{R}^{d_2}$ such that $\xi \in \tilde{U}$, $\tilde{\psi}$ equals ψ on $\tilde{U} \cap U$, and $\tilde{\psi}$ is continuously differentiable as an \mathbb{R}^{d_2} -valued function on \tilde{U} in the sense of elementary calculus.

Definition 19 (Diffeomorphisms with an arbitrary domain in Euclidean space). For any $d_1, d_2 \in \mathbb{N}$ an arbitrary set $U \subset \mathbb{R}^{d_2}$ and an open set $V \subset \mathbb{R}^{d_1}$, a mapping $\psi : V \rightarrow U$ is called a *diffeomorphism* if it is a homeomorphism between V and U , ψ is continuously differentiable in the usual sense as an \mathbb{R}^{d_2} -valued function on V , and its inverse ψ^{-1} on U is C^1 in the sense of definition 18. U and V are said to be *diffeomorphic* if such a map exists.

Definition 20 (d -dimensional differentiable manifold). For any $d_1 \leq d_2$ a set $\mathcal{A} \subset \mathbb{R}^{d_2}$ is called a d_1 -dimensional differentiable manifold if for each $a \in \mathcal{A}$, there exists a set $U \subseteq \mathcal{A}$ and an open set $V \subset \mathbb{R}^{d_1}$ such that $a \in U$, and U and V are diffeomorphic in the sense of definition 19.

C.2 Proofs and supporting results for section 4

Proposition 21. Assume [A1](#) and [A4](#). If for some given function $\eta : [0, 1] \rightarrow \mathcal{Z}$, the function $(s, s') \in [0, 1]^2 \mapsto f(\eta_s, \eta_{s'}) \in \mathbb{R}$ has a derivative $\frac{\partial^2}{\partial s \partial s'} f(\eta_s, \eta_{s'})$ which is continuous everywhere on $[0, 1]^2$, then $\gamma : [0, 1] \rightarrow \mathcal{M}$ defined by $t \mapsto \gamma_t := \phi(\eta_t)$ is a curve in \mathcal{M} . Furthermore,

$$\int_0^1 \|\dot{\gamma}_t\| dt = \int \left| \frac{\partial^2}{\partial s \partial s'} f(\eta_s, \eta_{s'}) \right|_{(t,t)}^{1/2} dt. \quad (40)$$

Proof of Proposition 21. The essence of the proof is to notice $(s, s') \mapsto f(\eta_s, \eta_{s'})$ is a kernel on $[0, 1]^2$ with feature map $s \mapsto \phi(\eta_s)$, and then apply a result of [Steinwart and Christmann \[2008, lemma 4.34\]](#) about differentiability of kernel feature maps. We provide all the details of the proof for completeness.

Throughout the proof η and γ are fixed as in the statement. Define

$$\Delta_h \phi(t) := \phi(\eta_{t+h}) - \phi(\eta_t).$$

We will prove that for any sequence $h_n \rightarrow 0$, the sequence $(\Delta_{h_n} \phi(t)/h_n)_{n \geq 1}$ is Cauchy, from which the existence of the limit $\dot{\gamma}_t$ follows by completeness of ℓ_2 . For any $t \in [0, 1]$, we have:

$$\|h_n^{-1} \Delta_{h_n} \phi(t) - h_m^{-1} \Delta_{h_m} \phi(t)\|_2^2 = \langle h_n^{-1} \Delta_{h_n} \phi(t), h_n^{-1} \Delta_{h_n} \phi(t) \rangle_2 \quad (41)$$

$$+ \langle h_m^{-1} \Delta_{h_m} \phi(t), h_m^{-1} \Delta_{h_m} \phi(t) \rangle_2 \quad (42)$$

$$- 2 \langle h_n^{-1} \Delta_{h_n} \phi(t), h_m^{-1} \Delta_{h_m} \phi(t) \rangle_2. \quad (43)$$

Defining the function $F(z) := f(\eta_{t+h_n}, z) - f(\eta_t, z)$, we have

$$\langle \Delta_{h_m} \phi(t), \Delta_{h_m} \phi(t') \rangle_2 = F(\eta_{t+h_m}) - F(\eta_{t'}).$$

By the mean value theorem, there exists $\xi'_{m,n} \in [-|h_m|, |h_m|]$ such that

$$\begin{aligned} \langle \Delta_{h_m} \phi(t), \Delta_{h_m} \phi(t') \rangle_2 &= \frac{\partial}{\partial s'} F(\eta_{s'}) \Big|_{t'+\xi'_{m,n}} \\ &= \frac{\partial}{\partial s'} f(\eta_{t+h_m}, \eta_{s'}) \Big|_{t'+\xi'_{m,n}} - \frac{\partial}{\partial s'} f(\eta_t, \eta_{s'}) \Big|_{t'+\xi'_{m,n}}. \end{aligned}$$

By a second application of the mean value theorem, there exists $\xi_{n,m} \in [-|h_n|, |h_n|]$ such that

$$\langle h_n^{-1} \Delta_{h_m} \phi(t), h_m^{-1} \Delta_{h_m} \phi(t') \rangle_2 = \frac{\partial^2}{\partial s \partial s'} f(\eta_s, \eta_{s'}) \Big|_{(t+\xi_{n,m}, t'+\xi'_{n,m})}.$$

By the continuity of $\frac{\partial^2}{\partial s \partial s'} f(\eta_s, \eta_{s'})$, for any $\epsilon > 0$ there exists n_0 such that for $m, n \geq n_0$,

$$\left| \langle h_n^{-1} \Delta_{h_m} \phi(t), h_m^{-1} \Delta_{h_m} \phi(t') \rangle_2 - \frac{\partial^2}{\partial s \partial s'} f(\eta_s, \eta_{s'}) \Big|_{(t,t')} \right| \leq \epsilon. \quad (44)$$

Applying (44) with $t = t'$ to each of the terms (41)-(43) we find that $(\Delta_{h_n} \phi(t)/h_n)_{n \geq 1}$ is a Cauchy sequence as required. With $\dot{\gamma}_t$ defined to be the associated limit in ℓ_2 , we have proved that:

$$\lim_{h \rightarrow 0} \left\| \frac{\gamma_{t+h} - \gamma_t}{h} - \dot{\gamma}_t \right\|_2 = 0.$$

Another consequence of (44) is that for any $t, t' \in [0, 1]$,

$$\langle \dot{\gamma}_t, \dot{\gamma}_{t'} \rangle_2 = \frac{\partial^2}{\partial s \partial s'} f(\eta_s, \eta_{s'}) \Big|_{(t,t')}. \quad (45)$$

The continuity of $t \mapsto \dot{\gamma}_t$ follows by combining the above identity with the assumed continuity of $\frac{\partial^2}{\partial s \partial s'} f(\eta_s, \eta_{s'})$ and:

$$\|\dot{\gamma}_t - \dot{\gamma}_{t'}\|_2^2 = \langle \dot{\gamma}_t, \dot{\gamma}_t \rangle_2 + \langle \dot{\gamma}_{t'}, \dot{\gamma}_t \rangle_2 - 2 \langle \dot{\gamma}_t, \dot{\gamma}_{t'} \rangle_2.$$

A special case of (45) is:

$$\|\dot{\gamma}_t\|_2^2 = \frac{\partial^2}{\partial s \partial s'} f(\eta_s, \eta_{s'}) \Big|_{(t,t)},$$

from which it follows that:

$$\int_0^1 \|\dot{\gamma}_t\|_2 dt = \int_0^1 \left| \frac{\partial^2}{\partial s \partial s'} f(\eta_s, \eta_{s'}) \Big|_{(t,t)} \right|^{1/2} dt.$$

□

Lemma 22. Assume **A1**, **A4** and **A5**. Then ϕ^{-1} is Lipschitz continuous as a map from \mathcal{M} (equipped with the distance $\|\cdot - \cdot\|_2$) to $\mathbb{R}^{\tilde{d}}$ (equipped with the distance $\|\cdot - \cdot\|_{\mathbb{R}^{\tilde{d}}}$).

Proof. See [Whiteley et al., 2021, Appendix A.1, Proposition 4]. □

When assumption **A5** holds let $\tilde{\mathbf{H}}_{z,z'} \in \mathbb{R}^{\tilde{d} \times \tilde{d}}$ be the matrix whose elements are the partial derivatives:

$$(\tilde{\mathbf{H}}_{z,z'})_{ij} := \frac{\partial^2 f}{\partial z_i \partial z'_j} \Big|_{(z,z')}, \quad (46)$$

so $\tilde{\mathbf{H}}_{z,z} = \mathbf{H}_z$.

Proposition 23. Assume **A1**, **A4** and **A5**. Fix any $z_0 \in \mathcal{Z}$. Then:

- i) there exists a set $U \subseteq \mathcal{Z}$, an open set $V \subset \mathbb{R}^d$ and a mapping $\psi : V \rightarrow U$ such that $z_0 \in U$ and ψ is a diffeomorphism between V and U ;
- ii) for each $i = 1, \dots, d$ and ξ in V , there exists $\partial_i(\phi \circ \psi)_\xi \in \ell_2$ such that

$$\lim_{h \rightarrow 0} \left\| \frac{\phi \circ \psi(\xi + he_i) - \phi \circ \psi(\xi)}{h} - \partial_i(\phi \circ \psi)_\xi \right\|_2 = 0$$

and $\xi \mapsto \partial_i(\phi \circ \psi)_\xi$ is continuous on V ;

iii) with $\partial(\phi \circ \psi)_\xi := [\partial_1(\phi \circ \psi)_\xi | \dots | \partial_d(\phi \circ \psi)_\xi]$, i.e. a matrix with d columns and rows indexed by \mathbb{N} ; and with $\mathbf{J}_\xi \in \mathbb{R}^{\tilde{d} \times d}$ defined to be the Jacobian matrix of ψ evaluated at ξ in the open set V , we have for any $\xi, \xi' \in V$,

$$\partial(\phi \circ \psi)_\xi^\top \partial(\phi \circ \psi)_{\xi'} = \mathbf{J}_\xi^\top \tilde{\mathbf{H}}_{\psi(\xi), \psi(\xi')} \mathbf{J}_{\xi'}, \quad (47)$$

and the matrix $\mathbf{J}_\xi^\top \mathbf{H}_{\psi(\xi)} \mathbf{J}_\xi$ is positive-definite.

Proof. As per the statement of the proposition, fix some $z_0 \in \mathcal{Z}$. The reader is alerted to the fact that U, V, ψ and other quantities appearing in the statement of the proposition and in the proof below may depend on z_0 ; this dependence is not shown in the notation but is not problematic since z_0 is fixed throughout.

Part i) of the proposition holds since by assumption **A5**, \mathcal{Z} is a d -dimensional differentiable manifold in the sense of definition 20. Part ii) is proved by exactly similar arguments to those used in the proof of proposition 21 applied to the map $\xi \mapsto \phi \circ \psi(\xi)$ for ξ in the open set V , so the details are omitted. Those arguments combined with the chain rule of elementary calculus also yield the identity (47) in part iii), which is the analogue of the identity (45).

To complete the proof of part iii) it remains to show that $\mathbf{J}_\xi^\top \mathbf{H}_{\psi(\xi)} \mathbf{J}_\xi$ is positive-definite, for any $\xi \in V$. Fix any $\xi_0 \in V$, set $u_0 = \psi(\xi_0)$ and note $u_0 \in U$. Since $\mathbf{H}_{\psi(\xi_0)}$ is positive-definite by assumption **A5**, we need to show that the columns of \mathbf{J}_{ξ_0} are linearly independent. To this end we note that by part i), $\psi : V \rightarrow U$ is a diffeomorphism, hence by definitions 19 and 18 there exists an open set $\tilde{U} \subset \mathbb{R}^{\tilde{d}}$ and a mapping $\widetilde{\psi^{-1}} : \tilde{U} \rightarrow \mathbb{R}^d$ such $u_0 \in \tilde{U}$, $\widetilde{\psi^{-1}}$ agrees with ψ^{-1} on $U \cap \tilde{U}$, and $\widetilde{\psi^{-1}}$ is continuously differentiable on \tilde{U} . Then

$$\psi^{-1} \circ \psi(\xi_0) = \widetilde{\psi^{-1}} \circ \psi(\xi_0) = \xi_0,$$

and writing \mathbf{K}_{u_0} for the Jacobian matrix of $\widetilde{\psi^{-1}}$ evaluated at u_0 , the chain rule gives:

$$\mathbf{K}_{u_0} \mathbf{J}_{\xi_0} = \mathbf{I}_d. \quad (48)$$

Suppose that for some $v \in \mathbb{R}^d$ such that $v \neq 0$, we have $\mathbf{J}_{\xi_0} v = 0$. Then $\mathbf{K}_{u_0} \mathbf{J}_{\xi_0} v = \mathbf{K}_{z_0} 0 = 0 \neq v$, contradicting (48). Therefore the columns of \mathbf{J}_{ξ_0} must be linearly independent, and as ξ_0 was an arbitrary point in V , the proof is complete. \square

Proof of proposition 3. For part i), let η be a curve in \mathcal{Z} . Then under assumption **A5**, $\left. \frac{\partial^2}{\partial s \partial s'} f(\eta_s, \eta_{s'}) \right|_{(t, t')}$ exists and is a continuous function of t, t' by the chain rule of elementary calculus, indeed:

$$\left. \frac{\partial^2}{\partial s \partial s'} f(\eta_s, \eta_{s'}) \right|_{(t, t')} = \left\langle \dot{\eta}_t, \widetilde{\mathbf{H}}_{\eta_t, \eta_{t'}, \dot{\eta}_{t'}} \right\rangle.$$

Hence by proposition 21, $\gamma : [0, 1] \rightarrow \mathcal{M}$ defined by $\gamma_t := \phi(\eta_t)$ is a curve in \mathcal{M} . Thus part i) of the proposition holds.

For part ii), let γ be a curve in \mathcal{Z} . Let $\eta : [0, 1] \rightarrow \mathcal{Z}$ be defined by $\eta_t := \phi^{-1}(\gamma_t)$. Taking the metric associated with \mathcal{Z} to be $d_{\mathcal{Z}}(\cdot, \cdot) = \|\cdot - \cdot\|_{\mathbb{R}^{\tilde{d}}}$, by lemma 2, ϕ is a homeomorphism, so $t \mapsto \eta_t$ is continuous as an $\mathbb{R}^{\tilde{d}}$ -valued function.

Fix any $t_0 \in [0, 1]$. Applying proposition 23 with z_0 there taken to η_{t_0} , there exist a set $U \subseteq \mathcal{Z}$, an open set $V \subset \mathbb{R}^d$ and a $\psi : V \rightarrow U$ such that $\eta_{t_0} \in U$ and ψ is a diffeomorphism between V and U , and hence by definitions 18–19, ψ is a homeomorphism between V and U , there exists an open set $\tilde{U} \subset \mathbb{R}^{\tilde{d}}$ and $\widetilde{\psi^{-1}} : \tilde{U} \rightarrow \mathbb{R}^d$ such that $\eta_{t_0} \in \tilde{U}$, $\widetilde{\psi^{-1}}$ equals ψ^{-1} on $U \cap \tilde{U}$ and $\widetilde{\psi^{-1}}$ is continuously differentiable on \tilde{U} . We shall now prove that $t \mapsto \widetilde{\psi^{-1}}(\eta_t)$ is continuously differentiable at t_0 , and then continuous differentiability of η_t at t_0 will follow from the identity $\eta_{t_0} = \psi \circ \widetilde{\psi^{-1}}(\eta_{t_0})$ and the continuous differentiability of ψ .

Since $t \mapsto \eta_t$ is continuous on $[0, 1]$, U is homeomorphic to the open set V and \tilde{U} is open, for $|h|$ small enough we have $\eta_{t_0+h} \in U \cap \tilde{U}$. Therefore $\widetilde{\psi^{-1}}(\eta_{t_0+h})$ is well-defined and equal to $\psi^{-1}(\eta_{t_0+h})$. With $\xi_0 := \widetilde{\psi^{-1}}(\eta_{t_0})$ and $\xi_h := \widetilde{\psi^{-1}}(\eta_{t_0+h})$, consider the decomposition:

$$\begin{aligned} & \frac{\widetilde{\psi^{-1}}(\eta_{t_0+h}) - \widetilde{\psi^{-1}}(\eta_{t_0})}{h} - (\mathbf{J}_{\xi_0}^\top \mathbf{H}_{\eta_{t_0}} \mathbf{J}_{\xi_0})^{-1} \partial(\phi \circ \psi)_{\xi_0}^\top \dot{\gamma}_{t_0} \\ &= (\mathbf{J}_{\xi_0}^\top \mathbf{H}_{\eta_{t_0}} \mathbf{J}_{\xi_0})^{-1} \partial(\phi \circ \psi)_{\xi_0}^\top \left[\partial(\phi \circ \psi)_{\xi_0} \frac{(\xi_h - \xi_0)}{h} - \frac{\phi \circ \psi(\xi_h) - \phi \circ \psi(\xi_0)}{h} \right] \end{aligned} \quad (49)$$

$$+ (\mathbf{J}_{\xi_0}^\top \mathbf{H}_{\eta_{t_0}} \mathbf{J}_{\xi_0})^{-1} \partial(\phi \circ \psi)_{\xi_0}^\top \left[\frac{\gamma_{t_0+h} - \gamma_{t_0}}{h} - \dot{\gamma}_t \right], \quad (50)$$

where the inverse of the matrix $\mathbf{J}_{\xi_0}^\top \mathbf{H}_{\eta_{t_0}} \mathbf{J}_{\xi_0}$ exists by part iii) of proposition 23, and the identities (47) and $\gamma_{t_0+h} = \phi(\eta_{t_0+h}) = \phi \circ \psi(\widetilde{\psi^{-1}}(\eta_{t_0+h})) = \phi \circ \psi(\xi_h)$ have been used.

Using the fundamental theorem of calculus, the bracketed term in (49) can be written as:

$$\begin{aligned} &= \partial(\phi \circ \psi)_{\xi_0} \frac{(\xi_h - \xi_0)}{h} - \frac{\phi \circ \psi(\xi_h) - \phi \circ \psi(\xi_0)}{h} \\ &= \int_0^1 [\partial(\phi \circ \psi)_{\xi_0} - \partial(\phi \circ \psi)_{\xi_0+s(\xi_h-\xi_0)}] \frac{(\xi_h - \xi_0)}{h} ds. \end{aligned} \quad (51)$$

Defining $\mathbf{B}_{s,h} := \partial(\phi \circ \psi)_{\xi_0} - \partial(\phi \circ \psi)_{\xi_0+s(\xi_h-\xi_0)}$ and applying (47) with the shorthand $\mathbf{G}_{\xi,\xi'} := \mathbf{J}_\xi^\top \tilde{\mathbf{H}}_{\xi,\xi'} \mathbf{J}_{\xi'}$,

$$\begin{aligned} &\left\| [\partial(\phi \circ \psi)_{\xi_0} - \partial(\phi \circ \psi)_{\xi_0+s(\xi_h-\xi_0)}] \frac{(\xi_h - \xi_0)}{h} \right\|_2^2 \\ &= \frac{1}{h^2} \langle (\xi_h - \xi_0), \mathbf{B}_{s,h}^\top \mathbf{B}_{s,h} (\xi_h - \xi_0) \rangle_2 \\ &= \frac{1}{h^2} \langle (\xi_h - \xi_0), [\mathbf{G}_{\xi_0,\xi_0} - \mathbf{G}_{\xi_0,\xi_0+s(\xi_h-\xi_0)}] (\xi_h - \xi_0) \rangle_2 \\ &\quad + \frac{1}{h^2} \langle (\xi_h - \xi_0), [\mathbf{G}_{\xi_0,\xi_0+s(\xi_h-\xi_0)} - \mathbf{G}_{\xi_0+s(\xi_h-\xi_0),\xi_0+s(\xi_h-\xi_0)}] (\xi_h - \xi_0) \rangle_2. \end{aligned} \quad (52)$$

$$+ \frac{1}{h^2} \langle (\xi_h - \xi_0), [\mathbf{G}_{\xi_0,\xi_0+s(\xi_h-\xi_0)} - \mathbf{G}_{\xi_0+s(\xi_h-\xi_0),\xi_0+s(\xi_h-\xi_0)}] (\xi_h - \xi_0) \rangle_2. \quad (53)$$

Using the continuous differentiability of ψ and the continuity in ξ, ξ' of the elements of $\tilde{\mathbf{H}}_{\xi,\xi'}$, for any $\epsilon > 0$, there exists δ such that $\|\xi_h - \xi_0\|_{\mathbb{R}^d} \leq \delta$ implies for all $s \in [0, 1]$, the sum of (52) and (53) is less than $\epsilon \|\xi_h - \xi_0\|_{\mathbb{R}^d}^2$. Moreover, by lemma 22, the assumption of part ii) that γ is a curve and the C^1 property of ψ^{-1} , there exists a finite constant c such that $\|\xi_h - \xi_0\|_{\mathbb{R}^d} = \|\psi^{-1} \circ \phi^{-1}(\gamma_{t+h}) - \psi^{-1} \circ \phi^{-1}(\gamma_t)\|_{\mathbb{R}^d} \leq |h|c$. Combining this fact with (51) and the continuity of $t \mapsto \eta_t = \phi^{-1}(\gamma_t)$, we conclude that for any $\epsilon > 0$ there exists δ such that

$$|h| \leq \delta \Rightarrow \left\| \partial(\phi \circ \psi)_{\xi_0} \frac{(\xi_h - \xi_0)}{h} - \frac{\phi \circ \psi(\xi_h) - \phi \circ \psi(\xi_0)}{h} \right\|_2 \leq \epsilon.$$

By definition of $\dot{\gamma}_t$, $\lim_{h \rightarrow 0} \|h^{-1}(\gamma_{t+h} - \gamma_t) - \dot{\gamma}_t\|_2 = 0$, and returning to (49)-(50) and applying the Cauchy-Schwartz inequality we conclude that:

$$\lim_{h \rightarrow 0} \left\| \frac{\psi^{-1}(\eta_{t_0+h}) - \psi^{-1}(\eta_{t_0})}{h} - (\mathbf{J}_{\xi_0}^\top \mathbf{H}_{\eta_{t_0}} \mathbf{J}_{\xi_0})^{-1} \partial(\phi \circ \psi)_{\xi_0}^\top \dot{\gamma}_{t_0} \right\|_{\mathbb{R}^d} = 0,$$

thus $t \mapsto \psi^{-1}(\eta_t)$ is differentiable at t_0 . The continuity of the derivative at t_0 follows from the continuity of each of the terms constituting $(\mathbf{J}_{\xi_0}^\top \mathbf{H}_{\eta_{t_0}} \mathbf{J}_{\xi_0})^{-1} \partial(\phi \circ \psi)_{\xi_0}^\top \dot{\gamma}_{t_0}$, recalling the definition $\xi_0 = \psi^{-1}(\eta_{t_0})$. Applying the chain rule and the identity $\eta_t = \psi \circ \psi^{-1}(\eta_t)$ we find that $t \mapsto \eta_t$ is continuously differentiable at t_0 , which completes the proof of part ii) of the proposition.

For part iii) of the proposition, either η is a curve by assumption, or γ is a curve by assumption and then η is a curve by part ii) of the proposition. The required identity follows from proposition 21 combined with the chain rule. \square

D Supplementary figures for Section 5.8

E Code and data

All code for this paper and links to the data are available at: <https://github.com/anniegray52/pca>

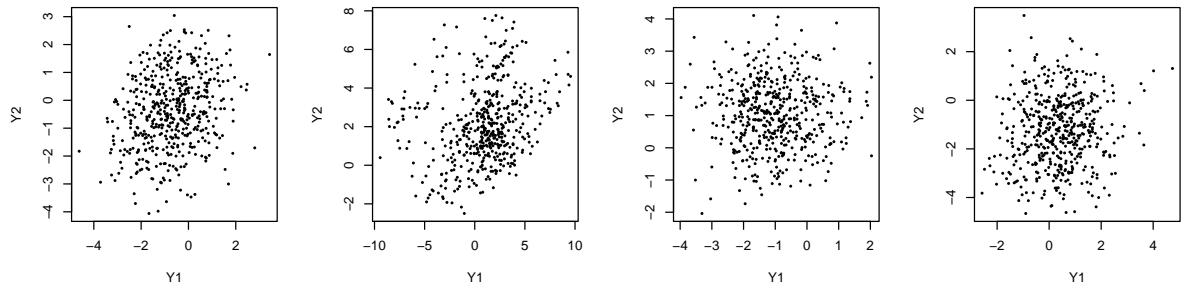


Figure 14: First two coordinates of the data matrices corresponding to figure 13, showing much less structure than the principal components.

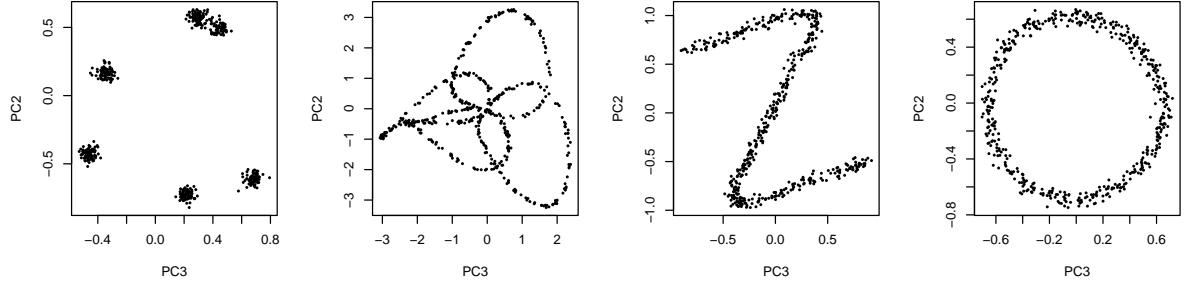


Figure 15: Third and second principal components of the data matrices corresponding to figure 13 (ordered like this to make the resemblance to \mathcal{Z} more obvious).

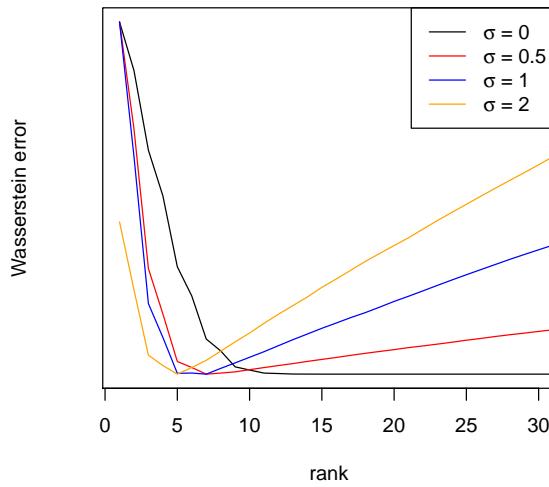


Figure 16: log-Wasserstein error for the fourth configuration in figure 13, for different error variances. As the variance increases, the optimal dimension (point achieving lowest error) decreases. The curves are shifted and rescaled so that their maxima and minima agree.