Matrix factorisation and the interpretation of geodesic distance

Nick Whiteley
University of Bristol
nick.whiteley@bristol.ac.uk

Annie Gray
University of Bristol
annie.gray@bristol.ac.uk

Patrick Rubin-Delanchy
University of Bristol
patrick.rubin-delanchy@bristol.ac.uk

Abstract

Given a graph or similarity matrix, we consider the problem of recovering a notion of true distance between the nodes, and so their true positions. Through new insights into the manifold geometry underlying a generic latent position model, we show that this can be accomplished in two steps: matrix factorisation, followed by nonlinear dimension reduction. This combination is effective because the point cloud obtained in the first step lives close to a manifold in which latent distance is encoded as geodesic distance. Hence, a nonlinear dimension reduction tool, approximating geodesic distance, can recover the latent positions, up to a simple transformation. We give a detailed account of the case where spectral embedding is used, followed by Isomap, and provide encouraging experimental evidence for other combinations of techniques.

1 Introduction

Assume we observe, or are given as the result of a computational procedure, data in the form of a symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ which we relate to unobserved vectors $Z_1, \ldots, Z_n \in \mathcal{Z} \subset \mathbb{R}^d$, where $d \ll n$, by

$$\mathbf{A}^{ij} = f(Z_i, Z_j) + \mathbf{E}^{ij}, \quad 1 \leq i \leq j \leq n, \tag{1}$$

for some real-valued symmetric function f which will be called a kernel, and a matrix of unobserved perturbations, $\mathbf{E} \in \mathbb{R}^{n \times n}$. As illustrative examples, $\mathbf{A}$ could be the observed adjacency matrix of a graph with n vertices, a similarity matrix associated with n items, or some matrix-valued estimator of covariance or correlation between n variates.

Broadly stated, our goal is to recover $Z_1, \ldots, Z_n$ given $\mathbf{A}$, up to identifiability constraints. We seek a method to perform this recovery which can be implemented *in practice* without any knowledge of f, nor of any explicit statistical model for $\mathbf{A}$, and to which we can plug in various matrix-types, be it adjacency, similarity, covariance, or other.

At face value, this may seem rather a lot to ask. Without assumptions, neither f, the Z_i's, nor even d are identifiable [24]. However, under quite general assumptions, we will show that a practical solution is provided by the key combination of two procedures: i) matrix factorisation of $\mathbf{A}$, such as spectral embedding, followed by ii) nonlinear dimension reduction, such as Isomap [57].

This combination is effective because of the following facts, only the first of which is already known. The matrix factorisation step approximates high-dimensional images of the Z_i's, living near a d-dimensional manifold [45]. Under regularity and non-degeneracy assumptions on the kernel, which

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

for our analysis is taken to be positive definite, geodesic distance in this manifold equals Euclidean geodesic distance on $\mathcal{Z}$, up to simple transformations of the coordinates, for example scaling. Thus a dimension reduction technique which approximates in-manifold geodesic distances can be expected to recover the Z_i's, up to a simple transformation. If those assumptions fail, as they must with real data, we may still find that those approximate geodesic distances are useful because they reflect a geometry implicit in the kernel.

There are many works which address recovery of the Z_i's when one *does* consider a particular matrix-type, an explicit statistical model, and/or specific form for f. When $\mathbf{A}$ is an adjacency matrix and $f(Z_i, Z_j)$ is the probability of an edge between nodes i and j, the construct (1) is a latent position model [22] and various scenarios have been studied: $\mathcal{Z}$ is a sphere and $f(x, y)$ is a function of $\langle x, y\rangle$ [10]; estimation by graph distance when $f(x, y)$ is a function of $\|x - y\|$ [11]; spectral estimation when $f(x, y)$ is a possibly indefinite inner-product [53, 55, 12, 46]; inference under a logistic model via MCMC [22] or variational approximation [48]. The case where $\mathcal{Z}$ is a discrete set of points corresponds to the very widely studied stochastic block model [23], see [5, 33, 44, 53, 8, 14] and references therein; by contrast the methods we propose are directed at the case where $\mathcal{Z}$ is continuous. The opposite problem of estimating f, with Z_i unknown but assumed uniform on $[0, 1]$, is known as graphon estimation [19].

Concerning the case when $\mathbf{A}$ is a covariance matrix, Latent Variable Gaussian Process models [31, 30, 59] use $f(Z_i, Z_j)$ to define the population covariance between the ith and jth variates under a hierarchical model. When $f(x, y) = \langle x, y\rangle$, this reduces to a latent variable optimisation counterpart of Probabilistic PCA [58] and the maximum likelihood estimate of the Z_i's is obtained from the eigendecomposition of the empirical covariance matrix, which in our setting could be $\mathbf{A}$. When f is nonlinear, but fixed e.g. to a Radial Basis Function kernel, gradient and variational techniques are available [31, 30, 59]. See the same references for onward connections to kernel PCA [50].

Our contribution and precursors. The general practice of using dimensionality reduction and geodesic distance to extract latent features from data is common in data science and machine-learning. Examples arise in genomics [38], neuroscience [63], speech analysis [28, 21] and cyber-security [9]. What sets our work apart from these contributions is that we establish a new rigorous basis for this practice.

It has been understood for several years that the high-dimensional embedding obtained by matrix factorisation can be related via a feature map to the Z_i of model (1) [26, 15, 34, 56, 62, 32]. However, it was only recently observed that the embedding must therefore concentrate about a low dimensional set, in the Hausdorff sense [45]. Our key mathematical contribution is to describe the topology and geometry of this set, proving it is a topological manifold and establishing how in-manifold geodesic distance is related to geodesic distance in $\mathcal{Z}$. Riemannian geometry underlying kernels was sketched in [7] but without consideration of geodesic distances or rigorous proofs, and not in the context of latent position estimation. The work [13, 61] was an inspiration to us, suggesting Isomap as a tool for analysing spectral embeddings, under a latent structure model in which $\mathcal{Z}$ is a one-dimensional curve in $\mathbb{R}^d$ and f is the inner product. A key feature of our problem setup is that f is unknown, in which case the manifold is not available in closed form and typically lives in an infinite-dimensional space. To our knowledge, we are the first to show why spectral embedding followed by Isomap might recover the true Z_i's, or a useful transformation thereof, in general.

We complement our mathematical results with experimental evidence, obtained from both simulated and real data, suggesting that alternative combinations of matrix-factorisation and dimension-reduction techniques work too. For the former, we consider the popular node2vec [20] algorithm, a likelihood-based approach for graphs said to perform matrix factorisation implicitly [43, 64]. For the latter, we consider the popular t-SNE [37] and UMAP [39] algorithms. As predicted by the theory, a direct low-dimensional matrix factorisation, whether using spectral embedding or node2vec, is less successful.

2 Proposed methods and their rationale

2.1 Spectral embedding, as estimating $\phi(Z_i)$

In our mathematical setup (precise details come later) the kernel f will be nonnegative definite and ϕ will be the associated Mercer feature map. It is well known that each point $\phi(Z_i)$ then lives in an

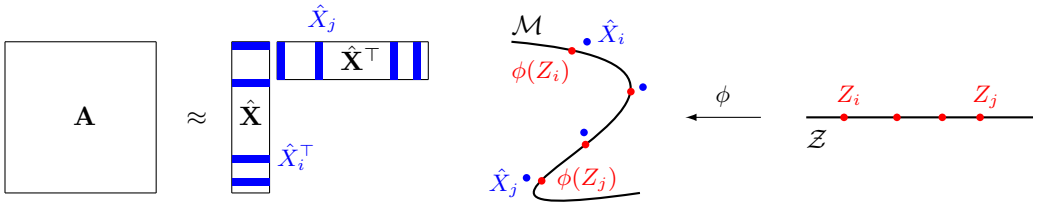

Figure 1: Illustration of theory in the case $d = 1$. Our analysis reveals how geodesic distance, along $\mathcal{M}$, between $\phi(Z_i)$ and $\phi(Z_j)$, is related to geodesic distance, along $\mathcal{Z}$, between Z_i and Z_j.

infinite-dimensional Hilbert space, which will be denoted ℓ_2, and the inner-product $\langle \phi(Z_i), \phi(Z_j) \rangle_2$ in this space equals $f(Z_i, Z_j)$.

The spectral embedding procedure. For $p \leq n$, we define the p-dimensional spectral embedding of $\mathbf{A}$ to be $\hat{\mathbf{X}} = [\hat{X}_1, \ldots, \hat{X}_n]^\top = \hat{\mathbf{U}}|\hat{\mathbf{S}}|^{1/2} \in \mathbb{R}^{n \times p}$, where $|\hat{\mathbf{S}}| \in \mathbb{R}^{p \times p}$ is a diagonal matrix containing the absolute values of the p largest eigenvalues of $\mathbf{A}$, by magnitude, and $\hat{\mathbf{U}} \in \mathbb{R}^{n \times p}$ is a matrix containing corresponding orthonormal eigenvectors, in the same order. The R packages irlba and RSpectra provide fast solutions which can exploit sparse inputs.

One should think of $\hat{X}_i$ as approximating the vector of first p components of $\phi(Z_i)$, denoted $\phi_p(Z_i)$, up to orthogonal transformation, and this can be formalised to a greater or lesser extent depending on what assumptions are made. There are several situations, e.g. f any polynomial [45], the cosine kernel used in Section 4.1, the degree-corrected [29] or mixed-membership [6] stochastic block model, in which only the first p_0 (say) components of $\phi(\cdot)$ are nonzero, where typically $p_0 \geq d$. If, after n reaches p_0, we embed into $p = p_0$ dimensions, then with $\| \cdot \|$ the Euclidean norm, we have [18]:

$$\max_{i=1,\ldots,n} \|\mathbf{Q}\hat{X}_i - \phi_p(Z_i)\| = O_{\mathbb{P}}\left(\frac{(\log n)^c}{n^{1/2}}\right), \tag{2}$$

for a universal constant $c \geq 1$, orthogonal matrix $\mathbf{Q}$, under regularity assumptions on the Z_i's, f and $\mathbf{E}$ (that Z_i are i.i.d., $f(Z_i, Z_j)$ has finite expectation, and the perturbations $\mathbf{E}^{ij}$ are independent and centered with exponential tails). This encompasses the case where $\mathbf{A}$ is binary, for example a graph adjacency matrix [36, 46]. Similar results are available in the cases where $\mathbf{A}$ is a Laplacian [46, 40], covariance matrix [17], or the matrix implicitly factorised by node2vec [64]. The methods of this paper are based, in practice, on the distances $\|\hat{X}_i - \hat{X}_j\|$, which are invariant to orthogonal transformations and so for purposes of validating $\|\hat{X}_i - \hat{X}_j\| \approx \|\phi_p(Z_i) - \phi_p(Z_j)\|$ the presence of $\mathbf{Q}$ in (2) is immaterial.

For $\hat{X}_i$ to converge to $\phi(Z_i)$ more generally, we must let its dimension p grow with n and, at least given the present state of literature, accept weaker consistency results, for example, convergence in Wasserstein distance between $\mathbf{Q}\hat{X}_1, \ldots \mathbf{Q}\hat{X}_n$ and $\phi_p(Z_1), \ldots, \phi_p(Z_n)$ [32]. Uniform consistency results, in the style of (2), are also available for indefinite [46], bipartite, and directed graphs [25]. These are left for future work because of the complications of $\mathbf{Q}$ no longer being orthogonal.

Rank selection. In real data, where n is typically fixed, there is no 'best' way of selecting p, as discussed for example in [42]. The method of [65], based on profile-likelihood, provides a popular, practical choice, taking as input the spectrum of $\mathbf{A}$, and is implemented in the R package igraph.

2.2 Isomap, as estimating Z_i

In Theorem 1 below we establish that under realistic and quite general assumptions on f, ϕ is a bi-Lipschitz homeomorphism, meaning $\mathcal{M} := \phi(\mathcal{Z})$ and $\mathcal{Z}$ are topologically equivalent, $\mathcal{M}$ has Hausdorff dimension *exactly* equal to that of $\mathcal{Z}$ (as opposed to an upper bound [45]) and $\mathcal{M}$ can be

called a topological manifold in the conventional mathematical sense. Our analysis makes precise how geodesic distance in $\mathcal{M}$ between any $\phi(Z_i)$ and $\phi(Z_j)$, written $\mathbf{D}_{\mathcal{M}}^{ij}$, is related to Euclidean geodesic distance in $\mathcal{Z}$ between Z_i and Z_j, written $\mathbf{D}_{\mathcal{Z}}^{ij}$ (illustrated in Figure 1). In Section 3.3, for various families of kernel, we show that $\mathbf{D}_{\mathcal{M}}^{ij}$ is equal to $\mathbf{D}_{\mathcal{Z}}^{ij}$ up to some simple transformations, or even exactly equal in special cases. One can estimate $\mathbf{D}_{\mathcal{M}}$ from $\hat{X}_1, \ldots, \hat{X}_n$ (Steps 1-2 below), and through this attempt to reconstruct $Z_1, \ldots, Z_n$, by nonlinear dimension reduction (Step 3).

Algorithm 1 Isomap procedure

input p-dimensional points $\hat{X}_1, \ldots, \hat{X}_n$

1: Compute the neighbourhood graph of radius ϵ: a weighted graph connecting i and j, with weight $\|\hat{X}_i - \hat{X}_j\|$, if $\|\hat{X}_i - \hat{X}_j\| \leq \epsilon$

2: Compute the matrix of shortest paths on the neighbourhood graph, $\hat{\mathbf{D}}_{\mathcal{M}} \in \mathbb{R}^{n \times n}$

3: Apply classical multidimensional scaling (CMDS) to $\hat{\mathbf{D}}_{\mathcal{M}}$ into $\mathbb{R}^d$

return d-dimensional points $\hat{Z}_1, \ldots, \hat{Z}_n$

Dimension and radius selection. For visualisation in our examples we will pick $d = 1$ or $d = 2$. For other applications, d can be estimated as the approximate rank of $\hat{\mathbf{D}}_{\mathcal{M}}$ after double-centering [60], e.g. via [65], again. In practice, we suggest picking ϵ just large enough for the neighbourhood graph to be connected or, if the data have outliers, a fixed quantile (such as 5%) of $\hat{\mathbf{D}}_{\mathcal{M}}$, removing nodes outside the largest connected component. The same recommendations apply if the k-nearest neighbour graph is used instead of the ϵ-neighbourhood graph.

3 Theory

3.1 Setup and assumptions

The usual inner-product and Euclidean norm on $\mathbb{R}^d$ are denoted $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$. ℓ_2 is the set of $x = [x^{(1)}\, x^{(2)}\, \cdots]^\top \in \mathbb{R}^{\mathbb{N}}$ such that $\|x\|_2 := (\sum_{k=1}^{\infty} |x^{(k)}|^2)^{1/2} < \infty$. For $x, y \in \ell_2$, $\langle x, y \rangle_2 := \sum_{k=1}^{\infty} x^{(k)} y^{(k)}$. With $d \geq 1$, let $\mathcal{Z}$ be a compact subset of $\mathbb{R}^d$, and let $\widetilde{\mathcal{Z}}$ be a closed ball in $\mathbb{R}^d$, centered at the origin, such that $\mathcal{Z} \subset \widetilde{\mathcal{Z}}$. Let $f : \widetilde{\mathcal{Z}} \times \widetilde{\mathcal{Z}} \to \mathbb{R}$ be a symmetric, continuous, nonnegative-definite function.

By Mercer's Theorem, e.g., [51, Thm 4.49], there exist nonnegative real numbers $(\lambda_k)_{k \geq 1}$ and functions $(u_k)_{k \geq 1}$, with each $u_k : \widetilde{\mathcal{Z}} \to \mathbb{R}$, which are orthonormal with respect to the inner-product $(u_j, u_k) \mapsto \int_{\tilde{\mathcal{Z}}} u_j(x) u_k(x) \mathrm{d}x$ and such that

$$f(x, y) = \sum_{k=1}^{\infty} \lambda_k u_k(x) u_k(y), \quad x, y \in \widetilde{\mathcal{Z}}, \tag{3}$$

where the convergence is absolute and uniform. For $x \in \mathcal{Z}$ let $\phi(x) := [\lambda_1^{1/2} u_1(x)\ \lambda_2^{1/2} u_2(x)\ \cdots]^\top \in \mathbb{R}^{\mathbb{N}}$. The image of $\mathcal{Z}$ by ϕ is denoted $\mathcal{M}$. Observe from (3) that for any $x, y \in \mathcal{Z}$, $f(x, y) = \langle \phi(x), \phi(y) \rangle_2$, $\|\phi(x)\|_2^2 = f(x, x)$, and the latter is finite for any $x \in \mathcal{Z}$ since f is continuous and $\mathcal{Z}$ is compact, hence $\mathcal{M} \subset \ell_2$.

The following definitions are standard in metric geometry [16]. For any $a, b \in \ell_2$, a *path* in $\mathcal{M}$ with end-points a, b is a continuous function $\gamma : [0, 1] \to \mathcal{M}$ such that $\gamma_0 = a$ and $\gamma_1 = b$. With $n \geq 1$, a non-decreasing sequence $t_0, \ldots, t_n$ such that $t_0 = 0$ and $t_n = 1$, is called a *partition.* Given a path γ and a partition $\mathcal{P} = (t_0, \ldots, t_n)$, define $\chi(\gamma, \mathcal{P}) := \sum_{k=1}^{n} \|\gamma_{t_k} - \gamma_{t_{k-1}}\|_2$. The *length* of γ (w.r.t. $\| \cdot \|_2$) is $l(\gamma) := \sup_{\mathcal{P}} \chi(\gamma, \mathcal{P})$, where the supremum is over all possible partitions. The geodesic distance in $\mathcal{M}$ between a and b is defined to be the infimum of $l(\gamma)$ over all paths γ in $\mathcal{M}$ with end-points a, b. $\mathbf{D}_{\mathcal{M}}^{ij}$ denotes geodesic distance in $\mathcal{M}$ between $\phi(Z_i)$ and $\phi(Z_j)$. Similarly a path η in $\mathcal{Z}$ with end-points x, y is a continuous function $\eta : [0, 1] \to \mathcal{Z}$ such that $\eta_0 = x, \eta_1 = y$, and with $\chi(\eta, \mathcal{P}) := \sum_{k=1}^{n} \|\eta_{t_k} - \eta_{t_{k-1}}\|$ the length of η (w.r.t. $\| \cdot \|$) is $l(\eta) := \sup_{\mathcal{P}} \chi(\eta, \mathcal{P})$. $\mathbf{D}_{\mathcal{Z}}^{ij}$ denotes geodesic distance in $\mathcal{Z}$ between Z_i and Z_j. If $\mathcal{Z}$ is convex, $\mathbf{D}_{\mathcal{Z}}^{ij} = \|Z_i - Z_j\|$.

Assumption 1. *For all $x, y \in \mathcal{Z}$ with $x \neq y$, there exists $a \in \mathcal{Z}$ such that $f(x,a) \neq f(y,a)$.*

Assumption 2. *f is C^2 on $\widetilde{\mathcal{Z}}$ and for every $z \in \mathcal{Z}$, the matrix $\mathbf{H}_z \in \mathbb{R}^{d\times d}$ with elements*

$$\mathbf{H}_z^{ij} := \left.\frac{\partial^2 f}{\partial x^{(i)} \partial y^{(j)}}\right|_{(z,z)}$$

is positive-definite.

Assumption 1 is used to show ϕ is injective. Assumption 2 has various implications, loosely speaking it ensures a non-degenerate relationship between path-length in $\mathcal{M}$ and path-length in $\mathcal{Z}$. Concerning our general setup, if f is not required to be nonnegative definite, but is still symmetric, then a representation formula like (3) is available under e.g. trace-class assumptions [45]. It is of interest to generalise our results to that scenario but technical complications are involved, so we leave it for future research.

3.2 Path-lengths in $\mathcal{M}$ and in $\mathcal{Z}$

Theorem 1. *Let assumptions 1 and 2 hold. Then ϕ is a bi-Lipschitz homeomorphism between $\mathcal{Z}$ and $\mathcal{M}$. Let a, b be any two points in $\mathcal{M}$ such that there exists a path in $\mathcal{M}$ with end-points a, b of finite length. Let γ be any such path and define $\eta : [0,1] \to \mathcal{Z}$ by $\eta_t := \phi^{-1}(\gamma_t)$. Then η is a path in $\mathcal{Z}$ with $l(\eta) < \infty$. For any $\epsilon > 0$ there exists a partition $\mathcal{P}_\epsilon$ such that for any partition $\mathcal{P} = (t_0, \ldots, t_n)$ satisfying $\mathcal{P}_\epsilon \subseteq \mathcal{P}$,*

$$\left| l(\gamma) - \sum_{k=1}^{n} \left\langle \eta_{t_k} - \eta_{t_{k-1}}, \mathbf{H}_{\eta_{t_{k-1}}}(\eta_{t_k} - \eta_{t_{k-1}}) \right\rangle^{1/2} \right| \leq \epsilon. \tag{4}$$

If γ and η are continuously differentiable, the following two equalities hold:

$$l(\gamma) = \int_0^1 \|\dot{\gamma}_t\|_2 \mathrm{d}t = \int_0^1 \langle \dot{\eta}_t, \mathbf{H}_{\eta_t} \dot{\eta}_t \rangle^{1/2} \, \mathrm{d}t. \tag{5}$$

The proof is in the appendix. In the terminology of differential geometry, the collection of inner-products $(x,y) \mapsto \langle x, \mathbf{H}_z y \rangle$, indexed by $z \in \mathcal{Z}$, constitute a *Riemannian metric* on $\mathcal{Z}$, see e.g. [41, Ch.1 and p.121] for background, and the right hand side of (5) is the length of the path η in $\mathcal{Z}$ with respect to this Riemannian metric. The theorem tells us that finding geodesic distance in $\mathcal{M}$, i.e. minimising $l(\gamma)$ with respect to γ for fixed end-points say $a = \phi(Z_i), b = \phi(Z_j)$, is equivalent to minimizing path-length in $\mathcal{Z}$ between end-points $\phi^{-1}(a) = Z_i, \phi^{-1}(b) = Z_j$ under the Riemannian metric. In section 3.3 we will show how, for various classes of kernels, this minimisation can be performed in closed form leading to explicit relationships between $\mathbf{D}_{\mathcal{M}}^{ij}$ and $\mathbf{D}_{\mathcal{Z}}^{ij}$. In Section 3.3 we focus on (5) rather than (4) only for ease of exposition, it is shown in the appendix that exactly the same conclusions can be derived directly from (4). To avoid repetition Assumption 1 will be taken to hold throughout section 3.3 without further mention.

3.3 Geodesic distance in $\mathcal{M}$ and in $\mathcal{Z}$

Translation invariant kernels. Suppose $\mathcal{Z} \subset \mathbb{R}^d$ is compact and convex, and $f(x,y) = g(x-y)$ where $g : \mathbb{R}^d \to \mathbb{R}$ is C^2. In this situation $\mathbf{H}_z$ is constant in z, and equal to the Hessian of $-g$ evaluated at the origin. The positive-definite part of Assumption 2 is therefore satisfied if and only if g has a local maximum at the origin. We now use Theorem 1 to find geodesic distance in $\mathcal{M}$ between generic points $a, b \in \mathcal{M}$ for this class of translation-invariant kernels. To do this we obtain a lower bound on $l(\gamma)$ over all paths γ in $\mathcal{M}$ with end-points a, b, then show there exists such a path whose length achieves this lower bound.

Let $\mathbf{G} = \mathbf{V}\mathbf{\Lambda}^{1/2}$ where $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$ is the eigendecomposition of the Hessian of $-g$ evaluated at the origin, so $\mathbf{H}_z = \mathbf{G}\mathbf{G}^\top$ for all z. Then from (5), for a generic path γ in $\mathcal{M}$ with end-points a, b,

$$l(\gamma) = \int_0^1 \langle \dot{\eta}_t, \mathbf{H}_{\eta_t} \dot{\eta}_t \rangle^{1/2} \, \mathrm{d}t = \int_0^1 \left\langle \dot{\eta}_t, \mathbf{G}\mathbf{G}^\top \dot{\eta}_t \right\rangle^{1/2} \mathrm{d}t = \int_0^1 \|\mathbf{G}^\top \dot{\eta}_t\| \mathrm{d}t \tag{6}$$

$$\geq \left\| \int_0^1 \mathbf{G}^\top \dot{\eta}_t \mathrm{d}t \right\| = \left\| \mathbf{G}^\top \int_0^1 \dot{\eta}_t \mathrm{d}t \right\| = \|\mathbf{G}^\top(\eta_1 - \eta_0)\| = \|\mathbf{G}^\top[\phi^{-1}(b) - \phi^{-1}(a)]\|, \tag{7}$$

where the inequality is due to the triangle inequality for the $\|\cdot\|$ norm. The right-most term in (7) is independent of γ, other than through the end-points a, b. To see there exists a path whose length equals this lower bound, take

$$\tilde{\eta}_t \coloneqq \phi^{-1}(a) + t[\phi^{-1}(b) - \phi^{-1}(a)], \quad t \in [0,1], \tag{8}$$

which is well-defined as a path in $\mathcal{Z}$, since for this class of kernels $\mathcal{Z}$ is assumed convex. With $\tilde{\gamma}_t \coloneqq \phi(\tilde{\eta}_t)$, $\tilde{\gamma}$ is clearly a path in $\mathcal{M}$. Differentiating $\tilde{\eta}_t$ w.r.t. t and substituting into (6) shows $l(\tilde{\gamma}) = \|\mathbf{G}^\top[\phi^{-1}(b) - \phi^{-1}(a)]\|$, as required. The following proposition summarises our conclusions in the case that $a = \phi(Z_i)$ and $b = \phi(Z_j)$, for any Z_i, Z_j.

Proposition 1. *If $\mathcal{Z}$ is compact and convex, and $f(x,y) = g(x-y)$ where $g : \mathbb{R}^d \to \mathbb{R}$ is C^2 and has a local maximum at the origin, then $\mathbf{D}^{ij}_{\mathcal{M}}$ is equal to Euclidean distance between Z_i and Z_j, up to their linear transformation by $\mathbf{G}^\top$. In the particular case where $g(x-y) = h(\|x-y\|^2)$ with $h'(0) < 0$, we have $\mathbf{D}^{ij}_{\mathcal{M}} \propto \mathbf{D}^{ij}_{\mathcal{Z}} = \|Z_i - Z_j\|$.*

Inner-product kernels. Suppose $\mathcal{Z} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ and $f(x,y) = g(\langle x, y\rangle)$, where $g : \mathbb{R} \to \mathbb{R}$ is C^2 and such that $g'(1) > 0$. In this case $\mathbf{H}_z = g'(1)\mathbf{I} + g''(1)zz^\top$, and by a result of [27], any kernel of the form $f(x,y) = g(\langle x, y\rangle)$ on the sphere is positive-definite iif $g(x) = \sum_{n=0}^{\infty} a_n x^n$ for some nonnegative $(a_n)_{n\geq 0}$, implying $g''(1) \geq 0$. Assumption 2 then holds. To derive the geodesic distance in $\mathcal{M}$, first write out (5) for a generic path γ in $\mathcal{M}$ with end-points a, b:

$$l(\gamma) = \int_0^1 \langle \dot{\eta}_t, \mathbf{H}_{\eta_t}\dot{\eta}_t\rangle^{1/2}\,\mathrm{d}t = \int_0^1 \left(g'(1)\|\dot{\eta}_t\|^2 + g''(1)\,|\langle \eta_t, \dot{\eta}_t\rangle|^2\right)^{1/2}\mathrm{d}t. \tag{9}$$

Since $\mathcal{Z}$ is a radius-1 sphere centered at the origin we must have $\|\eta_t\| = 1$ for all t, so $0 = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\eta_t\|^2 = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\langle \eta_t, \eta_t\rangle = \langle \eta_t, \dot{\eta}_t\rangle$. Therefore the r.h.s. of (9) is in fact equal to $g'(1)^{1/2}\int_0^1 \|\dot{\eta}_t\|\mathrm{d}t$. This is minimised over all possible paths η in $\mathcal{Z}$ with end-points $\phi^{-1}(a), \phi^{-1}(b)$ when η is a shortest (with respect to Euclidean distance) circular arc in $\mathcal{Z}$, in which case $\int_0^1 \|\dot{\eta}_t\|\mathrm{d}t = \arccos\left\langle \phi^{-1}(a), \phi^{-1}(b)\right\rangle$. Thus we have:

Proposition 2. *If $\mathcal{Z} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ and $f(x,y) = g(\langle x, y\rangle)$ where $g : \mathbb{R} \to \mathbb{R}$ is C^2 and such that $g'(1) > 0$, then $\mathbf{D}^{ij}_{\mathcal{M}} = g'(1)^{1/2}\mathbf{D}^{ij}_{\mathcal{Z}} = g'(1)^{1/2}\arccos\langle Z_i, Z_j\rangle$.*

Additive kernels. Suppose $\mathcal{Z}$ is the Cartesian product of intervals $\mathcal{Z}_i = [c_i^-, c_i^+]$, $i = 1, \ldots, d$ and $f(x,y) = \sum_{i=1}^d \alpha_i f_i(x^{(i)}, y^{(i)})$, $x = [x^{(1)} \cdots x^{(d)}]^\top$, where each $\alpha_i > 0$ and each f_i is positive-definite and C^2. For this class of kernels $\mathbf{H}_z$ is diagonal with $\mathbf{H}^{ii}_z = \alpha_i \left.\frac{\partial^2 f_i}{\partial x^{(i)}\partial y^{(i)}}\right|_{(z^{(i)}, z^{(i)})}$ where $z^{(i)}$ is the ith element of the vector z. The positive-definite part of Assumption 2 thus holds if these diagonal elements are strictly positive for all $z \in \mathcal{Z}$.

Let us introduce the vector of monotonically increasing transformations:

$$\psi(z) \coloneqq [\psi_1(z^{(1)}) \cdots \psi_d(z^{(d)})]^\top, \qquad \psi_i(z^{(i)}) \coloneqq \alpha_i^{1/2}\int_{c_i^-}^{z^{(i)}} \left.\frac{\partial^2 f_i}{\partial x^{(i)}\partial y^{(i)}}\right|^{1/2}_{(\xi,\xi)}\mathrm{d}\xi, \tag{10}$$

and write the vector of corresponding inverse transformations ψ^{-1}. For a generic path γ in $\mathcal{M}$ with end-points a, b and as usual $\eta_t \coloneqq \phi^{-1}(\gamma_t)$, define $\zeta_t \coloneqq \psi(\eta_t)$. Then:

$$l(\gamma) = \int_0^1 \langle \dot{\eta}_t, \mathbf{H}_{\eta_t}\dot{\eta}_t\rangle^{1/2}\,\mathrm{d}t = \int_0^1 \|\dot{\zeta}_t\|\mathrm{d}t \tag{11}$$

$$\geq \left\|\int_0^1 \dot{\zeta}_t\,\mathrm{d}t\right\| = \|\zeta_1 - \zeta_0\| = \|\psi\circ\phi^{-1}(b) - \psi\circ\phi^{-1}(a)\|, \tag{12}$$

where the second and last equalities hold due to the definition of ζ_t, and $\circ$ denotes elementwise composition. The quantity $\|\psi\circ\phi^{-1}(b) - \psi\circ\phi^{-1}(a)\|$ is thus a lower bound on path-length $l(\gamma)$ for any path γ in $\mathcal{M}$ with end-points a, b. To see that there exists a path whose length achieves this lower bound, and hence that it is the geodesic distance in $\mathcal{M}$ between a, b, define $\tilde{\gamma}_t \coloneqq \phi\circ\psi^{-1}(\tilde{\zeta}_t)$ where $\tilde{\zeta}_t \coloneqq \psi\circ\phi^{-1}(a) + t[\psi\circ\phi^{-1}(b) - \psi\circ\phi^{-1}(a)]$. Differentiating $\tilde{\zeta}_t$ w.r.t. t and substituting into (11) yields $l(\tilde{\gamma}) = \|\psi\circ\phi^{-1}(b) - \psi\circ\phi^{-1}(a)\|$, as required. Our conclusions in the case that $a = \phi(Z_i)$ and $b = \phi(Z_j)$ are summarised by:

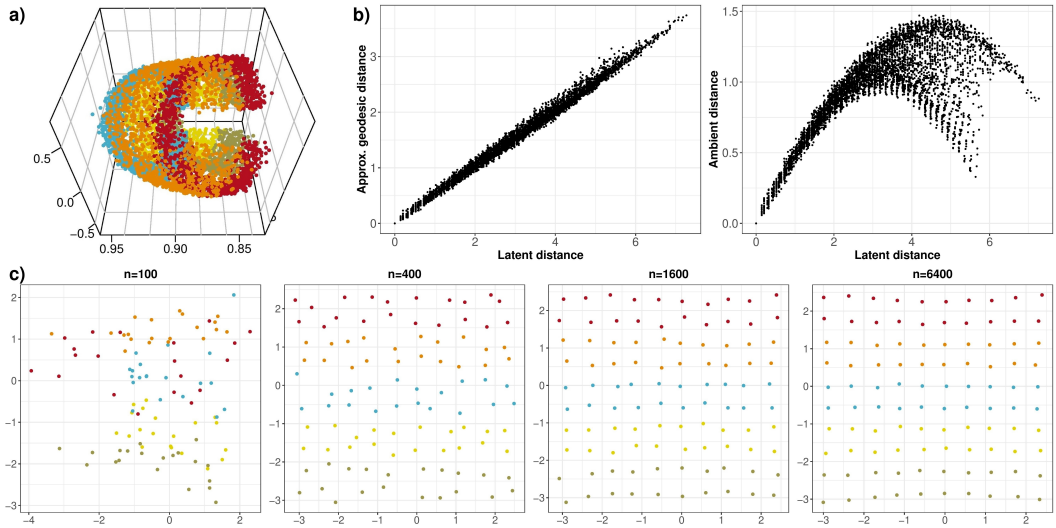

Figure 2: Simulated data example. a) Spectral embedding (first 3 dimensions), b) comparison of latent, approximate geodesic, and ambient distance and c) latent position recovery in the dense regime by spectral embedding followed by Isomap for increasing n. To aid visualisation, all plots in c) display a subset of 100 estimated positions corresponding to true positions on a sub-grid which is common across n. Estimated positions are coloured according to their true y-coordinate.

Proposition 3. *If $\mathcal{Z}$ is the Cartesian product of d intervals $\mathcal{Z}_i \subset \mathbb{R}$, $i = 1, \ldots, d$, and $f(x, y) = \sum_{i=1}^{d} \alpha_i f_i(x^{(i)}, y^{(i)})$ where for each i, $\alpha_i > 0$, f_i is positive definite and C^2, and $\frac{\partial^2 f_i}{\partial x^{(i)} \partial y^{(i)}}\Big|_{(z,z)} > 0$ for all $z \in \mathcal{Z}_i$, then $\mathbf{D}^{ij}_{\mathcal{M}}$ is equal to the Euclidean distance between Z_i and Z_j up to their coordinatewise-monotone transformation by ψ. If $\alpha_i \frac{\partial^2 f_i}{\partial x^{(i)} \partial y^{(i)}}\Big|_{(z,z)}$ is a positive constant over all $z \in \mathcal{Z}_i$, and over all $i = 1, \ldots, d$, then $\mathbf{D}^{ij}_{\mathcal{M}} \propto \mathbf{D}^{ij}_{\mathcal{Z}} = \|Z_i - Z_j\|$.*

4 Experiments

4.1 Simulated data: latent position network model

This section shows the theory at work in a pedagogical example, where $\mathbf{A}$ is an adjacency matrix. Consider an undirected random graph following a latent position model with kernel $f(x, y) = \rho_n\{\cos(x^{(1)} - y^{(1)}) + \cos(x^{(2)} - y^{(2)}) + 2\}/4$, operating on $\mathbb{R}^2 \times \mathbb{R}^2$. Here, the sequence ρ_n is a sparsity factor which will either be constant, $\rho_n = 1$, or shrinking to zero sufficiently slowly, reflecting dense (degree grows linearly) and sparse (degree grows sublinearly) regimes respectively. The kernel is clearly translation-invariant, satisfies Assumptions 1 and 2 and has finite rank, $p = 5$. The true latent positions $Z_i \in \mathbb{R}^2$ are deterministic and equally spaced points on a grid over the region $\mathcal{Z} = [-\pi + 0.25, \pi - 0.25] \times [-\pi + 0.25, \pi - 0.25]$, this range chosen to give valid probabilities and an interesting bottleneck in the 2-dimensional manifold $\mathcal{M}$. From Proposition 1, the geodesic distance between $\phi(Z_i)$ and $\phi(Z_j)$ on $\mathcal{M}$ is equal to the Euclidean distance between Z_i and Z_j, up to scaling, specifically $\mathbf{D}^{ij}_{\mathcal{M}} = \rho_n \mathbf{D}^{ij}_{\mathcal{Z}}/2$.

Focusing first on the dense regime, for each $n = 100, 400, 1600, 6400$, we simulate a graph, and its spectral embedding $\hat{X}_1, \ldots, \hat{X}_n$ into $p = 5$ dimensions. The first three dimensions of this embedding are shown in Figure 2a), and we see that the points gather about a two-dimensional, curved, manifold. To approximate geodesic distance, we compute a neighbourhood graph of radius ϵ from $\hat{X}_1, \ldots, \hat{X}_n$, choosing ϵ as small as possible subject to maintaining a connected graph. Figure 2b) shows that approximated geodesic distances roughly agree with the true Euclidean distances in latent space $\mathcal{Z}$ (up to the predicted scaling of 1/2), whereas there is significant distortion between ambient and

latent distance ($\|\hat{X}_i - \hat{X}_j\|$ versus $\|Z_i - Z_j\|$). Finally, we recover the estimated latent positions in $\mathbb{R}^2$ using Isomap, which we align with the original by Procrustes (orthogonal transformation with scaling), in Figure 2c). As the theory predicts, the recovery error vanishes as n increases.

In the appendix, we perform the same experiment in a sparse regime, $n\rho_n = \omega\{\log^4 n\}$, chosen to ensure the spectral embedding is still consistent, and the recovery error still shrinks, but more slowly. We also implement other related approaches: UMAP, t-SNE applied to spectral embedding, node2vec directly into two dimensions and node2vec in five dimensions followed by Isomap (Figure 6). We use default configurations for all other hyperparameters. Together, the results support the central recommendation of this paper: to use matrix factorisation (e.g. spectral embedding, node2vec), followed by nonlinear dimension reduction (e.g. Isomap, UMAP, t-SNE).

4.2 Real data: structural change in the global flight network

Air travel was, for several reasons, a prominent feature of the COVID-19 pandemic, and we will explore structural changes in the global flight network towards its start (the 11th March 2020 according to the WHO).

The dataset is a publically available, clean version [52] (with license details therein) of the crowd-sourced OpenSky dataset. Six undirected graphs are extracted, for the months November 2019 to April 2020, connecting any two airports over the world for which at least a single flight was recorded ($n \approx 10,000$ airports in each). For each graph the adjacency matrix $\mathbf{A}$ is spectrally embedded into $p = 10$ dimensions, degree-corrected by spherical projection [35, 33, 49], after which we apply Isomap with ϵ chosen as the 5% distance quantile. The hypothesis is that the resulting $\hat{Z}_i$ could estimate the geographical location of airport i. The dimension p was chosen as a loose upper bound on the dimension estimate returned by the method of [65] on any of the individual graphs, as to facilitate comparison we prefer to use the same dimension throughout (e.g. to avoid artificial differences in variance) and so follow the recommendation of [12] to err on the large side (experiments with different choices of p are in the appendix). In the degree-correction step we aim to remove network effects related to "airport popularity" (e.g. population, economic and airport size), after which geographic distance might be expected to be the principal factor deciding whether there should be a flight between two airports. The experiments take a few hours on a standard desktop (and the same is a loose upper bound on compute time for the other experiments in the paper).

The embedding for January is shown in Figures 3)a-b), before and after Isomap. Both recover real geographic features, notably the relative positioning of London, Paris, Madrid, Lisbon, Rome. However, the embedding of the US is warped, and only after using Isomap do we see New York and Los Angeles correctly positioned at opposite extremes of the country. As further confirmation, Figure 8 (appendix) shows the results restricted to North America, with the points coloured by latitude and longitude. For this continent, the accuracy of longitude recovery is particularly striking. In line with the recommendation to overestimate p [12], one obtains similar figure using $p = 20$ before applying isomap (Figure 11, appendix), whereas the corresponding figure with $p = 2$ is too poor to show. In the appendix, we also show the same figure applying t-SNE and UMAP to the spectral embedding, or using node2vec for the embedding step, either directly into two dimensions, or into ten followed by Isomap (Figure 9). Over these experiments, the conclusion is consistent with the theory: a plausible recovery of geography is only ever achieved by applying matrix factorisation, followed by dimension reduction.

Figure 3c) shows the next three months (six in the appendix), aligned using two-dimensional Procrustes, showing an important structural change in April. A statistical analysis (appendix) of inter-continental geodesic distances suggests the change reflects severe flight restrictions in South America and Africa (latter less visible in the figures).

4.3 Real data: correlations between daily temperatures around the globe

In this example $\mathbf{A}$ is a correlation matrix and we demonstrate a simple model-checking diagnostic informed by our theory. The raw data consist of average temperatures over each day, for several thousand days, recorded in cities around the globe. These data are open source, originate from Berkeley Earth [1] and the particular data set analyzed is from [3]. We used open source latitude and longitude data from [4]. See those references for license details. Removing cities and dates for which

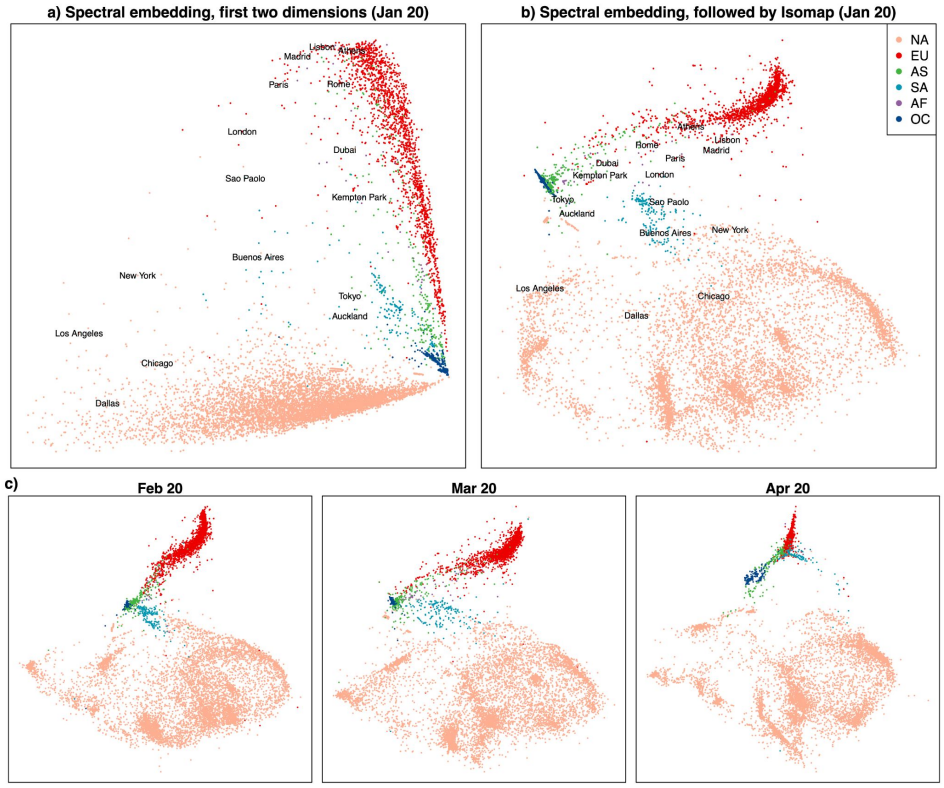

Figure 3: Visualisation of the global flight network. a) Spectral embedding, b) spectral embedding followed by Isomap, both for January 2020, c) spectral embedding followed by Isomap for the three following months. The colours indicate continents (NA = North America, EU = Europe, AS = Asia, AF = Africa, OC = Oceania) and a spread of cities with high-traffic airports are labelled.

data are missing yields a temperature time-series of 1450 common days for each of the $n = 2211$ cities. $\mathbf{A}$ is the matrix of Pearson correlation coefficients between the n time-series.

Figure 4b) shows the spectral embedding of $\mathbf{A}$, with $p = 2$ and points coloured by the latitude of the corresponding cities. Two visual features are striking: the concentration of points around a curved manifold, and a correspondence between latitude and location on the manifold. Figure 4c) shows latitude against estimated latent positions from Isomap using the k-nearest neighbour graph with $k = 200$, which is roughly 10% of n, and with $d = 1$. A clear monotone relationship appears. Our theoretical results can explain this phenomenon. Notice that when $d = 1$, the additive structure of the kernel in Proposition 3 disappears. Hence Proposition 3 shows that in the case $d = 1$, for *any* kernel (meeting the basic requirements of section 3.1, including Assumptions 1 and 2 of course) $\mathbf{D}^{ij}_{\mathcal{M}}$ is equal to Euclidean distance between Z_i and Z_j up to their monotone transformation by ψ defined in (10). In turn, this implies that if $\mathbf{A}$ did actually follow the model (1) for some f and some "true" latent positions $Z_1, \ldots, Z_n$ to which we had access, we should observe a monotone relationship between those true latent positions and the estimated positions $\hat{Z}_1, \ldots, \hat{Z}_n$ from Isomap.

The empirical monotonicity in Figure 4c) thus can be interpreted as indicating latitudes of the cities are plausible as "true" latent positions underlying the correlation matrix $\mathbf{A}$, without us having to specify anything in practice about f. In further analysis (appendix) no such empirical monotone relationship is found between longitude and estimated latent position; this indicates longitude does *not* influence the correlations captured in $\mathbf{A}$. A possible explanation is the daily averaging of temperatures in the underlying data: correlations between these average temperatures may be insensitive to longitude due to the rotation of the earth.

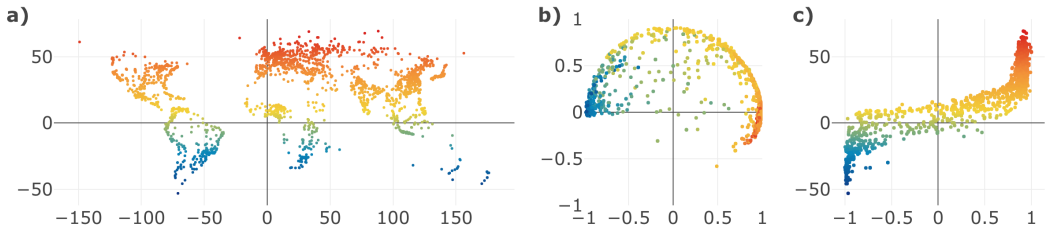

Figure 4: Temperature correlation example. a) Locations in degrees latitude and longitude of cities where temperatures were recorded, b) spectral embedding with $p = 2$, c) latitude of each city (vertical axis) against estimated latent position (horizontal axis) when $d = 1$. In all plots, points are coloured by latitude of the corresponding cities.

5 Conclusion

Our research shows how matrix-factorisation and dimension-reduction can work together to reveal the true positions of objects based only on pairwise similarities such as connections or correlations. For the sake of exposition and reproducibility, we have used public datasets which can be interpreted without specialist domain knowledge, but the methods are potentially relevant to any scientific application involving large similarity matrices. Thinking about societal impact, our results highlight the depth of information in principle available about individuals, given network metadata, and we hope to raise awareness of these potential issues of privacy.

Concerning the limitations of the methods we have discussed, the bound in (2) indicates that n should be "large" in order for $\mathbf{Q}\hat{X}_i$ to approximate $\phi_p(z_i)$ well. n also has an impact on the performance of Isomap: heuristically one needs a high density and a large number of points on or near the manifold to get a good estimate of the geodesic distance. Thus the methods we propose are likely to perform poorly when n is small, corresponding e.g. to a graph with a small number of vertices.

On the other hand, in applications involving large networks or matrices (e.g. cyber-security, recommender systems), the data encountered are often sparse. This is good news for computing the factorisation, in terms of speed, but bad news for statistical accuracy. In particular, for a graph, (2) can only be expected to hold under logarithmically growing degree, the information-theoretic limit below which no algorithm can obtain asymptotically exact embeddings [5]. What this means in practice (as illustrated in our numerical results for the simulated data example in the appendix) is that the manifold may be very hard to distinguish, even for a large graph. Missing data may have a similarly negative impact on discerning manifold structure. There could be substantial estimation gains in better integrating the factorisation and manifold estimation steps to overcome these difficulties.

Another limitation is that our theory restricts attention to the case of positive-definite kernels. When $\mathbf{A}$ is say a correlation or covariance matrix the positive-definite assumption on the kernel is of course natural, but e.g. when $\mathbf{A}$ is an adjacency matrix, it is a less natural assumption. The implications of removing the positive-definite assumption are the subject of ongoing research.

Funding Transparency Statement

Nick Whiteley and Patrick Rubin-Delanchy's research was supported by Turing Fellowships from the Alan Turing Institute. Annie Gray's research was supported by a studentship from Compass, the EPSRC Centre for Doctoral Training in Computational Statistics and Data Science.

References

[1] Berkeley Earth. `http://berkeleyearth.org`. Accessed: 2021-05-27.

[2] Countries slammed their borders shut to stop coronavirus. but is it doing any good? `https://www.npr.org/sections/goatsandsoda/2020/05/15/855669867/countries-slammed-their-borders-shut-to-stop-coronavirus-but-is-it-doing-any-goo`. Accessed: 2021-05-20.

[3] Climate change: Earth surface temperature data. https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data. Accessed: 2021-05-27.

[4] Simplemaps free entire world database. https://simplemaps.com/resources/free-country-cities. Accessed: 2021-05-27.

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

[6] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014, 2008.

[7] Shun-ichi Amari and Si Wu. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12(6):783–789, 1999.

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

[9] Thomas Anglade, Christophe Denis, and Thierry Berthier. A novel embedding-based framework improving the User and Entity Behav- ior Analysis. working paper or preprint, October 2019. URL https://hal.sorbonne-universite.fr/hal-02316303.

[10] Ernesto Araya Valdivia and De Castro Yohann. Latent distance estimation for random geometric graphs. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/c4414e538a5475ec0244673b7f2f7dbb-Paper.pdf.

[11] Ery Arias-Castro, Antoine Channarond, Bruno Pelletier, and Nicolas Verzelen. On the estimation of latent distances using graph distances. *Electronic Journal of Statistics*, 15(1):722 – 747, 2021. doi: 10.1214/21-EJS1801. URL https://doi.org/10.1214/21-EJS1801.

[12] Avanti Athreya, Donniell E Fishkind, Minh Tang, Carey E Priebe, Youngser Park, Joshua T Vogelstein, Keith Levin, Vince Lyzinski, and Yichen Qin. Statistical inference on random dot product graphs: a survey. *The Journal of Machine Learning Research*, 18(1):8393–8484, 2017.

[13] Avanti Athreya, Minh Tang, Youngser Park, and Carey E Priebe. On estimation and inference in latent structure random graphs. *Statistical Science*, 36(1):68–88, 2021.

[14] Peter Bickel, David Choi, Xiangyu Chang, and Hai Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Annals of Statistics*, 41 (4):1922–1943, 2013.

[15] Béla Bollobás, Svante Janson, and Oliver Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, 2007.

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

[17] 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.

[18] Ian Gallagher, Andrew Jones, Anna Bertiger, Carey Priebe, and Patrick Rubin-Delanchy. Spectral embedding of weighted graphs. *arXiv preprint arXiv:1910.05534*, 2019.

[19] Chao Gao, Yu Lu, and Harrison H Zhou. Rate-optimal graphon estimation. *Annals of Statistics*, 43(6):2624–2652, 2015.

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

[21] Souleiman Hasan and Edward Curry. Word re-embedding via manifold dimensionality retention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 321–326, 2017.

[22] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

[23] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[24] Svante Janson and Sofia Olhede. Can smooth graphons in several dimensions be represented by smooth graphons on $[0, 1]$? *arXiv preprint arXiv:2101.07587*, 2021.

[25] Andrew Jones and Patrick Rubin-Delanchy. The multilayer random dot product graph. *arXiv preprint arXiv:2007.10455*, 2020.

[26] Olav Kallenberg. On the representation theorem for exchangeable arrays. *Journal of Multivariate Analysis*, 30(1):137–154, 1989.

[27] Purushottam Kar and Harish Karnick. Random feature maps for dot product kernels. In *Artificial intelligence and statistics*, pages 583–591. PMLR, 2012.

[28] Zahi N. Karam and William M. Campbell. Graph-embedding for speaker recognition. In *Proc. Interspeech 2010*, pages 2742–2745, 2010. doi: 10.21437/Interspeech.2010-726.

[29] Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

[30] 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.

[31] Neil D Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *Nips*, volume 2, page 5. Citeseer, 2003.

[32] Jing Lei. Network representation using graph root distributions. *The Annals of Statistics*, 49(2): 745–768, 2021.

[33] Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *Ann. Statist.*, 43(1):215–237, 2015. ISSN 0090-5364. doi: 10.1214/14-AOS1274. URL `https://doi.org/10.1214/14-AOS1274`.

[34] László Lovász. *Large networks and graph limits. American Mathematical Society Colloquium Publications*, volume 60. Amer. Math. Soc. Providence, RI, 2012.

[35] Vince Lyzinski, Daniel L. Sussman, Minh Tang, Avanti Athreya, and Carey E. Priebe. Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electron. J. Stat.*, 8(2):2905–2922, 2014. doi: 10.1214/14-EJS978. URL `https://doi.org/10.1214/14-EJS978`.

[36] 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, 2017.

[37] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[38] Ashot Margaryan, Daniel J Lawson, Martin Sikora, Fernando Racimo, Simon Rasmussen, Ida Moltke, Lara M Cassidy, Emil Jørsboe, Andrés Ingason, Mikkel W Pedersen, et al. Population genomics of the viking world. *Nature*, 585(7825):390–396, 2020.

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

[40] Alexander Modell and Patrick Rubin-Delanchy. Spectral clustering under degree heterogeneity: a case for the random walk Laplacian. *arXiv preprint arXiv:2105.00987*, 2021.

[41] Peter Petersen. *Riemannian geometry*, volume 171 of *Graduate Texts in Mathematics*. Springer, 2006.

[42] Carey E Priebe, Youngser Park, Joshua T Vogelstein, John M Conroy, Vince Lyzinski, Minh Tang, Avanti Athreya, Joshua Cape, and Eric Bridgeford. On a two-truths phenomenon in spectral graph clustering. *Proceedings of the National Academy of Sciences*, 116(13):5995–6000, 2019.

[43] Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. Network embedding as matrix factorization: Unifying Deepwalk, LINE, PTE, and node2vec. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 459–467, 2018.

[44] Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *Annals of Statistics*, 39(4):1878–1915, 2011.

[45] Patrick Rubin-Delanchy. Manifold structure in graph embeddings. In *Proceedings of the Thirty-fourth Conference on Neural Information Processing Systems*, 2020.

[46] Patrick Rubin-Delanchy, Joshua Cape, Minh Tang, and Carey E Priebe. A statistical interpretation of spectral embedding: the generalised random dot product graph. *arXiv preprint arXiv:1709.05506*, 2020.

[47] Walter Rudin. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1976.

[48] Michael Salter-Townshend and Thomas Brendan Murphy. Variational bayesian inference for the latent position cluster model for network data. *Computational Statistics & Data Analysis*, 57(1):661–671, 2013.

[49] Francesco Sanna Passino, Nicholas A Heard, and Patrick Rubin-Delanchy. Spectral clustering on spherical coordinates under the degree-corrected stochastic blockmodel. *arXiv preprint arXiv:2011.04558*, 2020.

[50] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

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

[52] Martin Strohmeier, Xavier Olive, Jannis Lübbe, Matthias Schäfer, and Vincent Lenders. Crowd-sourced air traffic data from the opensky network 2019–2020. *Earth System Science Data*, 13 (2):357–366, 2021.

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

[54] Wilson A Sutherland. *Introduction to metric and topological spaces*. Oxford University Press, 2009.

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

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

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

[58] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[59] 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.

[60] Warren S Torgerson. Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4): 401–419, 1952.

[61] Michael W Trosset, Mingyue Gao, Minh Tang, and Carey E Priebe. Learning 1-dimensional submanifolds for subsequent inference on random dot product graphs. *arXiv preprint arXiv:2004.07348*, 2020.

[62] Jiaming Xu. Rates of convergence of spectral methods for graphon estimation. In *International Conference on Machine Learning*, pages 5433–5442. PMLR, 2018.

[63] Abubakar Yamin, Michael Dayan, Letizia Squarcina, Paolo Brambilla, Vittorio Murino, V Diwadkar, and Diego Sona. Comparison of brain connectomes using geodesic distance on manifold: A twins study. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1797–1800. IEEE, 2019.

[64] Yichi Zhang and Minh Tang. Consistency of random-walk based network embedding algorithms. *arXiv preprint arXiv:2101.07354*, 2021.

[65] 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.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Abstract and Section 1.
   (b) Did you describe the limitations of your work? [Yes] See Sections 2.1 and 3.1.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3 and Appendix A.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section 3.3 and Appendix A.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 4 and `https://github.com/anniegray52/graphs`.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] For Section 4.1 'set.seed(123)' has been used and supplementary figures for Section 4.2 in the Appendix report error bars.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.2.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 4 for data and we list all R packages used in Appendix B.
   (b) Did you mention the license of the assets? [Yes] For details on licenses of data see Sections 4.2, 4.3, and for R packages see Appendix B.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] `https://github.com/anniegray52/graphs`
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] All datasets used are public datasets.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]