研究

未知流形上潜在结构网络中的半监督回归

Aranyak Acharyya[1*], Joshua Agterberg[2], Michael W. Trosset[3], Youngser Park[4] 和 Carey E. Priebe[1]

*通讯作者： aachary6@jh.edu[1]应用数学与统计系，约翰斯·霍普金斯大学，美国巴尔的摩。所有作者信息完整列表见文章末尾。

摘要

随机图日益成为广泛应用领域中网络建模的重要研究对象。潜在位置随机图模型假设每个节点都关联一个潜在位置向量，且这些向量在潜在空间中遵循某种几何结构。本文研究随机点积图，其中两个节点之间形成一条边的概率由它们各自潜在位置的内积决定。我们假设潜在位置向量位于一条未知的一维曲线上，并通过回归模型与响应协变量相关联。利用潜在位置向量的底层几何结构，我们提出一种流形学习与图嵌入技术，用于预测样本外节点的响应变量，并为这些响应建立了收敛保证。我们的理论结果得到了模拟和一项针对果蝇脑部数据的应用的支持。

关键词： 网络推断；顶点协变量；随机点积图；流形学习；回归

arXiv:2305.02473v1 [stat.ML] 4 May 2023

1 引言

随机图长期以来一直是不同学科科学家关注的研究领域，这主要是因为它们可用于网络建模（[1],[2]）。潜在位置随机图（[3]）是一类随机图，其中每个节点都关联一个不可观测向量，即潜在位置。一种广受欢迎的模型是随机点积图模型（[4]），它属于网络模型的一个子类，其中一对节点之间形成边的概率由它们各自潜在位置向量的内积决定。此模型随后被推广为广义随机点积图模型（[5]），该模型以不定内积取代内积（见 [5]）。关于随机点积模型下推断问题的综述见 [6]。文献[7], 指出，在某些正则条件下，潜在位置随机图可等价地视为节点位于低维流形上的广义随机点积图，这为本文研究的模型提供了动机。考虑观测到一个随机点积图，其潜在位置位于环境空间 $\mathbb{R}^d$中的一条未知一维流形上，并假设在其中某些节点处记录了响应。我们选择在半监督设置下开展研究，因为在现实情形中，收集观测数据比获取与这些观测相对应的标签更容易。我们假设响应通过回归模型与相应潜在位置的标量原像相关联。在此半监督设置下，我们旨在预测样本外节点的响应。

网络分析问题中的半监督学习框架已在此前的多项研究中得到探讨。在 [8], 中，提出了一种针对含有有标签节点和无标签节点的图的正则化框架，用于预测无标签节点的标签。在 [9]中，基于一种图算法提出了一种降维技术，该算法旨在将高维环境空间中的数据表示在低维流形上。在具有潜在低维流形结构的潜在位置网络背景下， [10] 讨论了如何对随机点积图的潜在位置分布进行推断的问题，并假设这些潜在位置位于高维环境空间中的一个已知低维流形上。此外，[11]研究了随机点积图中的双样本假设检验问题，即检验均值是否相等，其中潜在位置位于高维环境空间中的一维流形上，而该流形是未知的，因此必须进行估计。更准确地说， [11] 提出了一种用于学习底层流形的方法，并证明了基于所得嵌入构建的统计检验的检验功效可以

基于对真实流形的了解，接近该检验的检验功效。

在本文中，我们研究了半监督设置下预测响应协变量的问题：在一个随机点积图中，其潜在位置位于环境空间中的未知一维流形上 $\mathbb{R}^d$。我们的主要结果建立了预测响应的收敛保证，其中流形是通过一种特定的流形学习程序学习得到的（见第 *2.3*节）。作为主要结果的推论，我们推导出基于所得嵌入的模型有效性检验功效的收敛保证。为帮助建立直观认识，我们首先在假设底层流形已知的情况下考虑回归参数估计问题，并证明在这种设置下某个特定估计量是相合的。

我们给出一个应用理论结果的示例。一个由100个幼虫果蝇 肯扬细胞神经元组成的连接组数据集（详情见 [12]）表明存在潜在的低维流形结构。每个节点（即每个肯扬细胞）都与一个响应协变量相联系，并使用邻接谱嵌入将每个节点的潜在位置估计为一个六维向量（见第*2.2*节）。对于估计潜在位置的每一对维度，均可得到一个散点图，因而得到一个 6×6 散点图矩阵（图 *5*）。可以看出，每个维度都与另一个维度近似相关，因此假设潜在位置位于六维环境空间中的一维流形上。为捕捉底层结构，我们在估计的潜在位置上构造定位图，并将最短路径距离的差异度矩阵嵌入一维空间（嵌入方法的描述见第 *2.3* 节）。在图 *1*中给出了估计潜在位置前两个维度的散点图，其中点的大小随相应响应协变量的取值而变化。还给出了响应 y_i 与一维嵌入 $\hat{z}_i$ 之间的散点图，以及表明显著效应的拟合回归线。这些结果表明，假定响应通过简单线性回归模型与嵌入相联系是合理的。

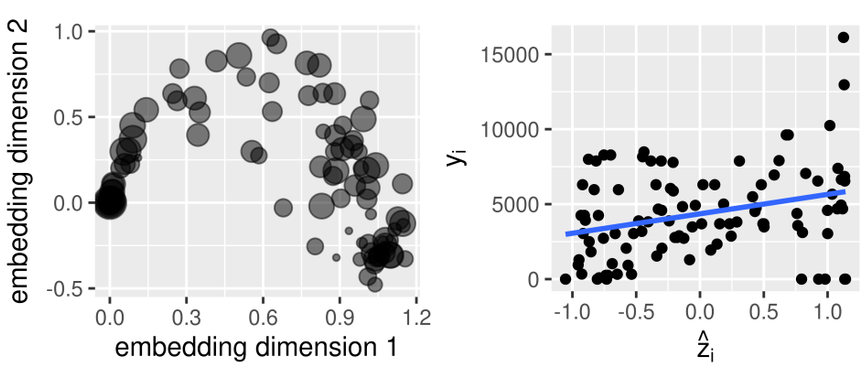

图 1 未知流形上潜在结构网络中的响应预测示意应用。我们将该方法应用于果蝇幼虫蘑菇体右半球的连接组。左图：100 个肯扬细胞神经元估计潜在位置的两个维度散点图，这些位置由网络的谱嵌入得到；点的大小表示响应变量 y_i （神经元 i 的束入口点与蘑菇体神经网 neuropil 之间的距离，单位为微米）。右图：100 个肯扬细胞神经元的响应 y_i 与学习得到的一维嵌入$\hat{z}_i$ 的关系图，该嵌入近似表示沿此曲线的测地距离，并附有回归线。左图显示，一维曲线捕捉到了谱嵌入中的非线性结构。右图显示，以测地距离为自变量对响应进行回归，表明存在显著效应（$p < 0.01$ 对于 H_0： $a = 0$，在 $y_i = a\hat{z}_i + b + \eta_i$中）。

我们的分析提出了一个问题：如何检验这样一个简单线性回归模型的有效性——该模型假设节点响应与潜在位置的标量原像之间存在联系。我们的理论表明，基于原始应力嵌入的模型有效性检验功效，近似于基于真实回归变量的检验功效。

在第*2*节中，我们讨论随机点积图和流形学习的关键要点。第*3* 节讨论模型以及如何对

相应的回归模型进行后续推断的方法。第4 节介绍已知流形和未知流形两种情形下的理论结果。第5 节介绍模拟结果。第6节回顾连接组应用。第7 节讨论结果，并提出一些有待进一步研究的问题。理论结果的证明见第8节。

2 背景符号、定义与结果

本节将介绍并解释全文中使用的符号。我们还将陈述与随机点积第 2.2节中的图第 2.3节中的流形学习）相关的定义和结果。

2.1 符号

我们用粗体小写字母表示向量，例如 $\mathbf{x}$。粗体大写字母（如 $\mathbf{P}$）用于表示矩阵。矩阵 $\mathbf{B}$ 的 Frobenius 范数和最大行范数分别记为 $\|\mathbf{B}\|_F$ 和 $\|\mathbf{B}\|_{2,\infty}$。矩阵 $\mathbf{B}$ 的第 i 行记为 $\mathbf{B}_{i*}$，$\mathbf{B}$ 的第 j 列记为 $\mathbf{B}_{*j}$。我们用 $\mathbf{I}_n$ 表示 $n \times n$ 单位矩阵，用 $\mathbf{J}_n$ 表示每个元素都等于 1 的 $n \times n$ 矩阵。此外，$\mathbf{H}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{J}_n$ 表示 $n \times n$ 中心化矩阵。除非另有说明，$\|\mathbf{x}\|$表示向量 $\mathbf{x}$ 的欧几里得范数。所有正交 $d \times d$ 矩阵构成的集合记为 $\mathcal{O}(d)$。所有正整数构成的集合记为 $\mathbb{N}$；对于任意 $n \in \mathbb{N}$，$[n]$ 表示集合 $\{1, 2, 3..., n\}$。

2.2 随机点积图预备知识

图是一个有序对 (V, E)，其中 V是顶点（或节点）的集合，$E \subset V \times V$是连接顶点的边的集合。图的邻接矩阵 $\mathbf{A}$ 定义为：若 $(i, j) \in E$，则 $\mathbf{A}_{ij} = 1$；否则，$\mathbf{A}_{ij} = 0$。这里我们讨论的是空对角且无向的图，因此 $\mathbf{A}$ 是对称矩阵，并且对所有 i，$\mathbf{A}_{ii} = 0$。潜在位置随机图是指每个节点都与一个向量相关联的随机图，该向量称为其潜在位置，记作 $\mathbf{x}_i$；第 i 个节点与第 j 个节点之间形成边的概率由 $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ 给出，其中 κ 是一个适当的核函数。

从任意概率分布中抽取的随机向量不能作为随机点积图的潜在位置，因为其模可能无界，而概率必须位于区间 $[0, 1]$ 内。以下定义使我们能够使用一类受限的分布，这类分布更适用于随机点积图。

定义 1（内积分布）：若 F 是定义在$\mathbb{R}^d$上的概率分布函数，且对于任意$\mathbf{x}$、$\mathbf{y} \in \text{supp}(F)$，$\mathbf{x}^T\mathbf{y} \in [0, 1]$，则称 F为 $\mathbb{R}^d$上的内积分布。

接下来，我们定义随机点积图，它是本文所研究模型的基础。

定义 2（随机点积图）：设 G 是一个无自环无向随机图，其潜在位置为$\mathbf{x}_1$、.....、$\mathbf{x}_n \in \mathbb{R}^d$。令$\mathbf{X} = [\mathbf{x}_1|\ldots|\mathbf{x}_n]^T$为其潜在位置矩阵，$\mathbf{A}$ 为其邻接矩阵。若对所有 $i < j$，$\mathbf{A}^{ij}$ ~服从伯努利分布$(\mathbf{x}_i^T\mathbf{x}j)$，且相互独立，则称图 G 为随机点积图。$\mathbf{A}$ 的概率分布为

$$P[\mathbf{A}] = \Pi_{i<j}(\mathbf{x}_i^T\mathbf{x}_j)^{\mathbf{A}_{ij}}(1 - \mathbf{x}_i^T\mathbf{x}_j)^{1-\mathbf{A}_{ij}}.$$

若$\mathbf{x}_1$，.....$\mathbf{x}_n \sim^{iid} F$ 是潜在位置，且 F 为内积分布，则记作$(\mathbf{A}, \mathbf{X}) \sim RDPG(F)$。给定$\mathbf{x}_1$，.....$\mathbf{x}_n$时，邻接矩阵的条件分布为

$$P[\mathbf{A}|\mathbf{X}] = \Pi_{i<j}(\mathbf{x}_i^T\mathbf{x}_j)^{\mathbf{A}_{ij}}(1 - \mathbf{x}_i^T\mathbf{x}_j)^{1-\mathbf{A}_{ij}}.$$

随机点积图的潜在位置通常是未知的，在实际应用中需要进行估计。以下定义通过邻接谱嵌入给出了这些潜在位置的一种估计。

定义 3 *(*邻接谱嵌入*)*：设 λ_i 是$\mathbf{A}$的第 i大（按绝对值计）特征值，令$\mathbf{v}_i$ 为对应的正交归一特征向量。定义$\hat{}\mathbf{S}$ $=\text{diag}(\lambda_1,\ldots\lambda_d)$ 和$\mathbf{V}$ $=[\mathbf{v}_1|\ldots|\mathbf{v}_d]$。我们将$\mathbf{X}$定义为$\mathbf{A}$ 到 $\mathbb{R}^d$的邻接谱嵌入，记作$\hat{}\mathbf{X}$ $=\mathbf{V}|\mathbf{S}|^{\frac{1}{2}}$。

现在，我们介绍文献中的两个结果，它们给出了随机点积图中潜在位置经适当旋转后的邻接谱估计的一致性和渐近正态性。

定理 *1*（ *[6]*中的定理 *8*）：设 x_1,*x2, ...x*$_n \in \mathbb{R}^d$ 表示一个包含 n 个节点的随机点积图的潜在位置，并令 $X^{(n)} = [x_1|x_2|...|x_n]^T$ 为潜在位置矩阵。假设对于所有充分大的 n， $rank(X^{(n)})$ $= d$，且存在 $a > 0$ 使得 b_n $= max_{i\in[n]} \sum_{j=1}^{d} x_i^T x_j$ $> (log(n))^{4+a}$。设 $A^{(n)}$ 表示相应的邻接矩阵，并令 $\hat{}X^{(n)}$ 为 $A^{(n)}$ 在 $\mathbb{R}^d$ 中的邻接谱嵌入。存在一个常数 $C > 0$ 和一个序列 $W^{(n)} \in \mathcal{O}(d)$, 使得

$$\lim_{n\to\infty} \mathbb{P}\left[\max_{i\in[n]} \left\|(\hat{\mathbf{X}}^{(n)}\mathbf{W}^{(n)} - \mathbf{X}^{(n)})_{i*}\right\| \leq \frac{C(\log(n))^2}{\sqrt{b_n}}\right] = 1 \tag{1}$$

注意，在假设 $b_n > \log(n))^{4+a}$ 的条件下，式（*1*）意味着 $\left\|\hat{\mathbf{X}}^{(n)}\mathbf{W}^{(n)} - \mathbf{X}^{(n)}\right\|_{2,\infty} \to^P 0 \ as \ n \to \infty.$

此后，我们将这个经过最优旋转的邻接谱嵌入记为 $\tilde{}\mathbf{X}^{(n)}$，即 $\tilde{}\mathbf{X}^{(n)} = \hat{}\mathbf{X}^{(n)}\mathbf{W}^{(n)}$。为简化记号，我们常常省略上标 n，并且将 $\tilde{}\mathbf{x}_i$ 用来表示$\mathbf{X}$的第 i 行。值得注意的是，在试图根据邻接矩阵估计真实潜在位置时，我们会遇到一个固有的不可识别性问题：对于任意$\mathbf{W}$ $\in \mathcal{O}$（d）， $\mathbb{E}$（$\mathbf{A}|\mathbf{X}$） $= \mathbf{X}\mathbf{X}^T =$ （$\mathbf{X}\mathbf{W}$）（$\mathbf{X}\mathbf{W}$）^{T}。这就是邻接谱嵌入需要经过适当旋转，以便近似真实潜在位置矩阵的原因。

定理 2（ *[6]* 中的定理 9）：设($\mathbf{A}^{(n)}$、$\mathbf{X}^{(n)}$) $\sim RDPG(F)$ 为一系列随机点积图的邻接矩阵和潜在位置矩阵，其中潜在位置由 F 上的内积分布生成。设

$$\mathbf{\Sigma}(\mathbf{x_0}) = \Delta^{-1}\mathbb{E}_{\mathbf{x}\sim F}\left[\mathbf{x_0}^T\mathbf{x}(1 - \mathbf{x_0}^T\mathbf{x})\mathbf{x}\mathbf{x}^T\right]\Delta^{-1} \tag{2}$$

其中 $\Delta = \mathbb{E}_{\mathbf{x}\sim F}[\mathbf{x}\mathbf{x}^T]$。于是，存在一个序列 $\mathbf{W}^{(n)} \in \mathcal{O}(d)$，使得对于所有 $\mathbf{u} \in \mathbb{R}^d$,

$$\lim_{n\to\infty} \mathbb{P}\left[\sqrt{n}(\hat{\mathbf{X}}^{(n)}\mathbf{W}^{(n)} - \mathbf{X}^{(n)})_{i*} \leq \mathbf{u}\right] = \int_{\mathbf{x}\in supp(F)} \Phi_{\mathbf{0},\mathbf{\Sigma}(\mathbf{x})}(\mathbf{u})dF(\mathbf{x}) \tag{3}$$

其中 $\Phi_{\mathbf{0},\mathbf{\Sigma}(\mathbf{x})}(.)$ 表示多元正态 $N(\mathbf{0}, \mathbf{\Sigma}(\mathbf{x}))$ 分布。

在适当的正则条件下，结合使用定理 2 和德尔塔方法，可以得到函数 $\gamma: \mathbb{R}^d \to \mathbb{R}$ 的 $\sqrt{n}(\gamma(\tilde{}\mathbf{x}_i) - \gamma(\mathbf{x}_i))$ 的渐近分布，该函数依赖于潜在位置的真实分布 F。因此，可以根据最优旋转的邻接谱估计 $\tilde{}\mathbf{x}_i$ 及其经验分布函数来近似 $\text{var}(\gamma(\tilde{}\mathbf{x}_i) - \gamma(\mathbf{x}_i))$。对于一种随机点积图，其潜在位置位于环境空间 $\mathbb{R}^d$中已知的一维流形上，且节点响应通过简单线性回归模型与潜在位置的标量原像相关联，我们可以使用 $(\gamma(\tilde{}\mathbf{x}_i) - \gamma(\mathbf{x}_i))$ 的近似方差（其依据是 [13]第 2、3 章中的研究），以提升回归参数朴素估计量的性能（以均方误差衡量）。这些朴素估计量是通过在回归参数的最小二乘估计中用 $\tilde{}\mathbf{x}_i$ 替换 $\mathbf{x}_i$ 得到的。我们将在第 7 节详细展示这一方法（见图 7）。

注 *1* 设 F 是一个满足 *supp(F)* $= K \subset \mathbb{R}^d$ 的概率分布，且 *z1,....z*$_n \sim^{iid} F$ 是具有相关核 κ 的空对称潜在位置随机图的潜在位置。 *[7]* 中的许多研究表明，若 $\kappa \in L^2(\mathbb{R}^d \times \mathbb{R}^d)$，则存在映射 $q : \mathbb{R}^d \to L^2(\mathbb{R}^d)$，使得该图可等价地表示为一个潜在位置为 $x_i = q(z_i) \in L(\mathbb{R}^d)$ 的广义随机点积图。若假设 κ 是指数为 c 的 *H¨older* 连续函数，则 $q(K)$ 的豪斯多夫维数可由 $\frac{d}{c}$ 界定，如 *[7]* 所示。 *[14]*，已证明，若 K 是黎曼流形，则在更强的假设下可得出结论： $q(K) \subset L^2(\mathbb{R}^d)$ 也是一个与 K 微分同胚的黎曼流形。因此，在适当的正则性假设下，任何潜在位置图都可以视为潜在位置位于低维流形上的广义随机点积图。

在阐述了随机点积图相关的定义和结果之后，我们将在下一节介绍本文所采用的流形学习技术。为避免歧义，需要说明的是，一般意义上的流形学习与随机点积图模型毫无关系；因此第 *2.2* 节 和 第 2.3 节可以独立阅读。

2.3 通过原始应力最小化进行流形学习

我们的主要模型基于随机点积图，其潜在位置位于一维黎曼流形上。由于一维黎曼流形与一维欧几里得空间等距，我们希望将潜在位置表示为实线上的点。这正是采用以下流形学习技术的动机；该技术依赖于通过定位图上的最短路径距离来近似测地线（[15], [16], [17]）。给定点 $\mathbf{x}_1, \dots \mathbf{x}_n \in \mathcal{M}$，其中 $\mathcal{M}$ 是环境空间 $\mathbb{R}^d$ 中一个未知的一维流形，目标是找到 $\hat{}z_1, \dots \hat{}z_n \in \mathbb{R}$，使得 $\hat{}z_i$ 之间的点间欧几里得距离近似等于 $\mathbf{x}_i$ 之间的点间测地距离。然而，$\mathbf{x}_i$ 之间的点间测地距离是未知的。以下结果表明，在适当的正则性假设下，如何估计这些未知的测地距离。

定理 3 （来自 *[17]* 的定理3 ）假设 $\mathcal{M}$ 是环境空间 $\mathbb{R}^d$中的一维紧致黎曼流形。设 r_0 和 s_0 分别为 $\mathcal{M}$的最小曲率半径和最小分支间距。假设给定 ν，并假设 $\lambda > 0$ 的选取满足 $\lambda < s_0$和 $\lambda < \frac{2}{\pi} r_0 \sqrt{24\nu}$。此外，假设 $\mathbf{x}_1$、.....$\mathbf{x}_n \in \mathcal{M}$ 满足，对于每个 $\mathbf{u} \in \mathcal{M}$，$d_M(\mathbf{u}, \mathbf{x}_i) < \delta$。按照以下规则，以 $\mathbf{x}_i$ 为节点构建定位图：如果 $\|\mathbf{x}_i - \mathbf{x}_j\| < \lambda$，则两个节点 $\mathbf{x}_i$ 和$\mathbf{x}_j$ 之间连接一条边。当 $\delta < \frac{\nu\lambda}{4}$ 时，对于所有 i、$j \in [n]$，以下条件成立，

$$(1-\nu) d_M(\mathbf{x}_i, \mathbf{x}_j) \leq d_{n,\lambda}(\mathbf{x}_i, \mathbf{x}_j) \leq (1+\nu) d_M(\mathbf{x}_i, \mathbf{x}_j),$$

其中，$d_{n,\lambda}(\mathbf{x}_i, \mathbf{x}_j)$ 表示$\mathbf{x}_i$ 与$\mathbf{x}_j$之间的最短路径距离。

给定差异度矩阵 $\mathbf{D} = (d_{n,\lambda}(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n$，$(z_1, \dots z_n)$ 处的原始应力函数定义为

$$\sigma(z_1, \dots z_n) = \sum_{i<j} w_{ij} (|z_i - z_j| - d_{n,\lambda}(\mathbf{x}_i, \mathbf{x}_j))^2$$

其中 $w_{ij} \geq 0$ 为权重。为了学习流形 $\mathcal{M}$，我们对所有i, j设定 $w_{ij} = 1$，并计算

$$(\hat{z}_1, \dots \hat{z}_n) = \arg\min \sigma(z_1, \dots z_n) = \arg\min \sum_{i<j} (|z_i - z_j| - d_{n,\lambda}(\mathbf{x}_i, \mathbf{x}_j))^2.$$

由于标量 $\hat{}z_i$ 是通过最小化原始应力，将$\mathbf{D}$ 嵌入一维空间而得到的，因此我们今后将 $\hat{}z_i$ 称为$\mathbf{D}$的一维原始应力嵌入。

注释 2 在实践中，*raw-stress* 通过迭代主化（ *[18]* 的第8 章）进行数值最小化。标准算法有时会陷入非全局最小值。然而，当使用经典多维尺度分析初始化配置时，通过重复迭代古特曼变换（*[18]* 的第8 章）可以获得近似最优值。在本文中，对于理论结果，我们假设已达到全局最小值。

3 模型与方法论

这里我们介绍在已知流形和未知流形两种假设下的模型。在这两种情况下，我们均假设观测到一个随机点积图，其中节点的潜在位置位于 d 维环境空间中的一维流形上。在底层流形已知的假设下，每个节点都与一个响应相关联，该响应通过回归模型与相应潜在位置的标量原像相关联，我们的目标是估计回归参数。当假设底层流形未知时，我们的模型涉及一个包含少量有标签节点和大量无标签节点的网络，我们的目标是在假设有标签节点的响应通过回归模型与各自潜在位置的标量原像相关联的情况下，预测给定无标签节点的响应。在流形未知的情形下，我们只能在尺度变换和平移变换意义下近似真实回归变量；由于这一不可辨识性问题，当流形未知时，我们选择预测响应，而非估计回归参数。

备注 3 我们想在此提醒读者，已知流形的设定并不现实。我们考虑已知流形的设定，是为了帮助读者熟悉最佳情形，以便与在未知流形这一现实设定下得到的结果进行比较。

3.1 已知流形上的回归参数估计

假设 $\psi: \mathbb{R} \to \mathbb{R}^d$ 是一个已知的双射函数，且令 $\mathcal{M} = \psi([0, L])$为一维紧黎曼流形。考虑一个随机点积图，其节点（潜在位置为 $\mathbf{x}_i$）位于已知的一维流形 $\mathcal{M}$ 中，该流形嵌入于 d维环境空间。令 $t_1, \dots t_n$为潜在位置的标量原像，使得对于所有 $i \in [n]$，$\mathbf{x}_i = \psi(t_i)$，其中 n 为图中的节点数。假设对于每个 i，第 i 个节点都与响应 y_i 相关联，该响应通过以下回归模型与潜在位置相联系。

$$y_i = \alpha + \beta t_i + \epsilon_i, i \in [n] \tag{4}$$

其中，对于所有 $i \in [n]$，$\epsilon_i \sim^{iid} N(0, \sigma_\epsilon^2)$。我们的目标是估计 α和 β。
如果已知真实回归变量 t_i，我们可以通过以下普通最小二乘估计来估计 α 和 β：

$$\hat{\beta}_{true} = \frac{\sum_{i=1}^n (y_i - \bar{y})(t_i - \bar{t})}{\sum_{i=1}^n (t_i - \bar{t})^2}, \quad \hat{\alpha}_{true} = \bar{y} - \hat{\beta}_{true}\bar{t}. \tag{5}$$

由于真实潜在位置 $\mathbf{x}_i$ 未知，我们通过 $\hat{}t_i = \arg\min_t \|\tilde{}\mathbf{x}_i - \psi(t)\|$ 来估计真实回归变量 t_i，其中 $\tilde{}\mathbf{x}_i$ 是第i个潜在位置 $\mathbf{x}_i$的最优旋转的邻接谱估计。由于流形$\mathcal{M} = \psi([0, L])$的紧致性，$\hat{}\ t_i$ 的存在性得以保证。随后，我们在 $\hat{}\alpha_{true}$ 和 $\hat{}\ \beta_{true}$ 中用 $\hat{}\ t_i$ 替代 t_i，从而得到以下替代估计量（或代入估计量）：

$$\hat{\beta}_{sub} = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{t}_i - \bar{\hat{t}})}{\sum_{i=1}^n (\hat{t}_i - \bar{\hat{t}})^2}, \quad \hat{\alpha}_{sub} = \bar{y} - \hat{\beta}_{sub}\bar{\hat{t}}. \tag{6}$$

$\hat{}$ 计算 $\hat{}\alpha_{sub}$ 和 β_{sub} 的步骤正式陈述于 算法 *1a*中。

算法1a EST($\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{W} \in \mathcal{O}(d)$ d, $\psi: [0, L] \to \mathbb{R}^d$, $\{y_i\}_{i=1}^n$)

1：将邻接矩阵$\mathbf{A}$ 进行邻接谱嵌入$\mathbf{X}$ $\hat{}$，嵌入到 $\mathbb{R}^d$ 中。2：使用给定的旋转矩阵 $\mathbf{W}$，得到最优旋转后的邻接谱嵌入 $\tilde{}\mathbf{X} = \hat{}\mathbf{X}\mathbf{W}$。3：通过 $\hat{}t_i = \arg\min_t \|\tilde{}\mathbf{x}_i - \psi(t)\|$，获得估计潜在位置在流形上的投影的原像。4：计算给定的替代估计量 $\hat{}\beta_{sub} = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{t}_i - \bar{\hat{t}}\hat{})}{\sum_{i=1}^n (\hat{t}_i - \bar{\hat{t}}\hat{})^2}$ $'\hat{\alpha}_{sub} = \bar{y} - \hat{}\beta_{sub}\bar{\hat{t}}$.5：**返回** ($\hat{\alpha}_{sub}$, $\hat{}\ \beta_{sub}$)。

3.2 未知流形上的响应预测

这里，我们假设

$$\psi \quad [0,L] \to \mathbb{R}^d \qquad \left\|\dot{\psi}\ t\ \right\| =: \text{未知且}$$

1 t

经过弧长参数化，即对所有，( )。此外，假设 $\mathcal{M} = \psi([0,L])$ 是一个紧黎曼流形。考虑一个 n 节点的随机点积图，其节点（潜在位置为$\mathbf{x}_i$）位于环境空间 $\mathbb{R}^d$ 中的未知流形 $\mathcal{M} = \psi([0,L])$ 上。假设图中的前 s 个节点与一个响应协变量相关联，并且第 i 个节点处的响应 y_i 通过线性回归模型与潜在位置相关联

$$y_i = \alpha + \beta t_i + \epsilon_i, i \in [s] \tag{7}$$

其中，对于所有 $i \in [s]$，$\epsilon_i \sim^{iid} N(0,\sigma_\epsilon^2)$。我们的目标是预测第 r 个节点的响应，其中 $r > s$。首先，我们计算所有 n 个节点的潜在位置的邻接谱估计 $\hat{\mathbf{x}}_i$ 。然后，按照以下规则，根据邻接谱估计 $\hat{\mathbf{x}}_i$构建定位图：当且仅当对于某个预先确定的 $\lambda > 0$ （称为邻域参数），$\|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\| < \lambda$时，连接两个节点 $\hat{\mathbf{x}}_i$ 和 $\hat{\mathbf{x}}_j$ 。记 $\hat{\mathbf{x}}_i$ 与 $\hat{\mathbf{x}}_j$ 之间的最短路径距离为 d_n，将差异度矩阵 $\mathbf{D} = (d_n,\lambda(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j))_i^l$ 嵌入一维空间，方法是最小化原始应力准则，从而得到

$$(\hat{z}_1, ...\hat{z}_l) = \arg\min \sum_{i=1}^{l}\sum_{j=1}^{l}(|z_i - z_j| - d_{n,\lambda}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j))^2 \tag{8}$$

其中 l 满足 $s < r < l \leq n$。随后，我们对二元数据 $(y_i, \hat{z}_i)_{i=1}^s$ 使用简单线性回归模型，以预测与 z_r 对应的 r-th 节点的响应。根据给定观测值预测 r-th 节点响应的上述过程，形式化描述于 算法 *1b*。

算法 1b PRED($\mathbf{A} \in \mathbb{R}^{n\times n}$, d, λ, l, $\{y_i\}_{i=1}^s$, r)

1：从邻接矩阵 A 中获取潜在位置的邻接谱估计 $\hat{\mathbf{x}}_1 \ldots \hat{\mathbf{x}}_n \in \mathbb{R}^d$。2：按照以下规则，以 $\hat{\mathbf{x}}_i$ 为顶点构造定位图：当且仅当 $\|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\| < \lambda$ 时，连接两个顶点 $\hat{\mathbf{x}}_i$、$\hat{\mathbf{x}}_j$ 。3：对于每个 i、$j \in [n]$，获取最短路径距离 d_n、$\lambda(\hat{\mathbf{x}}_i$、$\hat{\mathbf{x}}_j)$。4：获取 $(\hat{z}_1, ...\hat{z}_l) = \arg\min\sum_{i=1}^l \sum_{j=1}^l(|z_i - z_j| - d_n$、$\lambda(\hat{\mathbf{x}}_i$、$\hat{\mathbf{x}}_j))^2$。5：计算 $\hat{b} = \frac{\sum_{i=1}^s(y_i-\bar{y})(\hat{z}_i-\bar{\hat{z}})}{\sum_{i=1}^s(\hat{z}_i-\bar{\hat{z}})^2}$，以及 $\hat{a} = \bar{y} - \hat{b}\bar{\hat{z}}$。6：对于 $r > s$，计算 $\tilde{y}_r = \hat{a} + \hat{b}\hat{z}_r$。7：返回 $\tilde{y}_r$。

---

4 主要结果

本节介绍我们的理论结果，展示在已知流形上回归参数估计量的一致性，以及基于未知流形上的原始应力嵌入所得到的预测响应的收敛保证。在未知流形的情形下，作为预测响应一致性的推论，我们还推导出一个收敛保证，用于检验基于近似F-统计量的简单线性回归模型的有效性。

4.1 已知流形的情形

回顾一下，我们观测到一个具有 n 个节点的随机点积图，其潜在位置$\mathbf{x}_i$ 位于一维流形上。下面的结果表明，我们可以通过$(\hat{\alpha}_{sub}, \beta_{sub})$一致地估计$(\alpha, \beta)$。

定理 *4* 假设 $\psi: [0,L] \to \mathbb{R}^d$ 是双射，且其逆映射 γ 满足：对于所有 $w \in \psi([0,L])$，$\|\nabla\gamma(w)\| < K$，其中 $K > 0$ 为某个值。设 $x_i = \psi(t_i)$ 为具有 n 个节点的随机点积图中第 i 个节点的潜在位置，并假设对于所有 $i \in [n]$，$y_i = \alpha + \beta t_i + \epsilon_i, \epsilon_i \sim^{iid} N(0,\sigma_\epsilon^2)$。假设对于所有 i，$x_i \sim^{iid} F$，其中 F 是定义在 $\mathbb{R}^d$ 上的内积分布。令 $X^{(n)} = [x_1|\ldots|x_n]$ 为潜在位置矩阵，并设 $\hat{X}^{(n)}$ 是邻接矩阵 $A^{(n)}$ 在 $\mathbb{R}^d$ 维空间中的邻接谱嵌入。假设已知 $\hat{W}^{(n)} = arg\ min_{\mathbf{W}\in\mathcal{O}(d)} \left\|\hat{X}^{(n)}W - X^{(n)}\right\|_F$。那么，当 $n \to \infty$ 时，我们有 $\hat{\alpha}_{sub} \to^P \alpha$ 和 $\hat{\beta}_{sub} \to^P \beta$，其中 $(\hat{\alpha}_{sub}, \hat{\beta}_{sub}) = EST(A^{(n)}, d, \hat{W}^{(n)}, \psi, \{y_i\}_{i=1}^n)$（见算法*1a*）。

定理 4 的证明思路概要如下。注意，$\hat{t}_i = \arg\min_t \|\tilde{\mathbf{x}}_i - \psi(t)\|$，其中$\tilde{\mathbf{x}}_i$是 $\mathbf{x}_i$ 的最优旋转邻接谱估计，而 $t_i = \arg\min_t \|\mathbf{x}_i - \psi(t)\|$。回顾定理 1，它告诉我们 $\tilde{\mathbf{x}}_i$ 对 $\mathbf{x}_i$ 是相合的。这使我们能够证明，$\hat{t}_i$ 对 t_i 是相合的，最终可得 $\hat{\alpha}_{sub}$ 和 β_{sub} 的一致性，因为 $\hat{\alpha}_{sub}$ 和 $\hat{\beta}_{sub}$ 是通过在 $\hat{\alpha}_{true}$ 和 β_{true} 的表达式中用 $\hat{t}_i$ 替换真实回归变量 t_i 计算得到的，而这两个表达式分别对 α 和 β 是相合的。定理 4 在逆函数的梯度有界这一假设下，给出了回归参数替代估计量的一致性。由于连续可微函数在其定义域的紧子集上具有有界梯度，因此只要 $\gamma = \psi^{-1}$ 能表示为一个具有连续偏导数的函数在某个集合上的限制，我们就可以应用定理 4。作为定理 4 的直接推论，我们接下来的结果表明，在所有线性无偏估计量中，替代估计量具有最优渐近方差。

推论1：在真实回归变量条件下 t_i 在定理4的设置中，以下两个条件成立

(A) $\mathbb{E}(\hat{\alpha}_{sub}) \to \alpha$, $\mathbb{E}(\beta_{sub}) \to \beta$ 当 $n \to \infty$,$\tilde{}$时；*(B)* 对于任意两个线性无偏估计量 $\tilde{\alpha}$ 和 β 以及任意 $\delta > 0$，$\hat{}$ $\tilde{}\mathrm{var}(\hat{\alpha}_{sub}) \leq \mathrm{var}(\tilde{\alpha}) + \delta$，$\mathrm{var}(\beta_{sub}) \leq \mathrm{var}(\beta) + \delta$ 当 n足够大时。

4.2 未知流形的情形

回顾一下，我们的目标是在环境空间中的未知一维流形上的随机点积图上，在半监督设置下预测响应 $\mathbb{R}^d$。为此，我们通过证明算法*1b* 所得到的预测响应 $\tilde{y}_r$ 在第r个节点上基于原始应力嵌入的预测响应，随着 t_i 趋近于 $n \to \infty$，将其使用合理化。

直觉上，为了对回归模型进行推断，我们必须学习未知流形 $\mathcal{M}$。我们利用大量无标签节点的可用性，这些节点的潜在位置位于一维流形上，以学习该流形。注意，由于底层流形 $\psi([0, L])$ 是弧长参数化的，流形上任意两点之间的测地距离都与其对应原像之间的点间距离相同。文献中的研究结果（[16]）表明，如果在流形上足够多的点上构建恰当的定位图，那么两点之间的最短路径距离就能近似这两点之间的测地距离。因此，在基于邻接谱估计 $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}2, ..., \hat{\mathbf{x}}_n$ 构建的、邻域参数 λ 选取恰当的定位图上，$\hat{\mathbf{x}}_i$ 与 $\hat{\mathbf{x}}j$ 之间的最短路径距离 $d_{n,\lambda}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}j)$ 预计将近似测地距离 $d_M(\mathbf{x}_i, \mathbf{x}j) = |t_i - t_j|$。请注意，这里无需旋转邻接谱估计 $\hat{\mathbf{x}}_i$，因为最短路径距离是欧氏距离之和，而欧氏距离在正交变换下保持不变。因此，将差异度矩阵 $\mathbf{D} = (d_{n,\lambda}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}j))_{i,j=1}^{l}$ 通过最小化原始应力嵌入到一维空间后，我们得到嵌入 $\hat{z}_1, \ldots \hat{z}_l$，其点间距离$|\hat{z}_i - \hat{z}_j|$ 近似点间距离 $|t_i - t_j|$。换言之，对潜在位置的邻接谱估计应用原始应力嵌入所得到的估计距离，近似真实测地距离；以下结果将对此加以说明。这个论证是构造一系列基于原始应力嵌入的预测响应的基础；随着辅助节点数量趋于无穷，这些预测响应将趋近于基于真实回归变量得到的预测响应。

定理 *5*（*[11]*中的定理*4*）：假设函数 ψ*:* $[0, L] \to \mathbb{R}^d$满足以下条件：对所有 $t \in [0, L]$，均有 $\left\|\dot{\psi}(t)\right\| = 1$。令 $x_i \in \mathbb{R}^d$ 为随机点积图中第 i 个节点的潜在位置，并令 $t_i \in \mathbb{R}$ 满足：对所有 i，均有 $x_i = \psi(t_i)$；同时假设 $t_i \sim^{iid} g$，其中 g 是满足 $supp(g) = [0, L]$ 且满足 $g_{min} > 0$ 的绝对连续概率密度函数。令 $\hat{x}_i$ 为对所有 i 的真实潜在位置 x_i 的邻接谱估计。存在节点数序列 $\{n_K\}_{K=1}^{\infty}$ 和邻域参数序列 $\{\lambda_K\}_{K=1}^{\infty}$，满足$n_K \to \infty$, $\lambda_K \to 0$ （当 $K \to \infty$ 时），使得对于任意满足以下条件的固定整数 l：对所有K，均有 $s < l < n_K$，

$$(|\hat{z}_i - \hat{z}_j| - |t_i - t_j|) \to^P 0, \ for\ all\ i, j \in [l] \tag{9}$$

成立，其中$(\hat{}z_1, ...\hat{}z_l)$ 是原始应力准则的极小化解，即

$$(\hat{z}_1, ...\hat{z}_l) = arg\ min \sum_{i=1}^{l}\sum_{j=1}^{l}(|z_i - z_j| - d_{n_K,\lambda_K}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j))^2. \tag{10}$$

定理 5表明，一维原始应力嵌入 $\hat{}z_1$、.....$\hat{}z_l$ 满足 $(|\hat{}z_i - \hat{}z_j| \dashv |t_i - t_j|) \to 0$ 当 $K \to \infty$ 时，对所有 i，$j \in [l]$均成立。这意味着，对于每个 $i \in [l]$，$\hat{}z_i$ 都可以近似表示t_i ，且只相差尺度和平移变换。由于在简单线性回归中，预测响应的准确性不受尺度和平移变换影响，因此我们可以预期，基于 $\hat{}z_i$ 得到的特定节点的预测响应将趋近于基于真实回归变量 t_i得到的预测响应。本文的核心结果——下述定理——证明事实确实如此。

定理 6 设有一个随机点积图，其中每个节点都位于一维流形 $\psi([0, L])$ 上，该流形经过弧长参数化，且 ψ 未知。令 $x_i = \psi(t_i)$ 表示第 i 个节点的潜在位置，对所有 i 均成立。假设 $y_i = \alpha + \beta t_i + \epsilon_i$，$\epsilon_i \sim^{iid} N(0, \sigma_\epsilon^2)$ 对 $i \in [s]$ 均成立，其中 s 是一个固定整数。基于真实回归变量 $\hat{}$ 得到的第 r 个节点的预测响应为 $\hat{}y_r = \hat{}\alpha_{true} + \beta_{true} t_r$。存在节点数序列 $n_K \to \infty$ 和邻域参数序列 $\lambda_K \to 0$，使得对每个 $r > s$，$|\hat{}y_r - \tilde{y}_r^{(K)}| \to^P 0$，当 $K \to \infty$ 时成立，其中$\tilde{y}_r^{(K)} = PRED(A^{(K)}, d, \lambda_K, l, \{y_i\}_{i=1}^{s}, r)$（见算法 1b），$A^{(K)}$ 是节点数为 n_K 时的邻接矩阵，l 是满足 $l > r > s$ 的固定自然数。

回顾一下，简单线性回归模型的有效性可以通过 F-检验进行检验，其检验统计量取决于基于真实回归变量得到的预测响应。由于我们能够利用基于原始应力嵌入的预测响应来近似基于真实回归变量得到的预测响应，因此，我们也可以设计一种检验方法，使其检验功效能够近似于基于真实回归变量的 F-检验的检验功效，正如下面的结果所示。

推论2 在定理 6的设定下，假设 $\left\{(\tilde{y}_1^{(K)}, \tilde{y}_2^{(K)}, ...\tilde{y}_s^{(K)})\right\}_{K=1}^{\infty}$ 是基于原始应力嵌入 $\hat{}z_1$，...，$\hat{}z_s$，随机点积图前 s 个节点的预测响应向量序列。定义

$$F^* = (s-2)\frac{\sum_{i=1}^{s}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{s}(y_i - \hat{y}_i)^2}, \qquad \hat{F}^{(K)} = (s-2)\frac{\sum_{i=1}^{s}(\tilde{y}_i^{(K)} - \bar{y})^2}{\sum_{i=1}^{s}(y_i - \tilde{y}_i^{(K)})^2}. \tag{11}$$

在不存在真实回归变量 t_i 的情况下，考虑检验原假设 H_0： $\beta = 0$ 与 H_1： $\beta \neq 0$，决策规则为：若$\hat{}F^{(K)} > c_{\tilde{\alpha}}$，则在显著性水平 $\tilde{\alpha}$ 下拒绝 H_0，转而支持 H_1，其中 $c_{\tilde{\alpha}}$ 是 $F_{1,\ s-2}$ 分布的$(1-\tilde{\alpha})$ 分位数。若以 $\hat{}\pi^{(K)}$ 表示该检验的检验功效，则 $\lim_{K\to\infty}\hat{}\pi^{(K)} = \pi^*$，其中 π^* 是这样一个检验的检验功效，其决策规则为：若 $F^* > c_{\tilde{\alpha}}$，则在显著性水平 $\tilde{\alpha}$ 下拒绝 H_0，转而支持 H_1。

因此，如果在缺少真实回归变量t_i的情况下需要进行模型有效性检验，则可以改为采用基于原始应力嵌入$\hat{}z_i$ 的检验；当图的辅助节点数量足够多时，该检验的检验功效近似相同。

5 模拟

本节将介绍模拟结果，以证明我们的定理得到支持。我们对已知和未知的一维流形上的图进行了模拟，共使用了 100 个蒙特卡洛样本。

5.1 已知流形的情形

我们将流形设为 $\psi([0,1])$，其中 ψ： $[0,1] \to \mathbb{R}^3$ 是哈代-温伯格曲线，由 $\psi(t) = (t^2, 2t(1-t), (1-t)^2)$ 给出。节点数量 n 从 600 到 2500，以 100 为步长变化。对于每个 n，我们在 100 个蒙特卡洛样本上重复以下过程。生成一个样本 $t_1, ...t_n \sim^{iid} U[0,1]$，并通过回归模型 $y_i = \alpha + \beta t_i + \epsilon_i$，$\epsilon_i \sim^{iid} N(0, \sigma_\epsilon^2)$，$i \in [n]$ 对响应 y_i 进行采样，其中 $\alpha = 2.0$，$\beta = 5.0$，$\sigma_\epsilon = 0.1$。一个无向

生成了一个具有潜在位置 $\mathbf{x}_i = \psi(t_i)$ 的空心随机点积图。更具体地说，邻接矩阵 $\mathbf{A}$ 的第 (i, j) 个元素满足 $\mathbf{A}^{ij} \sim$ 伯努利分布$(\mathbf{x}_i^T \mathbf{x}j)$，对所有 $i < j$ 均成立；并且 $\mathbf{A}^{ij} = \mathbf{A}^{ji}$ 对所有 $i, j \in [n]$ 均成立，而 $\mathbf{A}_{ii} = 0$ 对所有 $i \in [n]$ 均成立。我们将真实的潜在位置矩阵记为 $\mathbf{X} = [\mathbf{x}_1|\mathbf{x}_2|...|\mathbf{x}_n]^T$，其邻接谱估计记为 $\hat{}\mathbf{X}$。我们计算的是

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W} \in \mathcal{O}(d)} \left\| \mathbf{X} - \hat{\mathbf{X}}\mathbf{W} \right\|_F$$

$\tilde{}$ $\hat{}\hat{}$ 最后，将 $\mathbf{X} = \mathbf{X}\mathbf{W}$ 设为潜在位置矩阵 $\mathbf{X}$ 的最优旋转的邻接谱估计。于是，我们得到 $\hat{}\ t_i = \arg\min_t \|\tilde{}\mathbf{x}_i - \psi(t)\|$ for $i \in [n]$，并得到 $\hat{}\alpha_{sub}$ 和 $\hat{}\hat{}\hat{}\hat{}\hat{}\beta_{sub}$。令 $\boldsymbol{\theta} = (\alpha, \beta)$、$\boldsymbol{\theta}_{true} = (\hat{}\alpha_{true}, \beta_{true})$ 和 $\boldsymbol{\theta}_{sub} = (\hat{}\alpha_{sub}, \beta_{sub})$，我们计算 $\boldsymbol{\theta}_{true}$ 和 $\boldsymbol{\theta}_{sub}$ 在 100 个蒙特卡洛样本上的样本均方误差（MSE），并将其对照 n 作图。图示见图2。

$\hat{}$**备注 4** 最优旋转矩阵$\mathbf{W}$ 需要根据真实潜在位置矩阵$\mathbf{X}$ 计算，这正是已知流形情形下对回归模型进行推断不切实际的原因，因为$\mathbf{X}$ 通常未知。

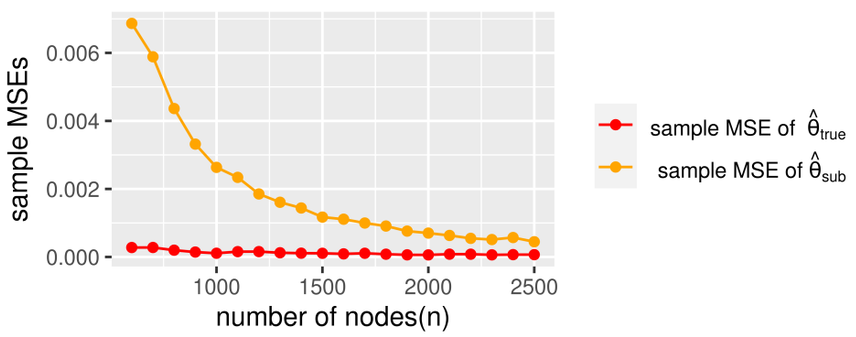

图2 展示了已知流形上回归参数向量的替代估计量的一致性。对于100个蒙特卡洛样本，使用潜在位置的最优旋转的邻接谱估计在流形上的投影计算替代估计，然后计算基于真实回归变量的估计量和替代估计量的样本均方误差。对于中等规模的图（$n \leq 2000$），替代估计量的表现明显差于基于真实回归变量的估计量。然而，随着节点数量增加，估计量之间的性能差异逐渐减小，且两种估计量的均方误差都趋近于零。

5.2 未知流形的情形

我们假设底层的弧长参数化流形为 $\psi([0, 1])$，其中 $\psi : [0, 1] \to \mathbb{R}^4$ 由 $\psi(t) = (t/2, t/2, t/2, t/2)$ 给出。我们将记录响应的节点数设为 $s = 20$。这里，m 表示辅助节点的数量，$n = m + s$ 表示节点总数。我们令 n 取遍集合 $\{500, 750, 1000, ...3000\}$。对于每个 n，我们对 100 个蒙特卡洛样本重复以下过程。生成一个样本 $t_1, ...t_s \sim^{iid} U[0, 1]$，并对所有 $i \in [s]$，根据回归模型 $y_i = \alpha + \beta t_i + \epsilon_i, \epsilon_i \sim^{iid} N(0, \sigma_\epsilon^2)$ 生成响应 y_i，其中 $\alpha = 2.0, \beta = 5.0, \sigma_\epsilon = 0.01$。此外，我们对辅助节点的原像进行抽样，$t_{s+1}, ...t_n \sim^{iid} U[0, 1]$。因此，生成一个潜在位置为 $\mathbf{x}_i = \psi(t_i)$、$i \in [n]$ 的随机点积图，并计算其潜在位置的邻接谱估计 $\hat{}\mathbf{x}_i$。按照以下规则，以邻接谱估计中的前 $n/10$个作为节点构建定位图：当且仅当 $\|\hat{}\mathbf{x}_i - \hat{}\mathbf{x}_j\| < \lambda$ 时，定位图中的两个节点 $\hat{}\mathbf{x}_i$ 和 $\hat{}\mathbf{x}j$ 才由一条边连接，其中$\lambda = 0.8 \times 0.99^{K-1}$ 在 n 为向量 (500, 750, 1000, . . . 3000) 中第 K 项时取值。然后，使用 R 中 smacof 软件包的多维尺度分析函数，通过原始应力最小化将前 $l = s + 1$ 个估计潜在位置的最短路径距离矩阵 $\mathbf{D} = (d_{n,\lambda}(\hat{}\mathbf{x}_i, \hat{}\mathbf{x}j))_{i,j=1}^{l}$ 嵌入一维空间，得到一维嵌入 $\hat{}z_i$。响应 y_i 将

以一维原始应力嵌入 $\hat{z}_i$ 为 $i \in [s]$进行回归，并计算第 ($\tilde{y}_{s+1}$ 个节点的预测响应 $s+1$。还根据真实回归变量 y_{s+1}，得到第 ($s+1$ 个节点的预测响应 $\hat{}\, t_i$ 。由于真实回归变量 t_i的分布相同，每个预测响应的分布不依赖于节点的标签；因此我们省略下标，并用 $\hat{y}_{true}$ 表示基于真实回归变量的预测响应 t_i，用 $\hat{y}_{sub}$ 表示基于原始应力嵌入 $\hat{z}_i$ 的预测响应（因为在预测响应时，原始应力嵌入被用作真实回归变量的替代）。最后，计算所有蒙特卡洛样本中平方距离 ($\hat{y}_{sub} - \hat{y}_{true}$) 的样本均值，并将其绘制为 n的函数。图见 图 3。

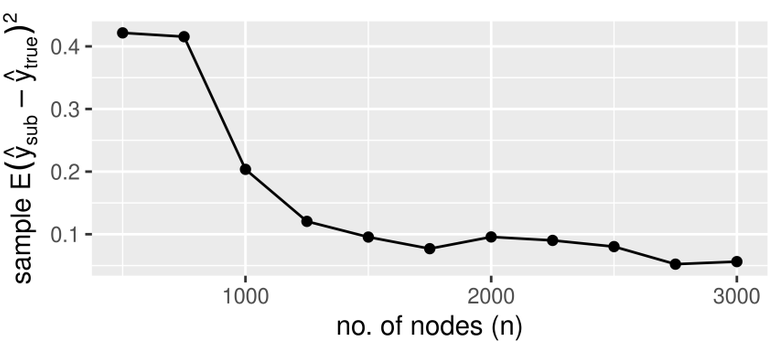

图3 展示基于未知流形上的原始应力嵌入所得预测响应一致性的图。弧长参数化流形取为 $\psi([0,1])$，其中 $\psi(t) = (t/2、\ t/2、\ t/2、\ t/2)$。将基于原始应力嵌入的预测响应$\hat{y}_{sub}$ 与基于真实回归变量的预测响应$\hat{y}_{true}$ 之间平方差的样本均值，对图中节点总数作图。当辅助节点数量适中时（$n \leq 1000$），替代预测响应$\hat{y}_{sub}$ 的表现显著差于 $\hat{}\, y_{true}$ 。然而，随着辅助节点数量进一步增加，$\hat{y}_{true}$ 与$\hat{y}_{sub}$ 之间的平方差趋于零。

接下来，我们重点讨论基于原始应力嵌入 $\hat{z}_i$ 的假设检验问题。为了检验模型的有效性

$$y_i = \alpha + \beta t_i + \epsilon_i$$

其中 $\epsilon_i \sim^{iid} N(0, \sigma_\epsilon^2)$、 $i \in [s]$，人们会进行如下假设检验 H_0： $\beta = 0$，针对H_1 ： $\beta \neq 0$。如果已知真实回归变量 t_i ，我们将使用公式（11）中的检验统计量 F^*，但推论2告诉我们，当辅助潜在位置的数量足够大时，可以基于一维原始应力嵌入 $\hat{z}_i$ 进行检验，其检验功效可近似于基于真实 F-检验统计量 F^* 的检验功效。我们在图4中给出了一幅支持推论2的图。我们展示了：对于所选的一对回归参数值，基于原始应力嵌入的检验功效趋近于基于真实回归变量的检验功效。该设置与前一个设置几乎相同，不同之处在于，此处节点数量 n 以50为步长从100变化到1000。对于每个 n，针对100个蒙特卡洛样本，分别计算真实 F-检验统计量（基于真实回归变量 t_i）和估计的 F-检验统计量（基于原始应力嵌入 $\hat{z}_i$），并通过统计量超过特定阈值的蒙特卡洛样本所占的比例来估计两种检验的检验功效。然后，对于每个 n，计算两种检验的经验估计检验功效之差（一种基于原始应力嵌入，另一种基于真实回归变量），并将其对 n作图。图见图4，其中显示，随着辅助节点数量增加，基于真实回归变量和原始应力嵌入的检验的经验估计检验功效之差趋近于零。

6 应用

在本节中，我们展示了将我们的方法应用于真实世界数据。霍华德·休斯医学研究所珍利亚重建了高阶的完整连接图。

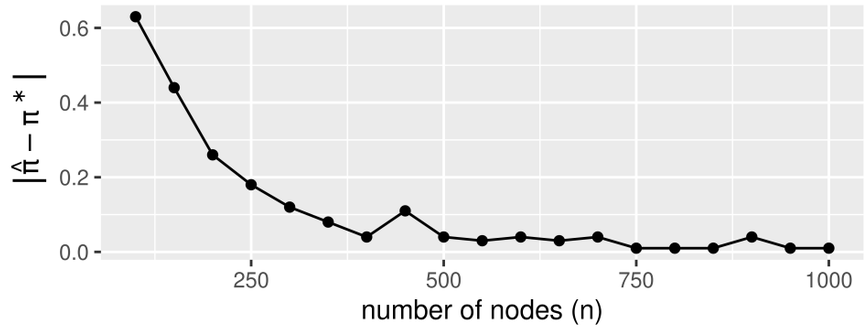

图4 基于一维原始应力嵌入和真实回归变量的模型有效性检验经验功效之差的绘图。弧长参数化流形取为 $\psi([0,1])$，其中 $\psi(t) = (t/2,\ t/2,\ t/2,\ t/2)$。对于固定的小数量（$s = 20$）的节点，响应 y_i 由 $y_i = \alpha + \beta t_i + \epsilon_i$，$\epsilon_i \sim^{iid} N(0,\ \sigma_\epsilon^2)$生成。在$\psi([0,1])$上生成大量（$n - s$）辅助节点，并在与前$n/2$个节点对应的邻接谱估计上构建定位图。当 n 为向量（100，150，200，..... 1000）的第 K项时，邻域参数取为 $\lambda = 0.9 \times 0.99^{K-1}$。通过最小化原始应力准则，将最短路径距离的差异度矩阵嵌入到一维空间。为了检验 $H_0:\ \beta = 0$ 对比 $H_1:\ \beta \neq 0$，比较基于真实回归变量 t_i 的检验统计量 $\hat{}F^*$ 和基于一维等距映射嵌入$\hat{z}_i$ 的检验统计量 F，其中 n 为图中的节点总数。对于每个从 100 到 500、步长为 25 的 n，通过统计 100 个蒙特卡洛样本中检验统计量拒绝 H_0的比例，对相应的检验功效进行经验估计。图中显示，两种检验的估计功效之差趋近于零，这表明当辅助节点数量足够大时，基于等距映射嵌入的检验几乎与基于真实回归变量的检验同样有效。

幼虫中用于联想学习的平行纤维系统 果蝇脑中的蘑菇体。共有 $n = 100$ 个数据点，对应于构成潜在空间中一个网络的 100 个肯扬细胞神经元。肯扬细胞神经元的束入口点与蘑菇体神经网 neuropil 之间的距离（以微米计）被视为该神经元对应的响应。我们进行假设检验，以检验一个简单线性回归模型是否将幼虫果蝇([12],[19],[6])右半球神经元的响应与神经元潜在位置的某种降维版本联系起来。

观测到一个表示由 100 个肯扬细胞神经元组成网络的有向图。由于所研究的图是有向图，因此构造邻接谱嵌入时，会同时考虑邻接矩阵的左右奇异向量。通过将按对应奇异值对角矩阵缩放后的前 3 个左奇异向量，与按对应奇异值对角矩阵缩放后的前 3 个右奇异向量拼接，得到一个 6 维向量，以此估计每个节点的潜在位置。对于每一对分量，我们绘制包含全部 100 个点的二元数据集散点图，从而得到 6×6 一个散点图矩阵，如图5所示。

图5显示，每个分量都与其他所有分量近似相关，这表明六维数据点可能位于一维流形上。我们使用邻域参数 $\lambda = 0.50$ 在潜在位置的 6 维估计值上构造定位图，并将最短路径距离的差异度矩阵$\mathbf{D} = (d_{100,0.5}(\hat{}\mathbf{x}_i, \hat{}\mathbf{x}j))_{i,j=1}^{100}$通过最小化原始应力准则嵌入一维，从而得到一维嵌入 $\hat{}z_i$。图6 展示了响应 y_i 与一维原始应力嵌入 $\hat{}z_i$的关系图。

中的图图 6 表明，简单线性回归模型将响应与原始应力嵌入联系起来。我们希望检验该模型的有效性

$$y_i = a + b\hat{z}_i + \eta_i$$

其中 $\eta_i \sim^{iid} N(0, \sigma_\eta^2),\ i \in [n]$。为此，我们检验 $H_0:\ b = 0$对 $H_1:\ b \neq 0$，显著性水平为 0。01。发现 F统计量的值为 9，自由度为 1 和 98。815。这得出 p值 $= P[F_{1,98} > 9.815] = 0.0023$，低于我们的显著性水平 0。01。因此，我们得出结论：存在一个简单线性回归模型，其中 y_i作为因变量的取值，$\hat{}z_i$ 作为自变量的取值。根据推论 2，我们得出结论，神经元上的响应与标量原像相关联

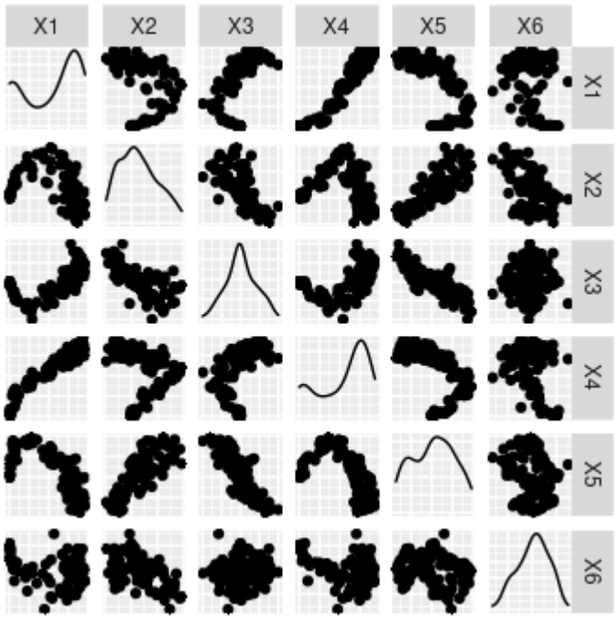

图5 散点图矩阵显示了幼年果蝇100个肯扬细胞神经元网络中潜在的低维结构。考虑一个有向图，其中每个节点代表一个神经元。通过将按相应奇异值缩放的3个主导左奇异向量与按相应奇异值缩放的3个主导右奇异向量结合，为每个节点获得一个6维邻接谱估计。随后针对这6个分量中的每一对生成一个散点图。由于每个维度似乎都通过某个函数与另一个维度近似相关，因此假设存在潜在的1维结构。

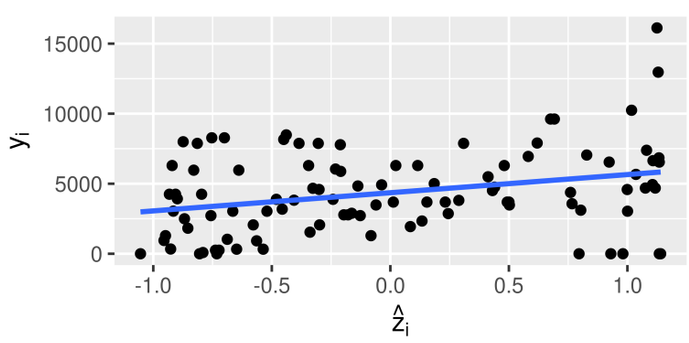

图6 散点图表明，响应与一维原始应力嵌入通过简单线性回归模型相关联。根据幼年果蝇中构成有向网络的100个肯扬细胞神经元对应的潜在位置的六维估计值，通过对最短路径距离进行原始应力最小化，获得一维嵌入$\hat{z}_i$。第 i 个神经元的束入口点与蘑菇体神经网 neuropil 之间的距离（y_i）被视为对应于第 i 个神经元的响应。带有拟合回归线 $y = 4356.1 + 1296.6x$ 的（y_i, $\hat{z}_i$）散点图表明存在显著效应（$p < 0.01$，针对 H_0： $a = 0$，而针对 H_1：$a \neq 0$，在 $y_i = a + b\hat{\varepsilon}_i + \eta_i$ 中）。

潜在位置的响应通过简单线性回归模型相联系。此外，如果未记录某些肯扬细胞神经元的束入口点与蘑菇体神经网 neuropil 之间的距离，则可使用一维原始应力嵌入 $\hat{z}_i$作为真实回归变量的替代指标来预测这些值。

7 结论

本文针对随机点积图的潜在位置位于高维环境空间中的一维流形上的模型，推导了理论结果并给出了数值结果。我们证明，对于已知流形，即使真实回归变量未知，也能相合地估计简单线性回归模型的参数；该模型将图中每个节点记录的响应变量与该节点潜在位置的标量原像联系起来。然而，我们的一项关键结果表明，即使流形未知（这也是更现实的情形），在

有利条件下，仍能以较好的效果学习流形，从而获得与基于真实回归变量得到的预测响应相近的预测响应。

我们利用原始应力嵌入（[11]）的收敛保证，在底层流形未知时获得回归变量点间距离的一致估计量。我们证明，随着辅助潜在位置的数量趋于无穷，对于每个辅助节点，基于原始应力嵌入得到的预测响应会趋近于基于真实回归变量得到的预测响应。

请注意，尽管回归参数的替代估计量（或基于原始应力嵌入得到的预测响应）的渐近表现可以接近基于真实回归变量得到的相应结果，但在现实情形中，我们可能面对小样本，此时替代估计量（或预测响应）的表现可能较差。当底层流形已知时，我们可以考虑测量误差（即估计回归变量与真实回归变量之间的差异），从而对回归参数的估计量进行调整（[13]）。我们在一个已知流形上进行模拟，比较基于真实回归变量的估计量、替代（或朴素）估计量以及经测量误差调整的估计量的表现。对于回归模型 $y_i = \beta t_i + \epsilon_i, \epsilon_i \sim^{iid} N(0, \sigma_\epsilon^2)$，其中回归变量 t_i 由 $\hat{t}_i$ 估计，经测量误差调整的估计量为 $\hat{\beta}_{adj,\sigma} = \frac{\sum_{i=1}^n y_i \hat{t}_i}{\sum_{i=1}^n \hat{t}_i^2 - \sum_{i=1}^n \Gamma_i}$，其中 $\Gamma_i = var(\hat{t}_i - t_i)$。在大多数现实情形中，真实值 Γ_i 并不可知。然而，如果它们存在相合估计量 $\hat{\Gamma}_i$，则我们可以使用代理量 $\hat{\beta}_{adj,\hat{\sigma}} = \frac{\sum_{i=1}^n y_i \hat{t}_i}{\sum_{i=1}^n \hat{t}_i^2 - \sum_{i=1}^n \hat{\Gamma}_i}$。为了比较这些估计量的表现，我们对随机点积图进行抽样，其节点位于高维环境空间中的一维曲线上。我们计算潜在位置的邻接谱估计，将其投影到流形上，并得到回归变量的估计值 $\hat{}\hat{}\hat{}$，随后用这些估计值计算 $\beta_{true}, \beta_{naive}, \beta_{adj,\sigma}$ 和 $\beta_{adj,\hat{\sigma}}$ 在 100 个蒙特卡洛样本中的取值。图 7 展示了基于 100 个蒙特卡洛样本计算得到的这些估计量取值的箱线图。

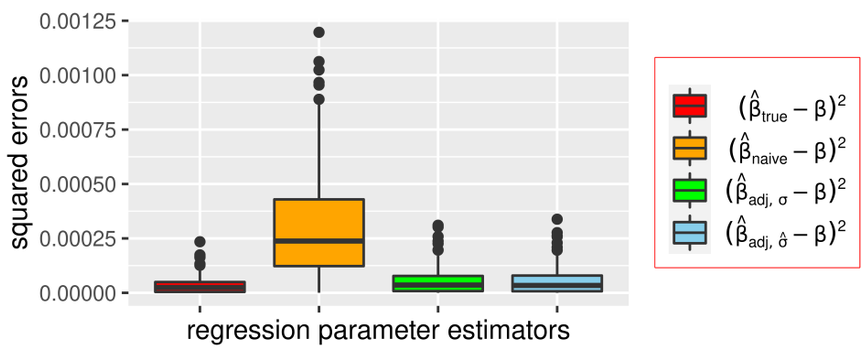

图 7 回归模型的截距项为零时，4 个回归斜率参数估计量的平方误差箱线图。对于 100 个蒙特卡洛样本中的每一个，均生成一个包含 $n = 800$ 个节点的随机图，其中第 i 个节点的潜在位置为 $\mathbf{x}_i = (t_i^2, 2t_i(1-t_i), (1-t_i)^2)$，其中 $t_i \sim^{iid} U[0,1]$。在第 i 个节点处，通过回归模型 $y_i = \beta t_i + \epsilon_i$, $\epsilon_i \sim^{iid} N(0, \sigma_\epsilon^2)$ 生成响应 y_i，其中 $\beta = 5.0$, $\sigma_\epsilon = 0.1$。朴素估计量通过将潜在位置的最优旋转邻接谱估计的投影的原像代入计算得出。为计算 $\hat{\beta}_{adj,\sigma}$，我们代入另一组蒙特卡洛样本所得到的样本方差之和；这些样本中的图均由同一模型生成。我们对潜在位置的最优旋转邻接谱估计的渐近方差（见 2）使用德尔塔方法，得到 $\sum_{i=1}^n \hat{\Gamma}_i$，进而计算 $\hat{\beta}_{adj,\hat{\sigma}}$。

$\hat{}\hat{}$图 7清楚地表明，测量误差校正估计量 $\beta_{adj,\sigma}$ 和 $\beta_{adj,\hat{\sigma}}$ 的表现优于朴素估计量 β_{naive}。此外，显然，$\hat{}\hat{}\beta_{adj,\sigma}$ 和 $\beta_{adj,\hat{\sigma}}$ 的表现差异并不显著。这使我们确信，使用测量误差校正估计量 $\beta_{adj,\hat{\sigma}}$ 是务实且有效的，因为在许多现实情形中，由于缺乏对 Γ_i 真实值的了解，无法计算 $\beta_{adj,\sigma}$。

遗憾的是，对于未知流形，无法轻易应用同样的方法，因为正如 第4节中所讨论的那样，保留下来的只有点间距离，而非嵌入。我们认为，这是一个值得在未来探讨的有趣问题。未来研究的另一个方向是，探究一维流形的结果能否推广到 k 维流形，其中 $k > 1$。

数据和材料的可用性
本研究中分析的数据集已收录于已发表的文章中 [12]。

利益冲突
作者声明不存在任何利益冲突。

作者贡献
AA 制定了理论，开展了实验并撰写了手稿。JA制定了理论并编辑了手稿。MWT 制定了理论并编辑了手稿。YP 开展了实验并编辑了手稿。CEP 提出了问题，制定了理论并编辑了手稿。

致谢
本研究部分得到约翰斯·霍普金斯数学数据科学研究所（MINDS）研究员奖学金的资助。

作者详情[1]

应用数学与统计学系，约翰斯·霍普金斯大学，美国巴尔的摩。 [2]数据工程与科学研究所、电气与系统工程系及统计与数据科学系，宾夕法尼亚大学，美国宾夕法尼亚州。 [3]统计学系，印第安纳大学，美国布卢明顿。 [4]成像科学中心，约翰斯·霍普金斯大学，美国巴尔的摩。

参考文献.

1 P. L. Erdos R´enyi, A.：关于随机图的演化。美国数学会会刊 286, 257–257 (1984) 2. Goldenberg, A., Zheng, A.X., Fienberg, S.E., Airoldi, E.M., 等：统计网络模型综述。机器学习基础与趋势 2(2), 129–233 (2010) 3. Hoff, P.D., Raftery, A.E., Handcock, M.S.：社会网络分析的潜在空间方法。美国统计学会杂志 97, 1090–1098 (2002) 4. Young, S.J., Scheinerman, E.R.：社会网络的随机点积图模型。载于：网络图算法与模型研讨会 (2007) 5. Rubin-Delanchy, P., Priebe, C.E., Tang, M., Cape, J.：谱嵌入的统计解释：广义随机点积图。英国皇家统计学会杂志：B辑（统计方法学）84, 1446–1473 (2022) 6. Athreya, A., Fishkind, D.E., Tang, M., Priebe, C.E., Park, Y., Vogelstein, J.T., Levin, K., Lyzinski, V., Qin, Y.：随机点积图的统计推断：综述。机器学习研究杂志 18(1), 8393–8484 (2017) 7. Rubin-Delanchy, P.：图嵌入中的流形结构。神经信息处理系统进展 33, 11687–11699 (2020) 8. Belkin, M., Matveeva, I., Niyogi, P.：大规模图上的 Tikhonov 正则化与半监督学习。载于：2004 IEEE声学、语音与信号处理国际会议，第3卷，第1000页 (2004)。doi:10 1326716 9 1109 2004 . /ICASSP. . . Belkin, M., Niyogi, P.：用于降维和数据表示的拉普拉斯特征映射。神经计算 15(6), 1373 089976603321780317 10 1396 2003 101162 – ( )。doi: . / . Athreya, A., Tang, M., Park, Y., Priebe, C.E.：潜在结构随机图中的估计与推断。统计科学 36(1), 68–88 (2021) 11. Trosset, M.W., Gao, M., Tang, M., Priebe, C.E.：学习一维子流形，以便随后对随机点积图进行推断。arXiv预印本 arXiv:2004.07348 (2020) 12. Eichler, K., Li, F., Litwin-Kumar, A., Park, Y., Andrade, I.V., Schneider-Mizell, C.M., Saumweber, T., Huser, A., Eschbach, C., Gerber, B., Fetter, R.D., Truman, J.W., Priebe, C.E., Abbott, L.F., Thum, A.S., Zlatic, M., Cardona, A.：昆虫脑中学习与记忆中枢的完整连接组。《自然》548, 175–182 (2017) 13. Fuller, W.A.：测量误差模型 (1987) 14. Whiteley, N., Gray, A., Rubin-Delanchy, P.：发现数据中的潜在拓扑与几何：大维数定律。arXiv预印本 arXiv:2208.11665 (2022) 15. Tenenbaum, J.B., De Silva, V., Langford, J.C.：用于非线性降维的全局几何框架。《科学》290(5500), 2319–2323 (2000) 16. Bernstein, M., De Silva, V., Langford, J.C., Tenenbaum, J.B.：嵌入流形上测地线的图近似。技术报告，Citeseer (2000) 17 10858 18 2021 2006 2005 . Trosset, M.W., Buyukbas, G.：重振 Isomap：测地结构的欧几里得表示 ( )。 . . Borg, I., Groenen, P.J.：现代多维尺度分析：理论与应用。施普林格，??? ( ) 19. Priebe, C.E., Park, Y., Tang, M., Athreya, A., Lyzinski, V., Vogelstein, J.T., Qin, Y., Cocanougher, B., Eichler, K., Zlatic, M., Cardona, A.：果蝇连接组的半参数谱建模。arXiv (2017)。doi:10.48550/ARXIV.1705.03297.20 https://arxiv.org/abs/1705.03297 . Alyakin, A.A., Agterberg, J., Helm, H.S., Priebe, C.E.：针对顶点数量不同的图，校正非参数双样本图假设检验 (2020 09434 21 2008)。 . . Athreya, A., Priebe, C., Tang, M., Lyzinski, V., Marchette, D., Sussman, D.：随机点积图缩放特征向量的极限定理。Sankhya A 1(78), 1–18 (2016) 22. Cape, J., Tang, M., Priebe, C.E.：二到无穷范数与奇异子空间几何及其在高维统计中的应用。统计年刊 (2019) 23. Tang, M., Athreya, A., Sussman, D.L., Lyzinski, V., Priebe, C.E.：随机点积图的非参数双样本假设检验问题。arXiv (2014 10 48550 1409 2344 https://arxiv.org/abs/1409.2344)。doi: . / 24 2019 ARXIV. . . . Trosset, M.W., Priebe, C.E.：统计子流形上的近似信息检验。arXiv ( )。doi: 10 48550 1903 08656 https://arxiv.org/abs/1903.08656 . /ARXIV. . . . Sekhon, S.：可测函数双重序列收敛性的 25 一条结果 (2021)。2104.09819

8 附录

定理*4*：设 $\psi:\ [0,L]\to\mathbb{R}^d$ 是双射，且其逆映射 γ 满足 $\|\nabla\gamma(w)\|<K$，对所有 $w\in\psi([0,L])$均成立，其中 $K>0$为某个值。令 $x_i=\psi(t_i)$ 为随机点积图中第 i 个节点的潜在位置，该图有 n 个节点，并假设 $y_i=\alpha+\beta t_i+\epsilon_i$, $\epsilon_i\sim^{iid}N(0,\sigma_\epsilon^2)$ 对所有 $i\in[n]$ 均成立。假设对所有 i，$x_i\sim^{iid}F$，其中 F 是 $\mathbb{R}^d$上的内积分布。设$\tilde{}x_i$ 是对所有 i 的 x_i 的最优旋转邻接谱估计，并令 $\hat{}t_i=arg\,min_t\,\|\tilde{}x_i-\psi(t)\|$。那么，当 $n\to\infty$ 时，$\hat{}\alpha_{sub}\to^P\alpha$, $\hat{}\beta_{sub}\to^P\beta$。

证明：对所有 $i\in[n]$，令 $\mathbf{u}_i=\tilde{}\mathbf{x}_i-\mathbf{x}_i$，并注意到根据定理 *1*，$\max_i\|\mathbf{u}_i\|\to^P0$ 当 $n\to\infty$ 时。令 $\mathbf{u}_i=\tilde{}\mathbf{x}_i-\mathbf{x}_i$，并令 $\mathbf{h}_i$ 为满足 $\tilde{}\mathbf{x}_i+\mathbf{h}_i\in\psi([0,L])$ 的最短向量。注意，对所有 i，$\|\mathbf{h}_i\|\le\|\mathbf{u}_i\|$。
令 $\mathbf{q}_i=\mathbf{h}_i+\mathbf{u}_i$，并利用泰勒定理，我们可以观察到对所有 i,

$$\hat{t}_i=\gamma(\mathbf{x}_i+\mathbf{q}_i)=t_i+\mathbf{q}_i^T\nabla\gamma(\mathbf{x}_i)+o(\|\mathbf{q}_i\|).$$

因此，根据柯西-施瓦茨不等式，

$$|\hat{t}_i-t_i|\le\|\mathbf{q}_i\|\,\|\nabla\gamma(\mathbf{x}_i)\|+o(\|\mathbf{q}_i\|)\le K\|\mathbf{q}_i\|+o(\|\mathbf{q}_i\|).$$

注意，根据三角不等式，$\|\mathbf{q}_i\|\le2\|\mathbf{u}_i\|$，因此 $\max_i\|\mathbf{q}_i\|\to^P0$ 当 $n\to\infty$ 时成立，这意味着 $\max_i|\hat{}t_i-t_i|\to^P0$ 当 $n\to\infty$ 时成立。
回顾一下，基于真实回归变量值的回归参数估计量是

$$\hat{\beta}_{true}=\frac{\frac{1}{n}\sum_{i=1}^n(y_i-\bar{y})(t_i-\bar{t})}{\frac{1}{n}\sum_{i=1}^n(t_i-\bar{t})^2},\qquad\hat{\alpha}_{true}=\bar{y}-\hat{\beta}_{true}\bar{t},$$

而替代估计量或代入估计量是

$$\hat{\beta}_{sub}=\frac{\frac{1}{n}\sum_{i=1}^n(y_i-\bar{y})(\hat{t}_i-\bar{\hat{t}})}{\frac{1}{n}\sum_{i=1}^n(\hat{t}_i-\bar{\hat{t}})^2},\qquad\hat{\alpha}_{sub}=\bar{y}-\hat{\beta}_{sub}\bar{\hat{t}}.$$

注意，根据三角不等式，当 $n\to\infty$ 时，$max_i|\hat{}t_i-t_i|\to^P0$ 意味着 $|\bar{t}-\hat{}t|\to^P0$，$|\frac{1}{n}\sum_{i=1}^n(t_i-\bar{t})^2-\frac{1}{n}\sum_{i=1}^n(\hat{}t_i-\hat{}\bar{t})^2|\to^P0$ 且 $|\frac{1}{n}\sum_{i=1}^n y_i(t_i-\bar{t})-\frac{1}{n}\sum_{i=1}^n y_i(\hat{}t_i-\hat{}t)|\to^P0$。因此，当 $n\to\infty$ 时，$|\hat{}\beta_{sub}-\hat{}\beta_{true}|\to^P0$ 且 $|\hat{}\alpha_{sub}-\hat{}\alpha_{true}|\to^P0$。回想一下，$\hat{}\alpha_{true}$ 和 $\hat{}\beta_{true}$ 分别是 α 和 β 的$\hat{}$相合估计量，我们得出结论，$\hat{}\alpha_{sub}$ 和 β_{sub} 也分别是 α 和 β的相合估计量。

推论*1*：在定理*4*的设定下，以真实回归变量 t_i 为条件，以下两个条件成立：$\hat{}$*(A)* $\mathbb{E}(\hat{}\alpha_{sub})\to\alpha$, $\mathbb{E}(\beta_{sub})\to\beta$，当 $n\to\infty$. 时，$\tilde{}$*(B)* 对任意两个线性无偏估计量 $\tilde{\alpha}$ 和 β 以及任意 $\delta>0$，$\hat{}\tilde{}var(\hat{}\alpha_{sub})\le var(\tilde{\alpha})+\delta$，$var(\beta_{sub})\le var(\beta)+\delta$，当 n 足够大时。

$\hat{}$证明：由定理*4*可直接得出，当 $n\to\infty$ 时，$E(\hat{}\alpha_{sub})\to\alpha$, $E(\beta_{sub})\to\beta$。此外，注意到 $(\hat{}\alpha_{sub}-\hat{}\alpha_{true})\to^P0$ 且 $(\hat{}\beta_{sub}-\hat{}\beta_{true})\to^P0$，当 $n\to\infty$ 时。因此，对任意 $\hat{}\hat{}\delta>0$, $var(\hat{}\alpha_{sub})\le var(\hat{}\alpha_{true})+\delta$, $var(\beta_{sub})\le var(\beta_{true})+\delta$，当 n 足够大时。回想一下，$\hat{}\alpha_{true}$ 和 β_{true} 分别是 α 和 β 的最佳线性无偏估计量，*(B)* 即得。

定理 *6*：考虑一个随机点积图，其中每个节点都位于弧长参数化的一维流形 $\psi([0,L])$ 上，且 ψ 未知。令 $x_i=\psi(t_i)$ 表示所有 i 中第 i 个节点的潜在位置。假设 $y_i=\alpha+\beta t_i+\epsilon_i$, $\epsilon_i\sim^{iid}N(0,\sigma_\epsilon^2)$ 对于 $i\in[s]$ 成立，其中 s 是一个固定整数。基于真实回归变量 $\hat{}$，第 r 个节点的预测响应为 $\hat{}y_r=\hat{}\alpha_{true}+\beta_{true}t_r$。存在节点数量序列 $n_K\to\infty$ 和邻域参数序列 $\lambda_K\to0$，使得对每个 $r>s$，$|\hat{}y_r-\tilde{y}_r^{(K)}|\to^P0$ 在 $K\to\infty$ 时成立，其中$\tilde{y}_r^{(K)}=PRED(A^{(K)},d,\lambda_K,l,\{y_i\}_{i=1}^s,r)$（见算法 *1b*），$A^{(K)}$ 表示节点数量为 n_K 时的邻接矩阵，且 l 是满足 $l>r>s$ 的固定自然数。

证明：固定 $l \in \mathbb{N}$ 使得 $s < r \leq l$。对于每个 $K \in \mathbb{N}$，选取节点数量 n_K 作为观测值，并选择适当的 λ_K，使得等式*9* 成立；并回顾等式*10* 可知，$(\hat{}z_1^{(K)}, ...\hat{}z_l^{(K)})$是原始应力准则的极小值点：

$$\sigma_l(z_1, ...z_l) = \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(|z_i - z_j| - d_{n_K,\lambda_K}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j))^2$$

根据定理 *5*，我们知道，对于所有 i、$j \in [l]$，当 $K \to \infty$ 时，

$$(|\hat{z}_i^{(K)} - \hat{z}_j^{(K)}| - |t_i - t_j|) \to^P 0. \tag{12}$$

定义

$$\tilde{\beta}^{(K)} = \frac{\sum_{i=1}^{s}(y_i - \bar{y})(\hat{z}_i^{(K)} - \bar{\hat{z}}^{(K)})}{\sum_{i=1}^{s}(\hat{z}_i^{(K)} - \bar{\hat{z}}^{(K)})^2}, \qquad \tilde{\alpha}^{(K)} = \bar{y} - \tilde{\beta}^{(K)}\bar{\hat{z}}^{(K)} \tag{13}$$

其中 $\bar{\hat{z}}^{(K)} = \frac{1}{s}\sum_{i=1}^{s}\hat{}\hat{}z_i^{(K)}$。然后，我们可以基于$\hat{}z_i^{(K)}$'定义 y_r 的预测量为

$$\tilde{y}_r^{(K)} = \tilde{\alpha}^{(K)} + \tilde{\beta}^{(K)}\hat{z}_r^{(K)} = \bar{y} + \tilde{\beta}^{(K)}(\hat{z}_r^{(K)} - \bar{\hat{z}}^{(K)}).$$

如果已知原始的 t_i，则 y_r 的预测量为

$$\hat{y}_r = \hat{\alpha} + \hat{\beta}t_r = \bar{y} + \hat{\beta}(t_r - \bar{t}).$$

现在，回想一下，对于所有 $i, j \in [l]$，$(|\hat{}z_i^{(K)} - \hat{}z_j^{(K)}| - |t_i - t_j|) \to^P 0$ 当 $K \to \infty$ 时成立。因此，对于任意 $\tau > 0$ 和 $\nu > 0$，存在 $K_0 \in \mathbb{N}$，使得对于所有 $K \geq K_0$，以至少 $(1 - \nu)$ 的概率，对所有 $i \in [l]$，

$$|a^{(K)}t_i + b^{(K)} - \hat{z}_i^{(K)}| \leq \tau \implies |a^{(K)}(t_i - \bar{t}) - (\hat{z}_i^{(K)} - \bar{\hat{z}}^{(K)})| \leq 2\tau \tag{14}$$

其中 $a^{(K)} \in \{-1, +1\}$, $b^{(K)} \in \mathbb{R}$。注意，$(a^{(K)})^2 = 1$ 对所有 K 均成立。因此，取足够大的 K，我们可以以任意高的概率使 $\sum_{i=1}^{s} y_i(t_i - \bar{t})(t_r - \bar{t})$ 与 $\sum_{i=1}^{s} y_i(\hat{}z_i^{(K)} - \bar{}\hat{}z^{(K)})(\hat{}z_r^{(K)} - \bar{}\hat{}z^{(K)})$ 任意接近。通过选择足够大的 K，我们也可以以任意高的概率使 $\sum_{i=1}^{s}(\hat{}z_i^{(K)} - \bar{}\hat{}z^{(K)})^2$ 与 $\sum_{i=1}^{s}(t_i - \bar{t})^2$ 任意接近。回想一下，

$$\begin{aligned}\tilde{y}_r^{(K)} &= \bar{y} + \frac{\sum_{i=1}^{s} y_i(\hat{z}_i^{(K)} - \bar{\hat{z}}^{(K)})(\hat{z}_r^{(K)} - \bar{\hat{z}}^{(K)})}{\sum_{i=1}^{s}(\hat{z}_i^{(K)} - \bar{\hat{z}}^{(K)})^2}, \\ \hat{y}_r &= \bar{y} + \frac{\sum_{i=1}^{s} y_i(t_i - \bar{t})(t_r - \bar{t})}{\sum_{i=1}^{s}(t_i - \bar{t})^2}.\end{aligned} \tag{15}$$

因此，通过选择足够大的 K，我们可以以任意高的概率使$\hat{}y_r$ 与 $\tilde{y}_r^{(K)}$任意接近，这意味着当 $K \to \infty$ 时，$|\tilde{y}_r^{(K)} - \hat{}y_r| \to^P 0$。

推论*2*：在定理 *6*的设定下，假设 $\left\{(\tilde{y}_1^{(K)}, \tilde{y}_2^{(K)}, ...\tilde{y}_s^{(K)})\right\}_{K=1}^{\infty}$ 是基于随机点积图的原始应力嵌入，在前 s 个节点处得到的预测响应向量序列 $\hat{}z_1$,..., $\hat{}z_s$。定义

$$F^* = (s-2)\frac{\sum_{i=1}^{s}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{s}(y_i - \hat{y}_i)^2}, \qquad \hat{F}^{(K)} = (s-2)\frac{\sum_{i=1}^{s}(\tilde{y}_i^{(K)} - \bar{y})^2}{\sum_{i=1}^{s}(y_i - \tilde{y}_i^{(K)})^2}. \tag{16}$$

在缺少真实回归变量 t_i 的情况下，考虑检验原假设 H_0： $\beta = 0$ 对立假设 H_1： $\beta \neq 0$，决策规则如下：如果$\hat{}F^{(K)} > c_{\tilde{\alpha}}$，则在显著性水平 $\tilde{\alpha}$ 下拒绝 H_0，支持 H_1；其中 $c_{\tilde{\alpha}}$ 是 $F_{1,s-2}$ 分布的第$(1 - \tilde{\alpha})$ 个分位数。如果将此检验的检验功效记为$\hat{}\pi^{(K)}$，则$\lim_{K\to\infty}\hat{}\pi^{(K)} = \pi^*$，其中 π^* 是如下检验的检验功效：该检验的决策规则是在显著性水平 $\tilde{\alpha}$ 下，当 $F^* > c_{\tilde{\alpha}}$ 时拒绝 H_0，支持 H_1。

证明：根据定理 *5*，我们有 最大值$_{i\in[s]}|\tilde{y}_i^{(K)} - \hat{}y_i| \to^P 0$，这意味着 $(\hat{}F^{(K)} - F^*) \to^P 0$当 $K \to \infty$ 时成立。因此，对于任意 $(\alpha, \beta) \in \mathbb{R}^2$，当 $K \to \infty$, $P_{\alpha,\beta}[\hat{}F^{(K)} > c_{\tilde{\alpha}}] \to P_{\alpha,\beta}[F^* > c_{\tilde{\alpha}}]$ 时成立。