毕业设计（论文）题目：数据的乘积空间模型与高维持续同调的计算

学　　院：数学与统计学院

专　　业：数学与应用数学

班　　级：17111401

姓　　名：曹越琦

指导教师：孙华飞

摘要

本文提出一种对数据进行乘积空间建模并计算其持续同调的方法。首先利用图贝叶斯理论计算图的后验概率，推断数据底空间的独立分量。然后对数据集进行投影并计算各分量空间的持续同调。最后利用 Künneth 定理推断出全空间的同调。

关键词：持续同调，图模型， Künneth 定理，蒙特卡洛马尔科夫链

Abstract

In this paper we propose a method to compute the high dimensional persistence of data sets using Künneth formula. First we view each point as an observation of a random vector and apply Bayesian inference to its graphical model. We introduce a prior on the space of disconnected graphs and compute the posterior probability using Markov chain Monte Carlo algorithm. Then we project the points onto its independent coordinates according to the graph and compute the persistence respectively. Finally the persistence of original data set is obtained by applying Künneth formula to the factor spaces.

Key words : persistent homology, graphical modeling, Künneth theorem, Monte Carlo Markov chain

目录

1 引言

1.1 问题引入

拓扑数据分析 为现代数据科学提供了一种全新的看待数据的观点。它旨在寻找大数据集中的整体特征。以线性代数为主要工具的传统方法在数据集具有较强的线性结构时效果良好。当数据集的结构高度非线性时（例如数据取自某个非线性流形），传统的方法只能用来发现其线性的局部，而无法探究数据集的整体性质。拓扑数据分析的主要思想在于收集数据的局部信息，并构造包含全局信息的几何对象。这样一个联系局部和整体的桥梁，在经典的代数拓扑理论中称为神经。神经是一个抽象单纯复形，它的单纯同调是一个同胚不变量，从而代表了数据的全局特征。

尽管拓扑数据分析为数据科学引入了新的技术，其本身仍有一些重要问题未解决。一个典型的问题就是随着数据集的数量增多，维数增高，单形的数量会呈指数增加，超出计算机的存储能力。目前持续同调的计算停留在0维，1维和2维。对于更高维的持续同调的计算，没有高效的方法。因此为一个数据集给出低维模型是方便的，而为数据集给出高维的忠实的模型则存在着困难。应该发展新的技术去简化数据的高维持续同调的计算。

一个自然的想法是在计算持续同调之前设法降低数据集的维数。如果已经知道数据集的内蕴空间是一个乘积空间，那么这个想法是容易实施的：对数据集进行投影，分别计算各个低维分量空间的持续同调，最后根据代数拓扑中的理论推断全空间的同调。因此问题化归为如何判别数据集来自乘积空间。在统计学中，这等价于判断数据各坐标之间的相关性。假设数据点独立地取自于一个随机向量$\mathbf{X} = (X_1, X_2, \cdots, X_n)$。如果这些随机变量$X_i$能够分解成独立的组，换言之，如果随机向量的概率密度函数有如下分解

$$f(x_1, x_2, \cdots, x_n) = f(x_1, \cdots, x_k) f(x_{k+1}, \cdots, x_n) \tag{1.1}$$

那么这些数据点就取自于两个独立的空间。为了推断随机变量之间的相关性，我们转向图模型（或概率图模型）理论。在这个理论中我们考虑一个更一般的关系，称为随机变量间的条件独立性；用这个关系构造一个无向图，它的每个顶点代表一个随机变量，每条边代表相邻的两个随机变量不是条件独立的。关于随机变量之间的相关性的推断化归为图的推断。而图的推断在许多统计学家的发展下，已有许多优秀成熟的算法。本文在第4节介绍一个由S.Lunagómez和S.Mukherjee等[21]发展的最新的算法，并对此算法进行改进。将改进的算法与之前的讨论相结合，便能对具有特定相关性结构的数据计算高维持续同调。

1.2 主要贡献

- 提出一种新的计算数据集高维持续同调的方法。这个方法联系了两个不同的理论，持续同调理论和图模型理论；
- 通过研究不连通图空间的性质，改进了[21]中计算图的后验概率的算法；
- 将本文算法应用于自然图像统计中，在[5]工作之上，进一步发现高密度高对比度光学图像块的内蕴空间具有乘积空间结构，而其非平凡同调均来自于第一个分量空间；
- 给出了域上的Künneth定理的一个初等证明。

1.3 各节内容

第2节介绍持续同调和图模型理论的一些基本概念。第3节介绍任意域上的Künneth定理并给出一个初等的线性代数的证明。第4节介绍[21]中发展的蒙特卡洛马尔科夫链算法（MCMC算法），证明不连通图空间的一个渐进性质，根据这个性质提出新的MCMC算法。第5节将算法应用于构造的数据和真实的数据。通过分析将算法应用于光学图像数据上的结果，为数据建立一个乘积空间模型。最后，在第6节总结本文，并对之后的工作进行展望。

2 预备知识

本节介绍持续同调与图模型的一些基本概念。方便起见，单纯同调中的链群，同调群等均以模2剩余类群$\mathbb{Z}_2$为系数，从而是$\mathbb{Z}_2$上的向量空间。所有的单纯复形都假设有限。本节最后简要介绍了图贝叶斯的最新进展。

2.1 持续同调

假设$a_0, a_1, \cdots, a_n$是$\mathbb{R}^N$中处于一般位置的点，也即n个向量$a_1-a_0, a_2-a_0, \cdots, a_n-a_0$线性无关。以$a_0, a_1, \cdots, a_n$为顶点的单纯形$\sigma = a_0a_1\cdots a_n$是这些顶点的凸包：$\{v \in \mathbb{R}^N | v = x^0a_0 + x^1a_1 + \cdots + x^na_n, \sum x^i = 1, x^i \geq 0\}$。$\sigma$的面是由$\{a_0, a_1, \cdots, a_n\}$的子集生成的单纯形。一个单纯复形$K$是一系列单纯形的集合，它满足：

1. 集合中任意两个单纯形的交或为空集，或为两个单纯形的公共面；
2. 如果单纯形τ在K中，它的所有面也在K中。

K的p链向量空间是由K中所有的p维单纯形在$\mathbb{Z}_2$上张成的向量空间，记为$C_p(K;\mathbb{Z}_2)$。定义向量空间同态$\partial_p : C_p(K;\mathbb{Z}_2) \to C_{p-1}(K;\mathbb{Z}_2)$，它在基上的作用为

$$\partial_p[a_0, a_1, \cdots, a_p] = \sum_i [a_0, \cdots, \hat{a_i}, \cdots, a_p] \tag{2.1}$$

∂_p一般被称为p维边缘同态算子。令$\mathscr{F}(\sigma)$记单纯形σ的所有$p-1$维面的集合。公式(2.1)等价于

$$\partial_p \sigma = \sum_{\sigma^i \in \mathscr{F}(\sigma)} \sigma^i \tag{2.2}$$

几何上∂_p把σ映为其边界$Bd\sigma = \cup\mathscr{F}(\sigma)$。容易验证$\partial_{p+1} \circ \partial_p = 0$。因此，$\partial_{p+1}$的像是$\partial_p$的核的子空间。$\partial_{p+1}$的像被称为$p$维边缘向量空间，记为$B_p(K;\mathbb{Z}_2)$。$\partial_p$的核被称为$p$维闭链向量空间，记为$Z_p(K;\mathbb{Z}_2)$。$B_p(K;\mathbb{Z}_2)$和$Z_p(K;\mathbb{Z}_2)$中的元素分别称为$p$维边缘和$p$维闭链。$p$维单纯同调向量空间$H_p(K;\mathbb{Z}_2)$定义为商空间$Z_p(K;\mathbb{Z}_2)/B_p(K;\mathbb{Z}_2)$。$H_p(K;\mathbb{Z}_2)$的维数被称为$p$维贝蒂数。

单纯复形K的底空间定义为K中所有元素的并。单纯同调向量空间是拓扑的不变量：假设K和L为单纯复形，g是将K的底空间映为L的底空间的同胚，那么g诱导了各维同调向量空间的同构$g_* : H_i(K;\mathbb{Z}_2) \to H_i(L;\mathbb{Z}_2)$。因此，一个单纯复形的各维贝蒂数蕴含了其底空间的重要的拓扑信息。

一个常用的用来构造单纯复形的方法是利用覆盖的神经。假设$\mathscr{U} = \{U_0, U_1, \cdots, U_n\}$是空间$S$的一个覆盖。以$U_i$为顶点。$U_{i1}, U_{i2}, \cdots, U_{ik}$生成一个k维单纯形当且仅当$U_{i1} \cap U_{i2} \cap \cdots \cap U_{ik} \neq \emptyset$。神经N($\mathscr{U}$)是所有这些单纯形的集合，根据定义，其本身构成一个单纯复形。在拓扑数据分析中，为了给数据S一个几何结构，在每个数据点处放置一个半径为ϵ的球。覆盖$\{B_\epsilon(x)|x \in S\}$的神经被称为数据集$S$的Čech复形，把它记为Čech$(\epsilon, S)$。通过增加半径$\epsilon$，得到一族单纯复形满足

$$\check{\text{C}}\text{ech}(\epsilon_1, S) \subseteq \check{\text{C}}\text{ech}(\epsilon_2, S),\ \epsilon_1 < \epsilon_2$$

更一般的，单纯复形K的子复形序列$\{K_i\}$构成了一个过滤，如果$\{K_i\}$满足

$$\emptyset = K_0 \hookrightarrow K_1 \hookrightarrow K_2 \hookrightarrow \cdots \hookrightarrow K_n = K$$

这些包含映射诱导了同调向量空间的同态

$$0 = H_i(K_0;\mathbb{Z}_2) \to H_i(K_1;\mathbb{Z}_2) \to H_i(K_2;\mathbb{Z}_2) \to \cdots \to H_i(K_n;\mathbb{Z}_2) = H_i(K;\mathbb{Z}_2)$$

在过滤中，从K_i到K_{i+1}添加了新的单纯形。添加一个新的p维单纯形σ或者增加一个新的p维闭链，此时称σ为正；或者减少一个$(p-1)$维边缘，此时称σ为负。[12]证明了每个负的p维单纯形对应唯一一个正的$(p-1)$维单纯形。将这样的正负单纯形进行配对，用一个半闭半开区间$[a_p, b_p)$记正负单纯形对进入过滤的时间，其中a_p 是正单纯形进入过滤的时间，b_p是负单纯形进入过滤的时间，如果一个正单纯形没有负单纯形与之配对，则b_p为∞，$[a_p, b_p)$称为p维条码(barcode)。所有p维条码的集合称为p维持续。

数据集的持续蕴含了其内蕴空间的重要拓扑信息。在一个持续中，长的条码解释为内蕴空间的真实的拓扑特征，而短的条码则解释为拓扑噪声。对于长短的定量的讨论见[7]。

用交换代数可以为持续同调给出形式的描述。在[29],[4]中，条码被解释为分次环上的分次模的标准同调中的不变量。在这个解释下G.Carlsson等给出了计算任意域上的持续同调的算法，并且证明了在不以域为系数的情形下，条码和持续的概念不再成立。

2.2 图模型

条件独立是更一般的衡量随机变量独立性的概念。给定三个随机变量X, Y和Z，如果X和Y在给定Z下的联合条件概率分布满足关系式$f(x, y|z) = f(x|z)f(y|z)$，则称X和Y在给定Z下条件独立（conditional independent），否则称X和Y在给定Z下条件相关（conditional dependent）。考虑一个随机向量$\mathbf{X} = (X_1, X_2, \cdots, X_p)$，其联合概率密度为$f$。定义这样一个无向图，其中每个随机变量定义了一个顶点，两个顶点之间连边当且仅当两个随机变量在给定其他所有随机变量下条件相关，也即，给定$\{X_k|k \neq i\}$下X_i的条件分布与给定$\{X_k|k \neq i, j\}$ 下X_i的条件分布不同。

$$f(x_i|x_1, \cdots, \hat{x_i}, \cdots, x_p) \neq f(x_i|x_1, \cdots, \hat{x_i}, \cdots, \hat{x_j}, \cdots, x_p) \tag{2.3}$$

其中$\hat{x_i}$表示去除x_i。设$I = \{1, 2, \cdots, p\}$是一个指标集。当f处处为正时，式(2.3)等价于

$$f(x_i, x_j|x_{I\setminus\{i,j\}}) \neq f(x_i|x_{I\setminus\{i,j\}})f(x_j|x_{I\setminus\{i,j\}}) \tag{2.4}$$

设$\mathcal{G} = (\mathcal{V}, \mathcal{E})$是无向简单图。$\mathcal{G}$被称为完全图如果$\mathcal{G}$中每两个顶点都连边。图$\mathcal{G}$的一个完全子图，如果相对于包含是极大的，被称为团（clique）。连接两个顶点v_i, v_j的道路是一个起始于v_i，终止于v_j 的边的序列。假设三个集合A, B, S构成了顶点集$\mathcal{V}$的划分。三元组(A, B, S)称为图$\mathcal{G}$的分解，如果S诱导的子图是完全的并且任何连接A中顶点v_a和B中顶点v_b的道路必定经过S。此时S被称为一个分解子。类似的，对$A \cup S$诱导的子图和$B \cup S$诱导的子图作分解，重复该步骤直至所有集合无法继续分解。最终得到的无法分解的子图称为图$\mathcal{G}$的素分支。一个图被称为可分解的，如果所有的素分支都是团，否则称其不可分解。记$\mathcal{P}$为所有素分支的集合，$\mathcal{S}$为所有分解子的集合。Hammersley-Clifford定理[3]称，在满足特定的条件下，联合概率密度函数f相对于它的图存在一个分解。

定理(Hammersley-Clifford). *假设联合概率密度函数f处处为正。$\mathcal{G}$是由f决定的图，并且$\mathcal{G}$是可分解的。那么f相对于$\mathcal{G}$有分解*

$$f(x) = \frac{\prod_{a\in\mathcal{P}} \psi_a(x_a|\theta_a)}{\prod_{b\in\mathcal{S}} \psi_b(x_b|\theta_b)}$$

其中$\mathcal{P}$和$\mathcal{S}$分别是素分支和分解子的集合，$\psi(x|\theta)$是边缘密度函数。

对随机变量条件独立结构的推断在图模型中等价于对图的推断。在标准的贝叶斯推断的框架中，先在图空间和参数空间上置先验分布，然后通过计算似然函

数得到图的后验分布。设$\mathbb{G}_p$是所有有p个顶点的图的集合。$\mathcal{G} \in \mathbb{G}_p$，$\Theta_{\mathcal{G}}$是$\mathcal{G}$ 上的参数空间。图和参数的联合先验分布为

$$p(\mathcal{G},\theta) = p(\mathcal{G})p(\theta|\mathcal{G}), \mathcal{G} \in \mathbb{G}_p, \theta \in \Theta_{\mathcal{G}} \tag{2.5}$$

$p(\mathcal{G})$通常取为$\mathbb{G}_p$上的均匀分布。A.P.David和S.L.Lauritzen对$p(\theta|\mathcal{G}), \theta \in \Theta_{\mathcal{G}}$的选取发展了严谨深刻的理论[8]，一类能够在参数水平遗传随机变量条件独立结构的先验分布扮演了至关重要的角色，这样的先验分布被称为超马尔科夫法则(hyper Markov law)。例如对于零均值的多元高斯分布，其精度矩阵（协方差矩阵Σ的逆）的先验分布$p(\Sigma^{-1}|\mathcal{G})$通常选为超逆威沙特分布（hyper inverse Wishart distribution）。超逆威沙特分布是目前研究最为广泛和深入的超马尔科夫法则。假设$x^{(1)}, x^{(2)}, \cdots, x^{(n)}$ 取样自分布f，每个$x^{(i)}$是$\mathbb{R}^p$中的一个向量。由贝叶斯公式可知

$$Pr(\mathcal{G}|x^{(1)}, \cdots, x^{(n)}) \propto \int_{\Theta_{\mathcal{G}}} f(x^{(1)}, \cdots, x^{(n)}|\theta, \mathcal{G}) p(\mathcal{G}) p(\theta|\mathcal{G}) \mathrm{d}\theta \tag{2.6}$$

这里$f(x^{(1)}, \cdots, x^{(n)}) = \prod_{i=1}^{n} f(x^{(i)}|\theta, \mathcal{G})$。积分$\mathcal{M}(\mathcal{G}) = \int_{\Theta_{\mathcal{G}}} f(x^{(1)}, \cdots, x^{(n)}|\theta, \mathcal{G})$ $p(\theta|\mathcal{G})\mathrm{d}\theta$被称为图$\mathcal{G}$的边缘似然。$\mathcal{M}(\mathcal{G})$的计算是推断中最关键的一步，也是最困难的一步。对于零均值多元正态分布，若精度矩阵服从超逆威沙特分布，$\mathcal{M}(\mathcal{G})$由(2.7)给出

$$\mathcal{M}(\mathcal{G}) = \frac{1}{(2\pi)^{\frac{np}{2}}} \frac{I_{\mathcal{G}}(\delta + n, D + \sum_{i=1}^{n} x^{(i)} x^{(i)t})}{I_{\mathcal{G}}(\delta, D)} \tag{2.7}$$

其中$I_{\mathcal{G}}(\delta, D)$是参数为$\delta$和$D$的超逆威沙特分布的归一化常数。$\delta > 2$是一正数，$D$是一个正定矩阵。$x^t$表示列向量$x \in \mathbb{R}^p$的转置。

如果$\mathcal{G}$可分解，归一化常数$I_{\mathcal{G}}(\delta, D)$有显式表达，从而$\mathcal{M}(\mathcal{G})$也能被显式的公式写出。令$\mathcal{C}$为团的集合。对每个团$C \in \mathcal{C}$，

$$I_C(\delta, D_C) = \frac{2^{\frac{nc}{2}} \Gamma_c(\frac{\delta+c-1}{2})}{|D_C|^{\frac{\delta+c-1}{2}}} \tag{2.8}$$

其中$\Gamma_c(\cdot)$是多元伽马函数

$$\Gamma_c(a) = \pi^{\frac{c(c-1)}{4}} \prod_{i=0}^{i=c-1} \Gamma(a - \frac{i}{2}) \tag{2.9}$$

对于可分解的图，$\mathcal{G}$的归一化常数为

$$I_{\mathcal{G}}(\delta, D) = \frac{\prod_{A \in \mathcal{P}} I_A(\delta, D_A)}{\prod_{B \in \mathcal{S}} I_B(\delta, D_B)} \tag{2.10}$$

如果图$\mathcal{G}$不可分解，那么归一化常数只能通过数值方法进行计算。[2]对正态分布的情况给出了一个高效的近似计算归一化常数的算法。

3 乘积空间的拓扑

本节讨论乘积空间的拓扑。首先考虑的问题是如何给单纯复形的乘积空间一个合理的单纯剖分，这里的方法来自[22],[17]。然后给出了域上的Künneth定理的一个初等的线性代数的证明。最后给出一个例子验证Künneth定理，这个例子也是高维持续同调计算算法的启发。

3.1 乘积空间的单纯剖分

设K和L为单纯复形。K和L的笛卡尔积$K \times L$中的元素有$\sigma \times \tau, \sigma \in K, \tau \in L$的形式，可以看出$K \times L$一般而言不是单纯复形。例如，两个单位区间（1维单纯形）的笛卡尔积是一个单位正方形，而单位正方形不是一个2维单纯形。必须先给乘积空间$|K| \times |L|$一个单纯剖分，才能讨论$|K| \times |L|$上的单纯同调。

考虑这样一个例子：令$\sigma = v_0v_1v_2$和$\tau = u_0u_1$为两个单纯形。σ是一个三角形，τ是一条线段。$\sigma \times \tau$是一个三棱柱。

$\sigma \times \tau$的顶点为$(v_i, u_j), i = 0, 1, 2; j = 0, 1$。令$(i, j)$ 表示2×1矩形网格中的格点。设α是网格中一条起始于$(0, 0)$结束于$(2, 1)$的道路，每次只能向右走或者向上走。网格上一共有三条不同的道路，分别为

$$\begin{aligned}\alpha_0 &= (0,0) \to (1,0) \to (2,0) \to (2,1)\\ \alpha_1 &= (0,0) \to (1,0) \to (1,1) \to (2,1)\\ \alpha_2 &= (0,0) \to (0,1) \to (1,1) \to (2,1)\end{aligned}$$

记顶点(v_i, u_j)为w_{ij}。定义三个单纯形

$$\kappa_0 = w_{00}w_{10}w_{20}w_{21},\ \kappa_1 = w_{00}w_{10}w_{11}w_{21},\ \kappa_2 = w_{00}w_{01}w_{11}w_{21}$$

这三个3维单纯形给出了$\sigma \times \tau$的剖分，换言之，$\sigma \times \tau$是这三个单纯形的并。

一般地，令$\sigma = v_0v_1 \cdots v_m, \tau = u_0u_1 \cdots u_n$。$\sigma \times \tau$的顶点记为$w_{ij}$。有$\binom{m+n}{m}$起始于$(0, 0)$终止于$(m, n)$的道路。记$\alpha_r$为这样的一条道路。定义

$$\kappa_r = w_{\alpha_r(0)} \cdots w_{\alpha_r(m+n)}$$

$\sigma \times \tau$是这些$m + n$维单纯形的并。

设K, L为单纯复形，在K和L的顶点集上分别置一个全序。定义$K \times L$为包含κ_r和它的所有面的集合。$K \times L$ 是一个单纯复形，它的底空间$|K \times L|$与$|K| \times |L|$有相同的拓扑。因此可以用$|K \times L|$表示$|K| \times |L|$。

令$\{\mathcal{C}(K; \mathbb{Z}_2), \partial_K\}$为$K$的链复形，$\{\mathcal{C}(L; \mathbb{Z}_2), \partial_L\}$为$L$的链复形。两个链复形的张

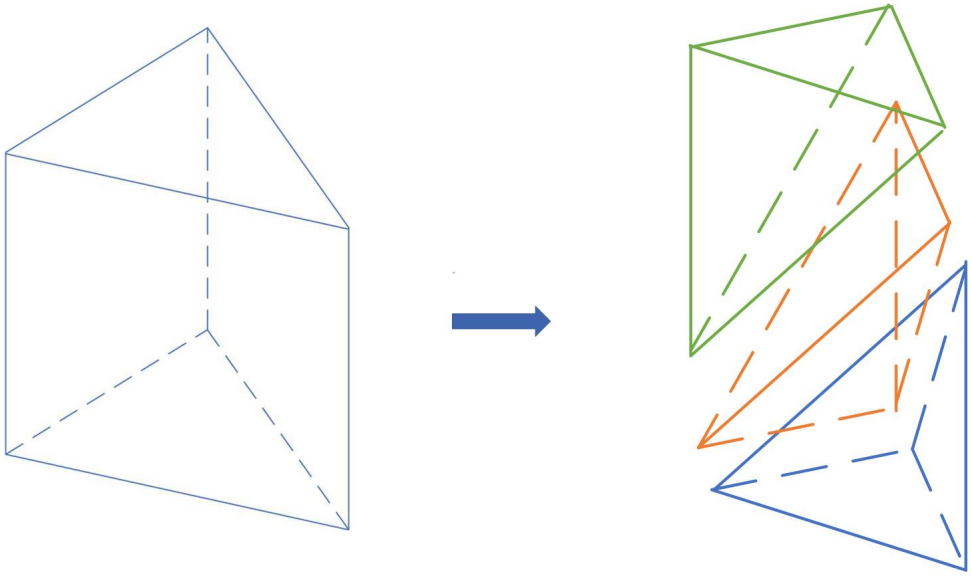


图 1: 三棱柱$\sigma \times \tau$分解为三个单纯形$\kappa_0, \kappa_1, \kappa_2$的并。

量积$\{\mathcal{C}(K;\mathbb{Z}_2) \otimes \mathcal{C}(L;\mathbb{Z}_2), \bar{\partial}\}$定义为

$$(\mathcal{C}(K;\mathbb{Z}_2) \otimes \mathcal{C}(L;\mathbb{Z}_2))_p = \bigoplus_{i+j=p} (C_i(K;\mathbb{Z}_2) \otimes C_j(L;\mathbb{Z}_2))$$

$$\bar{\partial}(\sigma \otimes \tau) = (\partial_K \sigma) \otimes \tau + \sigma \otimes (\partial_L \tau)$$

由公式(3.1)定义一个对应$\phi : (\mathcal{C}(K;\mathbb{Z}_2) \otimes \mathcal{C}(L;\mathbb{Z}_2))_p \to C_p(K \times L;\mathbb{Z}_2)$

$$\phi(\sigma \otimes \tau) = \sum_{\cup \kappa_r = \sigma \times \tau} \kappa_r \tag{3.1}$$

下面验证ϕ是一个链映射。也即，如下图表交换

$$\begin{array}{ccc} (\mathcal{C}(K;\mathbb{Z}_2) \bigotimes \mathcal{C}(L;\mathbb{Z}_2))_p & \xrightarrow{\phi} & C_p(K \times L;\mathbb{Z}_2) \\ \Big\downarrow{\scriptstyle \bar{\partial}} & & \Big\downarrow{\scriptstyle \partial_{K\times L}} \\ (\mathcal{C}(K;\mathbb{Z}_2) \bigotimes \mathcal{C}(L;\mathbb{Z}_2))_{p-1} & \xrightarrow{\phi} & C_{p-1}(K \times L;\mathbb{Z}_2) \end{array}$$

直接计算表明

$$\partial_{K\times L}(\phi(\sigma \otimes \tau)) = \partial_{K\times L}\left(\sum_{\cup \kappa_r = \sigma \times \tau} \kappa_r\right) = \sum_{\cup \kappa_r = \sigma \times \tau} \left(\sum_{\theta_r \in \mathscr{F}(\kappa_r)} \theta_r\right) \tag{3.2}$$

注意到在$\sigma \times \tau$内部的(p-1)面θ不会出现在和式中，因为θ必定是两个单纯形κ_α和κ_β的公共面。因此最后的和式只包含$Bd(\sigma \times \tau)$的(p-1)维单纯形。这里$Bd(\sigma \times \tau)$表示$\sigma \times$

τ的边界。

$$\begin{aligned}\phi(\bar{\partial}\sigma\otimes\tau) &= \phi((\partial_K\sigma)\otimes\tau+\sigma\otimes(\partial_L\tau))\\ &= \sum_{\eta\in\mathscr{F}(\sigma)}\phi(\eta\otimes\tau)+\sum_{\epsilon\in\mathscr{F}(\tau)}\phi(\sigma\otimes\epsilon)\\ &= \sum_{\eta\in\mathscr{F}(\sigma)}\left(\sum_{\theta\in Bd\eta\times\tau}+\sum_{\theta\in\eta\times Bd\tau}\right)\theta+\sum_{\epsilon\in\mathscr{F}(\tau)}\left(\sum_{\theta\in Bd\sigma\times\epsilon}+\sum_{\theta\in\sigma\times Bd\epsilon}\right)\theta\\ &= \sum_{\theta\in\sigma\times(Bd\tau)}\theta+\sum_{\theta\in(Bd\sigma)\times\tau}\theta\\ &= \partial_{K\times L}(\phi(\sigma\otimes\tau))\end{aligned}$$

这里用到了一个事实$Bd(\sigma\times\tau)=(Bd\sigma)\times\tau\cup\sigma\times(Bd\tau)$。

事实上，这个链映射是一个链同伦等价[14]。因此链映射ϕ诱导了向量空间的同构。

定理(Eilenberg). 设K和L为单纯复形，链复形ϕ在每个维数n都诱导了向量空间的同构$\phi_*: H_n(\mathcal{C}(K;\mathbb{Z}_2)\otimes\mathcal{C}(L;\mathbb{Z}_2))\to H_n(K\times L;\mathbb{Z}_2)$。

3.2 Künneth定理的证明

由于ϕ_*诱导了$H_n(K\times L;\mathbb{Z}_2)$和$H_n(\mathcal{C}(K;\mathbb{Z}_2)\otimes\mathcal{C}(L;\mathbb{Z}_2))$的同构，所以计算$K\times L$的同调等价于计算链复形的张量积的同调，后者则是经典的Künneth定理。这个定理的一般形式的表述和证明需要用到同调代数的相关知识[22]。但是域上的Künneth定理的表述和证明只需要线性代数。以下讨论均假设在一个域$\mathbb{F}$上，方便起见在所有记符中都省略了$\mathbb{F}$。

定理(Künneth). 设$\{\mathcal{C},\partial\}$和$\{\mathcal{D},\partial'\}$为两个链复形。$\{\mathcal{C}\otimes\mathcal{D},\bar{\partial}\}$ 是这两个链复形的张量积。对每个整数n，都有向量空间的同构

$$H_n(\mathcal{C}\otimes\mathcal{D})\cong\bigoplus_{p+q=n}(H_p(\mathcal{C})\otimes H_q(\mathcal{D}))$$

证明. 两个向量空间同构当且仅当它们的维数相同。令$h_p=\dim(H_p(\mathcal{C}))$，$h'_q=\dim(H_q(\mathcal{D}))$。注意到$\bigoplus_{p+q=n}(H_p(\mathcal{C})\otimes H_q(\mathcal{D}))$的维数为$\sum_{p+q=n}h_ph'_q$。只需要计算$H_n(\mathcal{C}\otimes\mathcal{D})$的维数。

由p维边缘同态$\partial_p: C_p\to C_{p-1}$可知有如下的向量空间同构

$$C_p\cong Z_p\oplus B_{p-1},\ Z_p\cong B_p\oplus H_p \tag{3.3}$$

令$r_p=\mathrm{rank}(\partial_\mathrm{p})$，那么$\dim(C_p)=r_p+r_{p+1}+h_p$。设$D_q$是$\mathcal{D}$的$q$维链向量空间，$r'_q=\mathrm{rank}(\partial'_\mathrm{q})$。相同的讨论表明$\dim(D_q)=r'_q+r'_{q+1}+h'_q$。

选择C_p和C_{p-1}的恰当的基底，使得∂_p的矩阵表示是对角的。事实上，从同构(3.3)可知存在C_p的基底$\{w_1,\cdots,w_{r_p},b_1,\cdots,b_{r_{p+1}},v_1,\cdots,v_{h_p}\}$和$C_{p-1}$的基底$\{\hat{w}_1,\cdots,\hat{w}_{r_{p-1}},\hat{b}_1,\cdots,\hat{b}_{r_p},\hat{v}_1,\cdots,\hat{v}_{h_{p-1}}\}$满足$\partial_p$在基底下具有形式

$$\begin{array}{c} \\ \hat{b}_1 \\ \vdots \\ \hat{b}_{r_p} \\ \hat{w}_1 \\ \vdots \\ \hat{v}_{h_{p-1}} \end{array}\begin{array}{c} \begin{array}{cccccc} w_1 & \cdots & w_{r_p} & b_1 & \cdots & v_{h_p} \end{array} \\ \left(\begin{array}{cccc|cccc} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ \hline & & & & & & & \\ & & & & & 0 & & \\ & & & & & & & \end{array}\right) \end{array}$$

用$E(p,q)$表示$C_p\otimes D_q$。考虑n维边缘同态$\bar{\partial}_n:\bigoplus_{p+q=n}E(p,q)\to\bigoplus_{p+q=n-1}E(p,q)$。这个线性映射的矩阵表示为

$$\begin{array}{r} \\ \vdots \\ E(p-1,q) \\ E(p,q-1) \\ E(p+1,q-2) \\ \vdots \end{array}\begin{array}{c} \begin{array}{ccccc} \cdots & E(p-1,q+1) & E(p,q) & E(p+1,q-1) & \cdots \end{array} \\ \begin{pmatrix} \cdots & & & & \\ & I_{p-1}\otimes\partial'_{q+1} & \partial_p\otimes I'_q & & \\ & & I_p\otimes\partial'_q & \partial_{p+1}\otimes I'_{q-1} & \\ & & & I_{p+1}\otimes\partial'_{q-1} & \\ & & & & \cdots \end{pmatrix} \end{array} \tag{3.4}$$

其中$I\otimes\partial'$和$\partial\otimes I'$是矩阵的张量积（克朗内克积）。这里I_p和I'_q表示C_p和D_q上的单位阵。以上讨论表明可以选择$E(p,q)$中恰当的基底，使得∂和∂'均为对角阵。因此$I\otimes\partial'$和$\partial\otimes I'$具有形式

$$I_p\otimes\partial'_q=\begin{pmatrix} I_p & & & & & \\ & \ddots & & & & \\ & & I_p & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix},\quad \partial_p\otimes I_q=\begin{pmatrix} I_{r_p} & 0 & & & & \\ 0 & 0 & & & & \\ & & \ddots & & & \\ & & & \ddots & & \\ & & & & I_{r_p} & 0 \\ & & & & 0 & I_{r_p} \end{pmatrix} \tag{3.5}$$

将这个表示带入式(3.4)中。用初等行变换将矩阵$\bar{\partial}$化为阶梯型，然后计算矩阵的秩。注意到$I\otimes\partial'$已经是对角阵。只需要用$I_p\otimes\partial'_q$消除$\partial_p\otimes I'_q$在同一列的元素即可。经过初等行变换h'_q个大小为r_p单位矩阵保留下来。因此$\bar{\partial}_n$的秩为$\bar{r}_n=\sum_{p+q=n}(r'_q\dim(C_p)+$

$h'_q r_p$)。

$$\begin{aligned}\dim(\bigoplus_{p+q=n} E(p,q)) &= \sum_{p+q=n} \dim(C_p)\dim(D_q)\\ &= \sum_{p+q=n} (r'_q + r'_{q+1} + h'_q)\dim(C_p)\\ &= \sum_{p+q=n} (r'_q \dim(C_p) + r'_{q+1}\dim(C_p)) + \sum_{p+q=n} h'_q(r_p + r_{p+1} + h_p)\\ &= \sum_{p+q=n} (r'_q \dim(C_p) + h'_q r_p) + \sum_{p+q=n} (r'_{q+1}\dim(C_p) + h'_q r_{p+1}) + \sum_{p+q=n} h_p h'_q\\ &= \bar{r}_n + \bar{r}_{n+1} + \sum_{p+q=n} h_p h'_q\end{aligned}$$

这个计算表明$\dim(H_n(\mathcal{C}\otimes\mathcal{D})) = \sum_{p+q=n} h_p h'_q$。证毕。

可以按如下方法构造$H_n(\mathcal{C}\otimes\mathcal{D})$ 和$\bigoplus(H_p(\mathcal{C})\otimes H_q(\mathcal{D}))$之间的一个显式同构。设$[\sigma]$和$[\tau]$分别是$H_p(\mathcal{C})$和$H_q(\mathcal{D})$中的同调类。将$[\sigma]\otimes[\tau]$映为$[\sigma\otimes\tau]$的线性映射$\psi$是良定义的。通过对投影映射$C_p \to H_p(\mathcal{C})$和$D_q \to H_q(\mathcal{D})$作张量积，得到从$\bigoplus(C_p\otimes C_q)$到$\bigoplus H_p(\mathcal{C})\otimes H_q(\mathcal{D})$的映射。这个映射下降为同调向量空间上的线性映射$\lambda : H_n(\mathcal{C}\otimes\mathcal{D}) \to \bigoplus H_p(\mathcal{C})\otimes H_q(\mathcal{D})$。由$\lambda(\psi([\sigma]\otimes[\tau])) = [\sigma]\otimes[\tau]$可知$\psi$是一个单同态。已经证明了$H_n(\mathcal{C}\otimes\mathcal{D})$和$\bigoplus H_p(\mathcal{C})\otimes H_q(\mathcal{D})$有相同的维数，从而$\psi$是一个同构。

推论. *设K和L为单纯复形。对每个整数n，存在向量空间同构$H_n(K\times L) \cong \bigoplus_{p+q=n} (H_p(K)\otimes H_q(L))$。*

3.3 例子

平坦环面$\mathbb{T}^2 = \{(x,y,z,w)\in\mathbb{R}^4 | x^2+y^2=\frac{1}{2}, z^2+w^2=\frac{1}{2}\}$是两个圆的乘积空间，其中每个圆都嵌入在一个2维平面里。因此有$\mathbb{T}^2 = \mathbb{S}^1\times\mathbb{S}^1 \subseteq \mathbb{R}^2\times\mathbb{R}^2$。设$\alpha$和$\beta$是服从$[0,1]$上的均匀分布的随机变量，通过变换

$$\begin{cases} x = \frac{1}{\sqrt{2}}\cos(2\pi\alpha) \\ y = \frac{1}{\sqrt{2}}\sin(2\pi\alpha) \end{cases} \quad \begin{cases} z = \frac{1}{\sqrt{2}}\cos(2\pi\beta) \\ w = \frac{1}{\sqrt{2}}\sin(2\pi\beta) \end{cases}$$

得到取样自$\mathbb{T}^2$的点集S。S的持续同调可以直接计算，或者先计算圆的持续同调，再使用Künneth定理。Künneth定理说明$\mathbb{T}^2$的n维贝蒂数是两个圆的贝蒂数的卷积。由图2可知，$\mathbb{T}^2$在0维，1维和2维的贝蒂数分别是1，2和1，其余维数维0。由圆的持续同调可知圆的0维和1维贝蒂数均为1，其余维数为0。对两个圆用Künneth定理得到的结果和直接计算的结果相同。

4 乘积空间持续同调的计算

由上节的例子可知，乘积空间的持续同调的计算在已知乘积结构的情况下会

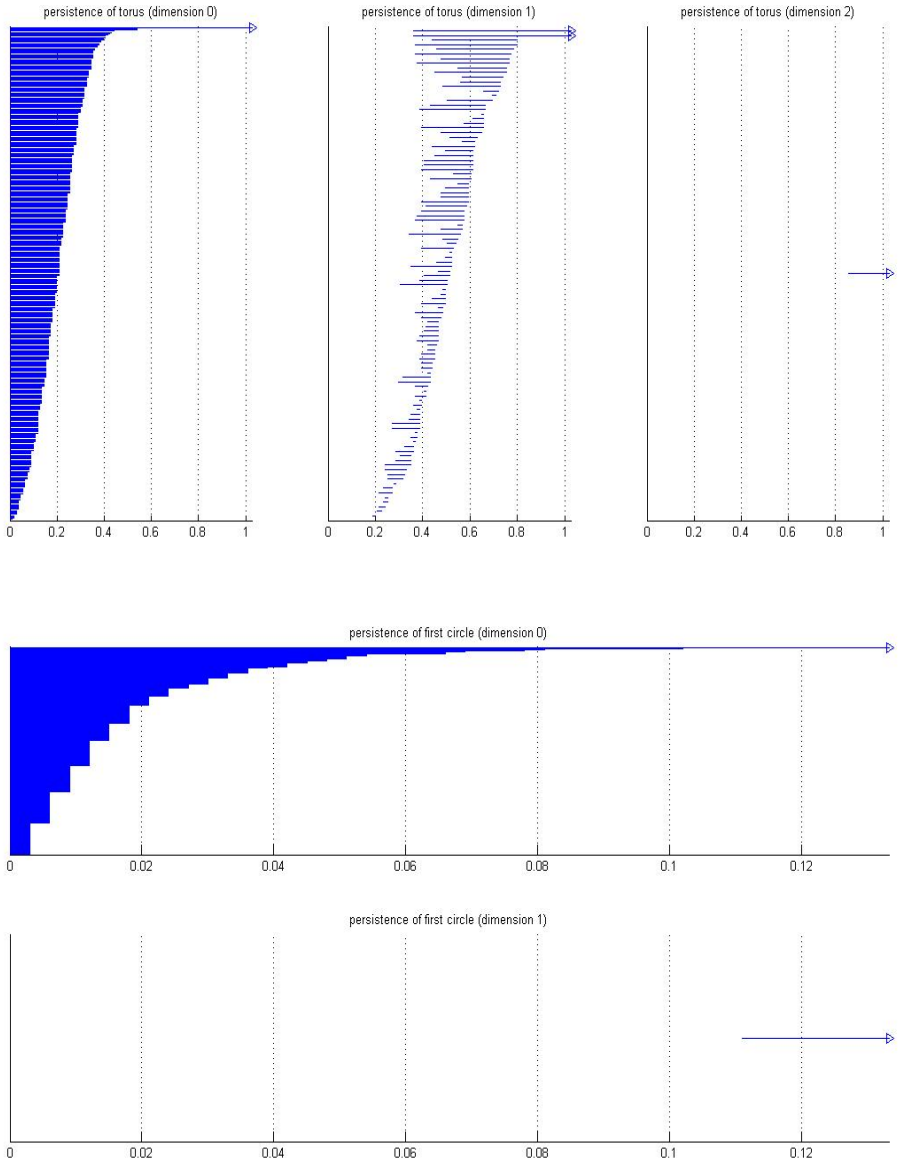


图 2: 平坦环面和圆的持续

变得简单。假设分量空间为M和N，通过投影至M和N并分别计算分量空间的持续同调，最后应用Künneth定理即可。为了推断乘积空间的结构，假设每个数据点来自某个随机向量$\mathbf{X} = (X_1, X_2, \cdots, X_p)$，并对$\mathbf{X}$的条件独立结构进行推断，由2.2节的讨论，这等价于对$\mathbf{X}$的图进行推断。本节首先介绍由S.Lunagómez和S.Mukherjee等[21]提出的图贝叶斯蒙特卡洛马尔科夫算法。然后为了改进这个算法，讨论了不连通图空间的一个渐进性质。最后给出改进的计算图的后验概率的算法。

4.1 图贝叶斯推断

假设$x^{(1)}, x^{(2)}, \cdots, x^{(n)}$取样自分布$f(\mathbf{x}|\theta)$。为了推断$f$的条件独立结构，我们计算每个图的后验概率，后验概率最大的图，就是与f的条件独立结构相匹配的图。由2.2节的讨论可知，需要计算积分(2.6)

$$\int_{\Theta_{\mathcal{G}}} f(x_1, \cdots, x_n|\theta, \mathcal{G}) p(\mathcal{G}) p(\theta|\mathcal{G}) \mathrm{d}\theta$$

一般而言先验分布$p(\mathcal{G})$通常取图空间上的均匀分布。假设$\mathbb{G}_p$所有有p个顶点的图的集合。从$\mathbb{G}_p$取到图$\mathcal{G}$ 的概率是$\frac{1}{|\mathbb{G}_p|}$。这样一个先验分布有它的好处和坏处。当p很小的时候，均匀分布是$\mathbb{G}_p$上一个简单而且自然的分布。然而当p很大时这个分布处理起来不够方便，因为$|\mathbb{G}_p| = 2^{\frac{p(p-1)}{2}}$的规模呈指数增长。如果要知道每个图的后验概率，就需要把积分(2.6)计算$2^{\frac{p(p-1)}{2}}$次。[21]提出了这样的处理方法，首先给每个图一个参数化，通过这样一个参数化，将图上的先验分布转化为欧氏空间中p个点的分布。同时，从图的后验分布取样化归为从点的后验分布取样。后者则通过对传统的蒙特克洛马尔科夫算法进行改进来实现。其中每次迭代的Metropolis-Hastings比例就是积分(2.6)。这个方法为数值计算带来了巨大的方便。

注意到每个单纯复形的1维骨架就是图。设$v_1, \cdots, v_p \in \mathbb{R}^m$是欧氏空间中任意p个点，构造一个单纯复形并取它的1维骨架，这样就在欧氏空间的点和图之间建立了对应。以Čech 复形为例。设$r > 0$为一个正数。Čech复形是通过取覆盖$\{B_r(v_i), i = 1, 2, \cdots, n\}$的神经得到的。其中$B_r(v_i)$是以$v_i$为球心，$r$为半径的闭球。它的1维骨架包括顶点$v_1, \cdots, v_p$ 和边$(v_i, v_j)(i, j = 1, 2, \cdots, p)$，$v_i$和$v_j$连边当且仅当$||v_i - v_j|| \leq 2r$。记这个图为$\mathcal{G}(v_1, \cdots, v_p, r)$。这个对应使得图空间成为欧氏空间$\mathbb{R}^{pm+1}$的子集。

固定$r > 0$，任意$\mathbb{R}^m$中p个向量的联合概率分布诱导了图空间上的概率分布。方便讨论假设每个顶点服从$\mathbb{R}^2$中单位圆盘上的均匀分布。从$v \in B^2$开始的随机游走给出了B^2上的提议分布。设(ρ, θ)是$\mathbb{R}^2$上的极坐标。令$v^k = (\rho^k, \theta^k)$是第k次移动。固定参数$\eta > 0$，第$(k+1)$次移动v^*由式(4.1)给出

$$\begin{cases} \rho^* = R([\rho^k]^2 + \xi_1^k \eta)^{1/2} \\ \theta^* = \theta^k + \xi_2^k \eta / \rho^k \end{cases} \tag{4.1}$$

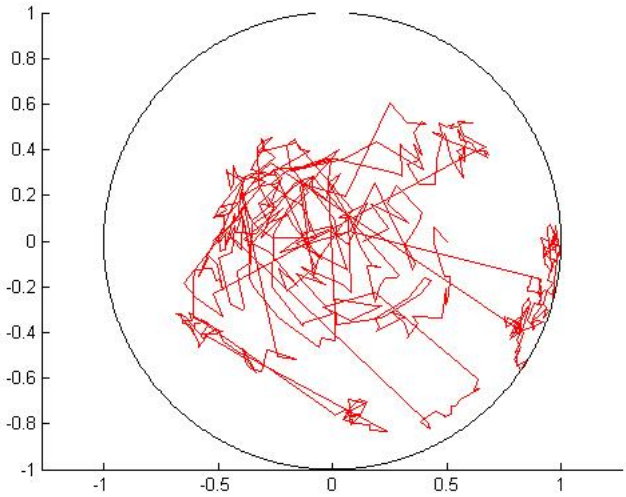


图 3: 单位圆盘中的500次随机游走

其中(ξ_1^k, ξ_2^k)是独立同分布的标准正态随机变量，$R(x) = |x - 2\lfloor 1/2(x+1) \rfloor|$。

令$V = (v_1, \cdots, v_p)$。分别对这p个点进行独立的随机游走，得到从V^k到V^*的移动。设$\mathcal{G}^k$和$\mathcal{G}^*$分别是由V^k和V^*决定的图。计算如下的Metropolis-Hastings比例

$$H^k = \frac{\mathcal{M}(\mathcal{G}^*)p(V^*)q(V^k|V^*)}{\mathcal{M}(\mathcal{G}^k)p(V^k)q(V^*|V^k)} \tag{4.2}$$

其中$p(V)$是V的先验分布，$q(V^*|V)$是从$V \in (B^2)^p$到$V^* \in (B^2)^p$的一步随机游走的勒贝格密度，$\mathcal{M}(\mathcal{G})$是$\mathcal{G}$的边缘似然。对于零均值多元正态分布，且精度矩阵满足超逆威沙特分布，将式(2.7)，(2.8)和(2.10)带入(4.2)，Metropolis-Hastings比例是

$$H^k = \frac{I_{\mathcal{G}^*}(\delta + n, D + \sum_{i=1}^n x^{(i)} x^{(i)t})}{I_{\mathcal{G}^k}(\delta + n, D + \sum_{i=1}^n x^{(i)} x^{(i)t})} \frac{I_{\mathcal{G}^k}(\delta, D)}{I_{\mathcal{G}^*}(\delta, D)} \tag{4.3}$$

移动V^*以概率$1 \wedge H^k$被接受。经过预烧期，通过后验取样，对每个样本计算其频数，便能估计图的后验概率。整个过程总结为算法1。

4.2 不连通图空间

在MCMC算法中，图的后验概率通过其频数来逼近。当$\mathbb{G}_p$规模很大的时候，这个方法仍然存在缺陷。事实上，实际应用中无法得到足够多的样本，使得每个图的频数充分接近它的后验概率。为了克服这个缺陷，可以考虑使用$\mathbb{G}_p$的某个子空间来降低规模。既然我们最关心的是图的连通分支的情况，考虑包含所有不连通的图的子空间是自然的。下面的讨论说明这个子空间确实能够简化计算。

$\mathbb{G}_p$是包含$2^{\frac{p(p-1)}{2}}$个元素的有限集。记所有有p个顶点的连通图的空间为$\mathbb{G}_p^c$，所有有p个顶点的不连通图的空间为$\mathbb{G}_p^d$。为了比较这两个空间的大小，定义空间$\mathbb{G}_p$的

Algorithm 1 图贝叶斯蒙特卡洛马尔科夫算法

初始化:

1. 取任意一个初始顶点集$V^{(0)} = (v_1^{(0)}, v_2^{(0)}, \cdots, v_p^{(0)}) \in \mathbb{R}^p$；
2. 置$r > 0, k = 0$；

迭代:

1. 用随机游走生成一个候选顶点集$V^* = (v_1^*, v_2^*, \cdots, v_p^*)$；
2. 计算接受概率$A(V^*|V^k) = \min\{1, H^k\}$；
3. 生成随机数$a \in [0, 1]$。若$a \leq A(V^*|V^k)$，接受候选顶点集并置$V^{k+1} = V^*$；否则拒绝候选顶点集并置$V^{k+1} = V^k$；
4. 置$k = k + 1$。

连通比为

$$R_p = \frac{\#(\mathbb{G}_p^d)}{\#(\mathbb{G}_p)} \tag{4.4}$$

下面这个命题说明了R_p的渐进性质。

命题. $\lim\limits_{p\to\infty} R_p = 0$

Proof. 令$K_c(p) = \#(\mathbb{G}_p^c), K_d(p) = \#(\mathbb{G}_p^d)$。因为有$p$个顶点的完全图有$\frac{p(p-1)}{2}$条边，而每个有$p$个顶点的图都是有$p$顶点的完全图的子图，故$K_c(p) + K_d(p) = 2^{\frac{p(p-1)}{2}}$。

设$v_0, v_1, \cdots, v_p$为图$\mathcal{G} \in \mathbb{G}_{p+1}$的$p$+1个顶点。考虑包含$v_0$的连通分支$C_0$。假设$C_0$包含了$k$个顶点。余下的$p + 1 - k$个顶点不能通过任何边连接到$C_0$。因为$\mathcal{G}$由$C_0$和任意一个有$p + 1 - k$个顶点的子图决定，故有如下递推式成立。

$$K_d(p+1) = \sum_{k=1}^{p} \binom{p}{k-1} K_c(k) 2^{\frac{(p-k)(p-k+1)}{2}} \tag{4.5}$$

式(4.5)两边同除以$2^{\frac{p(p+1)}{2}}$

$$R_{p+1} = \frac{K_d(p+1)}{2^{\frac{(p+1)p}{2}}} = \sum_{k=1}^{p} \binom{p}{k-1} (1 - R_k) 2^{-k(p+1-k)} \tag{4.6}$$

注意到$k > 1$时，$1 - R_k < 1$。式(4.6)两边同乘2^p

$$2^p R_{p+1} < \sum_{k=1}^{p} \binom{p}{k-1} 2^{-(p-k)(k-1)} = 1 + p + \sum_{k=2}^{p-1} \binom{p}{k-1} 2^{-(p-k)(k-1)} \tag{4.7}$$

当$2 \leq k \leq p - 1$时，$2^{-(p-k)(k-1)} \leq 2^{-(p-2)}$。因此

$$2^p R_{p+1} < 1 + p + 2^p 2^{-(p-2)} = p + 5 \tag{4.8}$$

故$R_{p+1} < \frac{p+5}{2^p}$。显然有$R_p \to 0,\ p \to \infty$。 □

这个命题说明当p充分大的时候，几乎每个有p个顶点的图都是连通的。对于充分大的p，如果f决定的图不是连通的，那么用算法1得到的估计值也很难正确。其主要原因是候选顶点集或者难以被接受，或者被接受时其决定的图为连通图。如果已经知道随机变量之间有较弱的相关性，那么只在不连通图的空间进行取样将克服这个困难，极大地提升计算的效率。

4.3 有限制的蒙特卡洛马尔科夫链算法

给定一个数据集$S = \{x^{(1)}, x^{(2)}, \cdots, x^{(n)}\}$，假设每个数据$x^{(i)}$取自随机向量$\mathbf{X}$。令$u$是一个给定的正整数。为了将MCMC算法限制在子空间$\mathbb{G}_p^d$上，把初始向量V^0随机地划分为u组。设$p_1, p_2, \cdots, p_u$为满足$p_1 + p_2 + \cdots + p_u = p$的$u$个数。对含$p_i$个元素的组，通过随机游动生成一个候选的有$p_i$个顶点的连通分支$\mathcal{G}_{p_i}^*$。所有的连通分支以概率$1 \wedge \widetilde{H}$被接受，其中Metropolis-Hastings比例为

$$\widetilde{H} = \prod_{i=1}^{m} H(\mathcal{G}_{p_i}) \tag{4.9}$$

因此得到有限制的MCMC算法

Algorithm 2 u连通图蒙特卡洛马尔科夫算法

初始化:

1. 取任意一个顶点集$V^{(0)} = (v_1^{(0)}, v_2^{(0)}, \cdots, v_p^{(0)}) \in \mathbb{R}^p$；
2. 固定一个划分$p_1 + p_2 + \cdots + p_u = p$。置$V_{p_i}^{(0)} = (v_{k_1}^{(0)}, v_{k_2}^{(0)}, \cdots, v_{k_{p_i}}^{(0)}), r > 0, k = 0$；

迭代:

1. 对每个p_i，用随机游走生成一个候选顶点集$V_{p_i}^* = (v_{k_1}^*, v_{k_2}^*, \cdots, v_{k_{p_i}}^*)$；
2. 计算接受概率$A(V^*|V^k) = \min\{1, \widetilde{H}^k\}$；
3. 生成一个随机数$a \in [0, 1]$。如果$a \leq A(V^*|V^k)$，接受候选顶点集并置$V^{k+1} = V^*$；否则拒绝候选顶点集并置$V^{k+1} = V^k$；
4. 置$k = k + 1$.

当$u = 1$时，算法2从整个图空间$\mathbb{G}_p$取样，从而等价于原始的MCMC算法。当$u = 2$时，算法2从不连通图空间$\mathbb{G}_p^d$取样。当$u > 2$时，算法2从包含所有至少有u个连通分支的图的集合中取样。u越大，取样的空间就越小，对随机变量的相关性的假设也更强。

只要知道了$\mathbf{X}$分为u个互相独立的随机变量组，就能推测数据集S的内蕴空间是u个因子空间的乘积。将数据集S投影到它的独立的组上得到u个数据集$S_i, i = 1, 2, \cdots, u$。S_i的内蕴空间的乘积就是S的内蕴空间。S的持续同调的计算下降为每个S_i上的计算。

假设S_i的k维贝蒂数是β_i^k，算法3表述了这个过程

Algorithm 3 乘积空间的持续同调算法

初始化: 对所有k，置$\Omega^k = \beta_1^k$；

迭代:

1: **while** 存在$\beta_i^k \neq 0$ **do**
2: **for** $i = 2 : m$ **do**
3: $\Omega^k = \sum_{p+q=k} \beta_i^p \Omega^q$
4: **end for**
5: **end while**

5 仿真结果

我们用两个例子来阐述4中的方法。在第一个例子里我们从两个2维球面的乘积空间中取样得到数据集，这是一个嵌入在6维欧式空间的4维流形，它的持续同调用现有的算法无法直接进行计算。在第二个例子里我们将方法应用于自然图像统计。G.Carlsson等人计算了高对比度，高密度的3×3光学图像块的持续同调，并提出其内蕴空间是一个三圆模型（three circle model）[5]。通过计算我们为其给出一个新的高维模型，它是三圆模型和$\mathbb{R}^4$中实心圆盘的乘积。这个模型和三圆模型有相同的伦型，从而有相同的拓扑。

所有的计算都是在Matlab上用`JavaPlex`软件进行的。这个软件可以在`http://appliedtopology.github.io/javaplex/`免费得到。

5.1 球面的乘积

考虑如下空间

$$W = \{(x_1, y_1, z_1, x_2, y_2, z_2) \in \mathbb{R}^6 | (x_1)^2 + (y_1)^2 + (z_1)^2 = 1, (x_2)^2 + (y_2)^2 + (z_2)^2 = 1\}$$

W是两个单位球面$\mathbb{S}^2$的乘积，其中每个球面嵌入在$\mathbb{R}^3$中。为了从W中取样，先从两个随机向量$\Theta = (\theta_1, \theta_2)$和$\Phi = (\phi_1, \phi_2)$取样，其中$\Theta$和$\Phi$都服从单位正方形$[0, 1] \times [0, 1]$上的均匀分布。然后由变换

$$\begin{cases} x_1 = \sin(\pi\theta_1)\cos(2\pi\theta_2) \\ y_1 = \sin(\pi\theta_1)\sin(2\pi\theta_2) \\ z_1 = \cos(\pi\theta_1) \end{cases} \begin{cases} x_2 = \sin(\pi\phi_1)\cos(2\pi\phi_2) \\ y_2 = \sin(\pi\phi_1)\sin(2\pi\phi_2) \\ z_2 = \cos(\pi\phi_1) \end{cases} \tag{5.1}$$

得到W上的数据。从W上取1000个样本，将每个样本视为取自随机向量$\mathbf{X} = (X_1, X_2, \cdots, X_6)$，下面对$\mathbf{X}$决定的图$\mathcal{G}$进行推断。令$u = 2$，在算法2中，置$\delta = 5$，$D = 0.5I$，在$10,000$步

表 1: 后验概率最高的三个图

图拓扑	后验概率
$\{\mathbf{1},\mathbf{2},\mathbf{3}\}\{\mathbf{4},\mathbf{5},\mathbf{6}\}$	0.2980
$\{1,2,3,4\}\{5,6\}$	0.0850
$\{1,2\}\{3,4,5,6\}$	0.0720

迭代的预烧期后取1,000个样本。后验概率最高的3个图列在表1中。括号表示图的不同的连通分支。

具有最大的后验概率的图表示$\mathbf{X}$的前三个随机变量和后三个随机变量独立，因此数据集的内蕴空间W最有可能是两个嵌入在$\mathbb{R}^3$ 的子空间的乘积。将点集分别投影到前三个坐标和后三个坐标上，得到数据集W_1和W_2。分别计算W_1和W_2的持续同调，最后的结果与2维球面的同调一致。由Künneth定理得到W的同调为

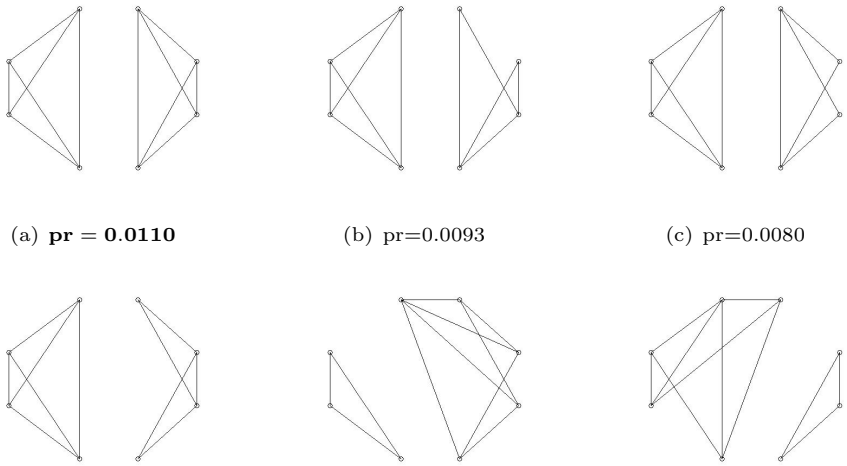
$$H_n(X) = \begin{cases} \mathbb{Z}_2, n = 0, 4 \\ \mathbb{Z}_2 \oplus \mathbb{Z}_2, n = 2 \\ 0, \text{其他情况} \end{cases}$$

5.2 自然图像统计

用数码照相机拍出的图像包含了上千万的像素，每个像素可以取到255个灰度值。因此一个图像可以视为一个极高维的线性空间$\mathbb{R}^P$中的向量，其中P是像素的数量。自然图像统计关心的问题是，一个数码照相机能够拍到的所有的图像构成的空间，蕴含了什么样的统计信息。A.Lee，K.Pedersen和D.Mumford提出，虽然没有好的办法直接考虑这个极高维的流形，可以考虑一个由3×3的图像块构成的低维子流形[19]。在[19]中，预处理后的高对比度图像块位于一个7维球面上，研究发现高对比度的图像块在球面上的分布是高度不均匀的，大量的图像块集中在球面上的一小部分区域上，而大部分区域的图像块都非常稀疏。

进一步，G.Carlsson等人研究了高密度区域的局部拓扑结构[5]。在计算了高密度区域图像块的持续同调后，他们指出，高密度高对比度的光学图像块的拓扑符合一个三圆模型，它的0维贝蒂数为1，1维贝蒂数为5，而且这个三圆模型自然地嵌入到克莱因瓶中。然而，无论是三圆模型还是克莱因瓶，它们都是低维模型。注意到数据在一个7维球面上，拟合一个高维模型是可能的。

[26]中包含了预处理过的光学图像块数据。预处理过程如下：将每个图像块视为$\mathbb{R}^9$中的向量，通过去中心化，原始数据被映到一个8维超平面$\mathbb{R}^8 \subseteq \mathbb{R}^9$上。通过离散余弦变换，数据点被变化到一个7维球面上$\mathbb{S}^7 \subseteq \mathbb{R}^8$。定义一个密度函数挑选出前15%密度最大的向量。最终的数据集包含15,000个点。我们将其记为S。



(a) **pr = 0.0110** (b) pr=0.0093 (c) pr=0.0080

(d) pr=0.0071 (e) pr=0.0070 (f) pr=0.0056

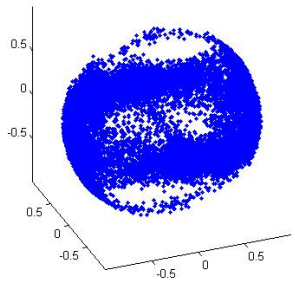
图 4: 前6个后验概率最大的图。后验概率在图的底部给出。

将每个S中的点视为随机向量$\mathbf{Z} = (Z_1, Z_2, \cdots, Z_8)$的样本。对$\mathbf{Z}$决定的图$\mathcal{G}$进行推断。取$u = 2$，在算法2中，置$\delta = 3$，$D = 0.3I$。经过$25,000$迭代的预烧期，取500个样本点。后验概率最大的6个图如图4。

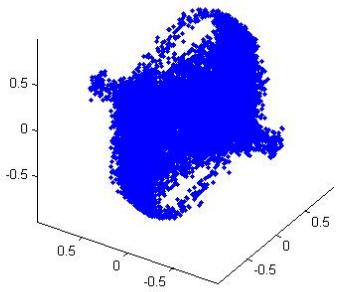
后验概率最大的图表明，S的内蕴空间可能是两个$\mathbb{R}^4$中的子空间的乘积。将S分别投影到前四个分量和后四个分量上得到两个数据集S_1和S_2。分别计算S_1和S_2的持续同调。图、ref结果表明S_1的内蕴空间的同调和S的内蕴空间的同调相同，其0维贝蒂数为1，1维贝蒂数为5。而S_2的持续同调显示，S_2的内蕴空间的同调平凡。通过S_2在$\mathbb{R}^3$的投影可以合理猜测，S_2的内蕴空间为一个实心球6。从而数据集S的内蕴空间为三圆模型和实心球的乘积。注意到实心球是可缩的，所以这个乘积空间同伦等价于三圆模型。它的同调和[5]中的结果相吻合。

6 结语

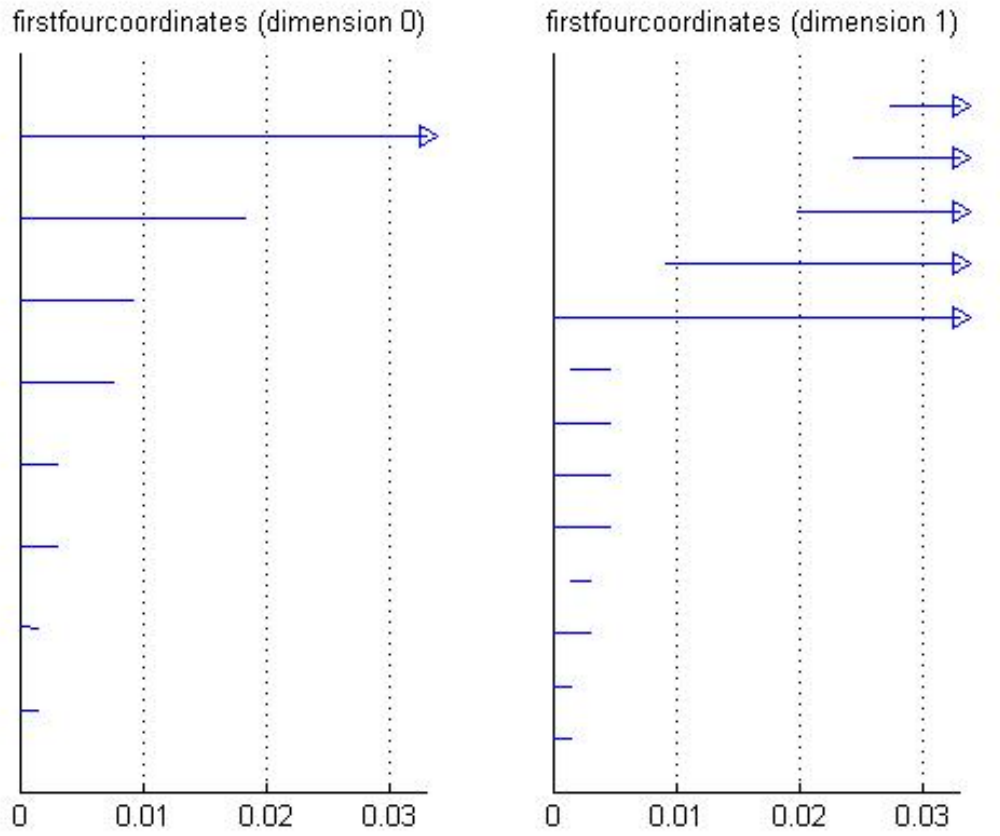
本文提出了用乘积空间对数据集的内蕴空间进行建模的方法。为此使用了持续同调和图贝叶斯的理论。本文介绍了一种蒙特卡洛马尔科夫算法来计算图的后验概率，并对其进行改进，提出了新的更高效的算法。新的算法基于文中证明的一个事实：当图的顶点趋于无穷的时候，不连通的图在整个图空间所占的比例趋于0。将新的算法应用于自然图像统计，我们为高密度，高对比度的光学图像块提出了新的高维的模型。此外文中还给出了域上的Künneth定理的一个线性代数的证明。



(a) S_1在第1，2，3个坐标分量上的投影

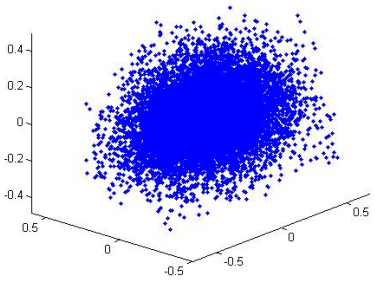


(b) S_1在第2，3，4个坐标分量上的投影

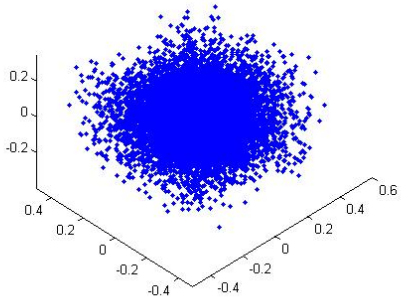


(c) S_1的持续同调

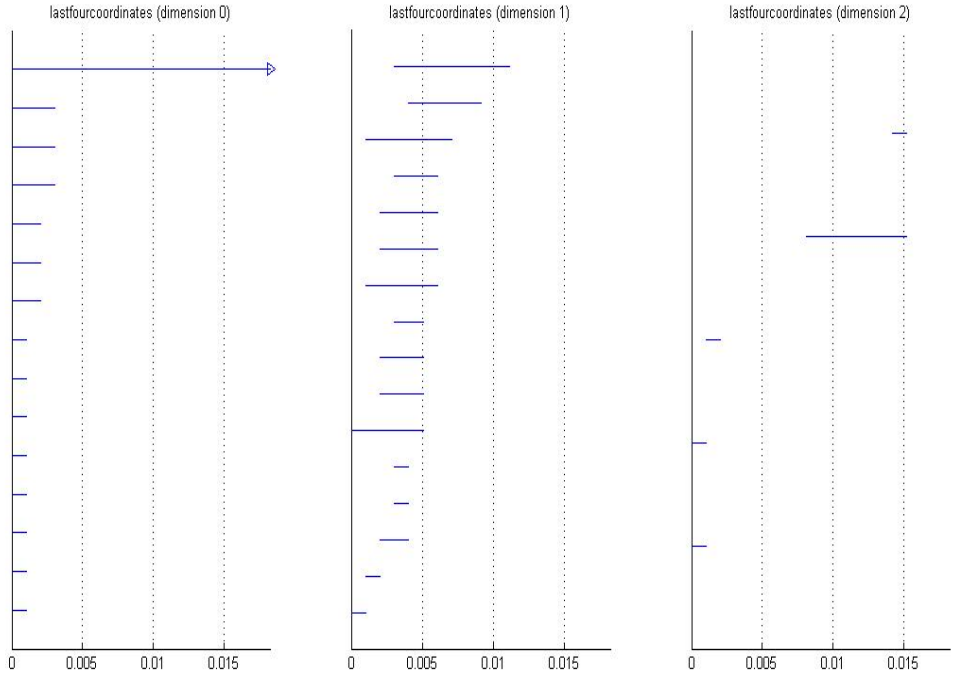
图 5: S_1的投影和持续同调



(a) S_2在第5，6，7个坐标分量上的投影



(b) S_2在第6，7，8个坐标分量上的投影



(c) S_2的持续同调

图 6: S_2的投影和持续同调

致谢

感谢孙华飞老师从大一以来对我的指导和照顾。没有孙老师的鼓励，我没有勇气看完一本又一本艰涩难懂的书籍和一篇又一篇佶屈聱牙的文章。孙老师教给我的不只是知识，还有成为一名合格研究者的素质品质。我应当始终牢记孙老师的教诲。

感谢Assadi教授教给我拓扑数据分析的知识，他一直非常关切我的研究，并且一直给与我帮助。

感谢父母的陪伴和鼓励，他们永远视我为骄傲，尽管有时候我没能做到最好。感谢室友的陪伴，我们一起度过了很多欢乐的时光。

参考文献

[1] Henry Adams and Carlsson Gunnar. On the nonlinear statistics of range image patches [J]. *SIAM Journal on Imaging Sciences*, 2(1):110–117, 2009.

[2] Aliye Atay-Kayis and Massam Hélène. Monte carlo method for computing the marginal likelihood in nondecomposable gaussian graphical models[J]. *Biometrika*, 92(2):317–335, 2005.

[3] Julian Besag. Spatial interaction and the statistical analysis of lattice systems[J]. *Journal of the Royal Statistical Society*, 36(2):192–236, 1974.

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

[5] Gunnar Carlsson, Tigran Ishkhanov, Vin de Silva, and Afra Zomorodian. On the local behavior of spaces of natural images[J]. *International Journal of Computer Vision*, 76(1):1–12, 2007.

[6] Carlos M. Carvalho, Massam Hélène, and West Mike. Simulation of hyper-inverse wishart distributions in graphical models[J]. *Biometrika*, 94(3):647–659, 2007.

[7] David Cohen-Steiner, Edelsbrunner Herbert, and Harer John. Stability of persistence diagrams[J]. *Discrete & Computational Geometry*, 37(1):103–120, 2007.

[8] A. Philip Dawid and Steffen L. Lauritzen. Hyper markov laws in the statistical analysis of decomposable graphical models[J]. *Annals of Statistics*, 23(5):1864–1864, 1995.

[9] Vin De Silva. A weak characterisation of the delaunay triangulation[J]. *Geometriae Dedicata*, 135(1):39–64, 2008.

[10] Vin De Silva and Carlsson Gunnar. Topological estimation using witness complexes[J]. In *Eurographics Conference on Point-Based Graphics*, pages 157–166.

[11] Cecil Jose A. Delfinado and Edelsbrunner Herbert. An incremental algorithm for betti numbers of simplicial complexes on the 3-sphere[J]. *Computer Aided Geometric Design*, 12(7):771–784, 1995.

[12] Herbert Edelsbrunner, Letscher David, and Afra Zomorodian. Topological persistence and simplification[J]. *Discrete & Computational Geometry*, 28(4):511–533, Nov 2002.

[13] Herbert Edelsbrunner and John Harer. *Computational topology : an introduction*[M]. American Mathematical Society, Providence, R.I., 2010.

[14] Samuel Eilenberg and Joseph A. Zilber. On products of complexes[J]. *American Journal of Mathematics*, 75(2):200–204, 1953.

[15] Robert Ghrist. Barcodes: The persistent topology of data[J]. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

[16] Leonidas J. Guibas and Steve Y. Oudot. Reconstruction using witness complexes[J]. *Discrete Comput Geom*, 40(3):325–356, 2008.

[17] Allen Hatcher. *Algebraic topology*[M]. Cambridge University Press, Cambridge ; New York, 2002.

[18] Steffen L. Lauritzen. *Graphical models*[M]. Oxford statistical science series. Clarendon Press ; Oxford University Press, Oxford New York, 1996.

[19] Ann B. Lee, Kim S. Pedersen, and David Mumford. The nonlinear statistics of high-contrast patches in natural images[J]. *International Journal of Computer Vision*, 54(1-2):83–103, 2003.

[20] Hanns-Georg Leimer. Optimal decomposition by clique separators[J]. *Discrete Mathematics*, 113(1-3):99–123, 1993.

[21] Simón Lunagómez, Sayan Mukherjee, Robert L. Wolpert, and Edoardo M. Airoldi. Geometric representations of random hypergraphs[J]. *Journal of the American Statistical Association*, 112(517):363–383, 2017.

[22] James R. Munkres. *Elements of algebraic topology*[M]. Addison-Wesley, Menlo Park, Calif., 1984.

[23] Nina Otter, Mason A. Porter, Ulrike Tillmann, Peter Grindrod, and Heather A. Harrington. A roadmap for the computation of persistent homology[J]. *EPJ Data Science*, 6(1):17, 2017.

[24] Alberto Roverato. Hyper inverse wishart distribution for non-decomposable graphs and its application to bayesian inference for gaussian graphical models[J]. *Scandinavian Journal of Statistics*, 29(3):391–411, 2002.

[25] Gurjeet Singh, Facundo Mémoli, and Gunnar Carlsson. *Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition*[M]. 2007.

[26] Andrew Tausz, Mikael Vejdemo-Johansson, and Henry Adams. JavaPlex: A research software package for persistent (co)homology[J]. In Han Hong and Chee Yap, editors, *Proceedings of ICMS 2014*, Lecture Notes in Computer Science 8592, pages 129–136, 2014. Software available at `http://appliedtopology.github.io/javaplex/`.

[27] Hao Wang and Mike West. Bayesian analysis of matrix normal graphical models[J]. *Biometrika*, 96(4):821–834, 2009.

[28] Afra Zomorodian. *Topology for computing*[M]. Cambridge monographs on applied and computational mathematics. Cambridge University Press, Cambridge, UK ; New York, 2005.

[29] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology[J]. *Discrete & Computational Geometry*, 33(2):249–274, 2005.