分层拓扑聚类

AnaCarpio[1]，Gema Duro[21]数学科学学院´，马德里康普顿斯大学，科学广场 *3* 号，马德里 *28040*，西班牙，以及[2]经济与商业科学学院´，马德里自治大学，托马斯·巴连特 *5* 号，马德里 *28049*，西班牙(日期：2026年1月6日)

arXiv:2601.00892v1 [cs.LG] 31 Dec 2025

摘要

拓扑方法能够在不对数据云的结构作出假设的情况下对其进行探索。本文提出了一种分层拓扑聚类算法，可结合任意距离选择加以实现。由此得到的层次结构可以推断离群点和任意形状簇的持久性。我们在若干选定的数据集上展示了该算法的潜力，这些数据集包含图像、医疗数据和经济数据，其中离群点发挥着重要作用。在其他技术无法有效聚类的情况下，这些方法能够提供有意义的簇。

I. 引言

聚类分析旨在根据对象之间的亲和性将其划分为不同的组。其应用包括模式识别、图像处理、市场调研和数据分析。所有聚类过程都会受到噪声和离群点存在的影响。噪声是随机误差。离群点是数据集中与其他点相比存在显著偏离的任何点。由错误导致的极端值只是噪声，可以予以舍弃；而有意义的离群点则是数据的一部分。它们可能表明存在不同的机制或异常元素。本文提出了一种分层拓扑聚类（HTC）算法，能够识别任意几何形状的簇，并以自动方式区分重要的离群点。我们的算法基于持久同调，这是一种拓扑数据分析工具，用于识别跨尺度持续存在的数据特征。

聚类技术通常以这样的方式对元素进行分组：同一组（一个簇）中的项目彼此之间的距离，比它们与其他组中项目之间的距离更近（在某种意义上）。人们针对不同目的提出了越来越多以不同方式定义簇的聚类策略，综述见 [2, 18]。常见方法基于划分、层次结构、密度或分布。基于划分的聚类算法将数据分成预先设定数量的集合，并力求最小化数据到簇中心的距离。K-中心点算法 [1] 和K均值 [3] 属于这一类。这些方法无法识别非凸簇，因为观测值总是被分配给最近的中心。层次聚类 [1] 依据连接准则构建簇的层次结构；该准则根据集合中观测值两两之间的距离来定义集合间的不相似度。其凝聚变体最初将每个对象视为一个单独的簇。随后，簇两两合并，直至形成一个包含所有对象的最终大簇。这一过程为数据提供了一种树状表示，称为树状图。K均值试图优化全局目标函数（簇的方差）并达到局部最优，而凝聚层次聚类则在每次簇合并时寻找最佳步骤，因此可能得到受诸多任意选择影响的次优解。尽管如此，使用随机初始化的K均值算法多次处理同一数据集时，仍可能得到不同结果。基于分布的算法旨在将观测到的数据结构模拟为预先设定的概率分布函数 [4]的叠加。其效果取决于所提出的概率分布能否重现数据。局部密度

数据点的局部密度被基于密度的带噪声应用空间聚类（DBSCAN） [5]直接利用。该方法将簇分配给密度超过选定阈值的任意形状点云。其余点则被视为离群点或噪声。然而，选择有意义的阈值往往并非易事。较新的聚类算法结合了密度和距离。文献 [6]探讨了一种基于以下理念的方法：簇中心具有较高的密度，并且与其他高密度点保持一定距离。其他新颖的聚类策略则依托图论、网格或网络结构、分形理论、统计或神经网络学习以及核[2, 18]。

这里我们关注基于形状概念的聚类算法。近年来，拓扑方法——即利用数据中的拓扑特征（如连通分支、空洞、隧道等）的方法——引起了广泛关注 [7–9]。将拓扑特征纳入分析的方式各不相同，由此产生了不同的拓扑聚类算法。如今常见的一种做法是用持久图表征点云 [10], ，计算它们之间的Wasserstein距离 [11] ，再应用标准聚类算法，根据相似性对数据集进行分组。应用包括网络聚类[8], 、函数差异度测量 [9], 、时间序列聚类 [12], 、界面、聚集体和模式的聚类 [13–15], ，以及对群集和鸟群结构的研究 [16, 17], ，这里只举几例。我们的目标并非根据数据集或点云的持久图对其进行分组，而是依据基本拓扑特征，对数据集本身的元素进行聚类。因此，我们以不同的方式运用同调信息。文献 [18] 讨论了一个相关思路，并提出了一种不同的计算算法，该算法用于各点具有指定密度的网络。本文介绍另一种算法，它只关注点云及其距离，以计算 H_0 同调[20] ，同时在构建层次结构的过程中跟踪每个分支中的各个元素。与DBSCAN一样，我们提出的算法能够检测任意形状的簇和极端值。不过，它不需要盲目选择参数。正如我们接下来将在选定的数据集上展示的那样，可以从所得层次结构中推断离群点和簇的持久性。与持久同调研究中的常见做法一样，选定一种衡量项目相似性的距离后，我们构造一个嵌套的单参数单纯复形族。然后，利用单纯复形定义聚类层次结构。对于控制参数的每个取值，与该参数对应的单纯复形的不相交分支

定义了任意形状的簇。随着控制参数变化，它们会合并成更大的簇，并演化为该族中的另一个单形。簇和离群点就这样自然地产生。主导簇通常标示出密集区域，而离群点可能代表有意义的稀有元素。这种方法对噪声并不十分敏感。

本文结构如下。首先，我们解释该算法，并通过一个几何示例展示结果：一个碎片化界面，代表恶性细胞对健康组织的侵袭。其他标准算法（K均值、具有最高共表型相关系数的层次聚类、DBSCAN）无法在所考虑的应用中识别出含义明确的簇，而我们的拓扑层次聚类算法能够区分健康组织与癌变组织之间的主界面，以及迁移到健康细胞中的恶性细胞分离岛。我们尤其关注该算法识别有意义离群点的潜力，并在图像分析、基因数据研究和经济学研究中的数据集上对其进行测试。附录汇集了用于比较的聚类算法、图像和基因数据所用的非欧几里得距离，以及图像测试中采用的压缩过程的技术细节。

II. 材料与方法

我们的方法基于从拓扑数据分析，尤其是持久同调中借鉴的思想。持久同调是对点数据云 $[10,\ 19]$ 上的同调的一种改编，它提供了一种方法，可使用任意选定的距离，在不同空间分辨率下计算其拓扑特征。高度持久的特征跨越较宽的空间尺度范围，因此相比于抽样、噪声或参数选择造成的伪影，更有可能代表所研究数据的真实特征 [20]。

a. 分层拓扑聚类。 要找到点数据云 X 的持久同调，首先必须用单纯复形来表示它。选定一种度量d 来测量数据点之间的距离后，我们便生成单纯复形的滤过。滤过是由一系列逐渐增大的嵌套子集组成的序列。Vietoris-Rips 过滤 $[10,\ 19]$ 是一种典型的选择。给定一个有限点云 X，*Vietoris-Rips* 过滤 VR(X, r) 是一个为 $r > 0$ 定义并构造的单纯复形

如下

- 数据点集合 X 即为顶点。
- 给定顶点 x_1 和 x_2，如果距离$d(x_1, x_2) \leq r$，则边 $[x_1, x_2]$ 包含在 VR(X, r) 中。
- 如果高维单形的所有顶点都包含在 VR(X, r) 中，则该单形属于 VR(X, r)。

根据这些条件，只要 $r \leq r'$，就有 VR(X, r) $\subset$ VR(X, r')。构造 Vietoris-Rips 过滤时，随着 r 增大，数据点通过边相连，形成连通分支。对于 r的每个取值，连通分支的数量就是贝蒂数 b_0(r)。如果将连通分支视为簇，这一过程便形成了分层拓扑聚类策略 (HTC)：

- 我们选择距离 d 以及过滤参数 r的范围

$$\left\{0, \frac{r_{\max}}{M}, \frac{2r_{\max}}{M}, \ldots, \frac{(M-1)r_{\max}}{M}, r_{\max}\right\},$$

  例如，其中 $r_{\max} = \max_{x,y \in X} d(x, y)$（点云直径），而 $\frac{r_{\max}}{M} \sim \min_{x,y \in X, x \neq y} d(x, y)$（点之间的最小距离）。

- 对于 $r = 0$，每个数据点本身构成一个组件/簇。
- 对于每个值 r，我们连接距离小于 r 的点，构成边。每组由边连接的点定义为一个簇。
- 当 r 达到点云直径时，所有可能的边都形成了，此时只剩下一个组件/簇。

持续处于孤立状态的点或簇最后才会合并，因此自然可以作为离群点（单个或集体）的候选对象。在指定的过滤值 r处，贝蒂数 b_0(r)给出该层级的簇数。随着 r变化，所有簇构成的整体簇族呈现出簇的拓扑层次结构。对于这一聚类过程，无需使用 Vietoris-Rips 过滤的更高阶层级，H_0 层（同调）就足够了。

***b. HTC*算法。**由于我们只使用 H_0 同调，算法可以用一种直接明了的方式实现。我们提出以下流程：

- 给定由 N 个点组成的点云 $X = \{x_1, \ldots, x_N\}$，计算所选度量 d 下的距离矩阵。

- 计算 $r_{\text{最大值}}$ =最大值$_{x,\ y\in X} d(x,\ y)$ 和 $r_{\text{最小值}}$ =最小值$_{x,\ y\in X,\ x\neq y} d(x,\ y)$。

- 选择 $M = \left[\frac{r_{\text{最大值}} - r_{\text{最小值}}}{}\right]$，并设置步长 $h = \frac{r_{\text{最大值}}}{M} \geq r_{\text{最小值}}$。

- 构建由递增过滤值组成的网格 $r_m = m\,h$, $m = 0,\ 1,\ \ldots\ldots,\ M$。

- 对于 $r_0 = 0$，定义 $N_0 = N$ 个簇，每个簇由一个点构成。

- 对于 $m = 1, \ldots M$– 从簇 C_i, $i = 1, \ldots, N_{m-1}$ 开始，并设置 $N_m = N_{m-1}$。 – 计算簇连接矩阵 $N_m \times N_m$ L_m, $L_m(i, j) = 1$，如果 $d(x_i, x_j) < r_m$，对于某个 $x_i \in C_i$ 和某个 $x_j \in C_j$, $L_m(i, j) = 0$；否则，– 如果 $m = 1$ 或 $m > 1$ 且 $L_m \neq L_{m-1}$，* 取第一个簇 C，将其标记为已访问 C，列出与其连接的簇 C'（$L_m(C, C') = 1$），对于每个未访问的簇 C'，将其标记为已访问，并合并到 C，列出与合并后的簇连接的簇 C''（$L_m(C, C'') = 1$），然后递归重复此过程，直到所有与其连接的未访问簇都合并到 C。* 对于每个未访问的簇 C，重复此过程，直到所有簇都已访问。* 将 N_m 设为最终簇数。* 如果 $N_m = 1$，则停止。

- O输出：每个滤过值对应的簇层次结构及其元素 ue.

计算簇连接矩阵 L_m 的一种简单方法是，先计算点连接矩阵 $N_m \times N_m$ P_m，其定义如下：若 $d(x_i, x_j) < r_m$，则 $P_m(i, j) = 1$，否则为 $P_m(i, j) = 0$。然后，对于所有 $x_{i'} \in C_i$ 和 $x_{j'} \in C_j$，若 $P_m(i', j') = 0$，则 $L_m(i, j) = 0$，否则为 $L_m(i, j) = 1$。该算法能够列出每个过滤值下各个簇的独立元素，并跟踪这些元素所属的簇。图 **1** 展示了这一过程。

该算法的复杂度至多为 MN^2，其下界为 $\Pi_{m=1}^{M} N_m^2$ 阶，其中 N_m是过滤值为 r_m时的簇数。请注意，随着过滤值增大，除非簇连接矩阵发生变化，否则不会执行任何操作。一旦形成单个簇，也不会再执行任何操作。该算法适用于小型和中型数据集。随着数据集规模增大，基于密度的算法可能会成为一种选择 [18]。

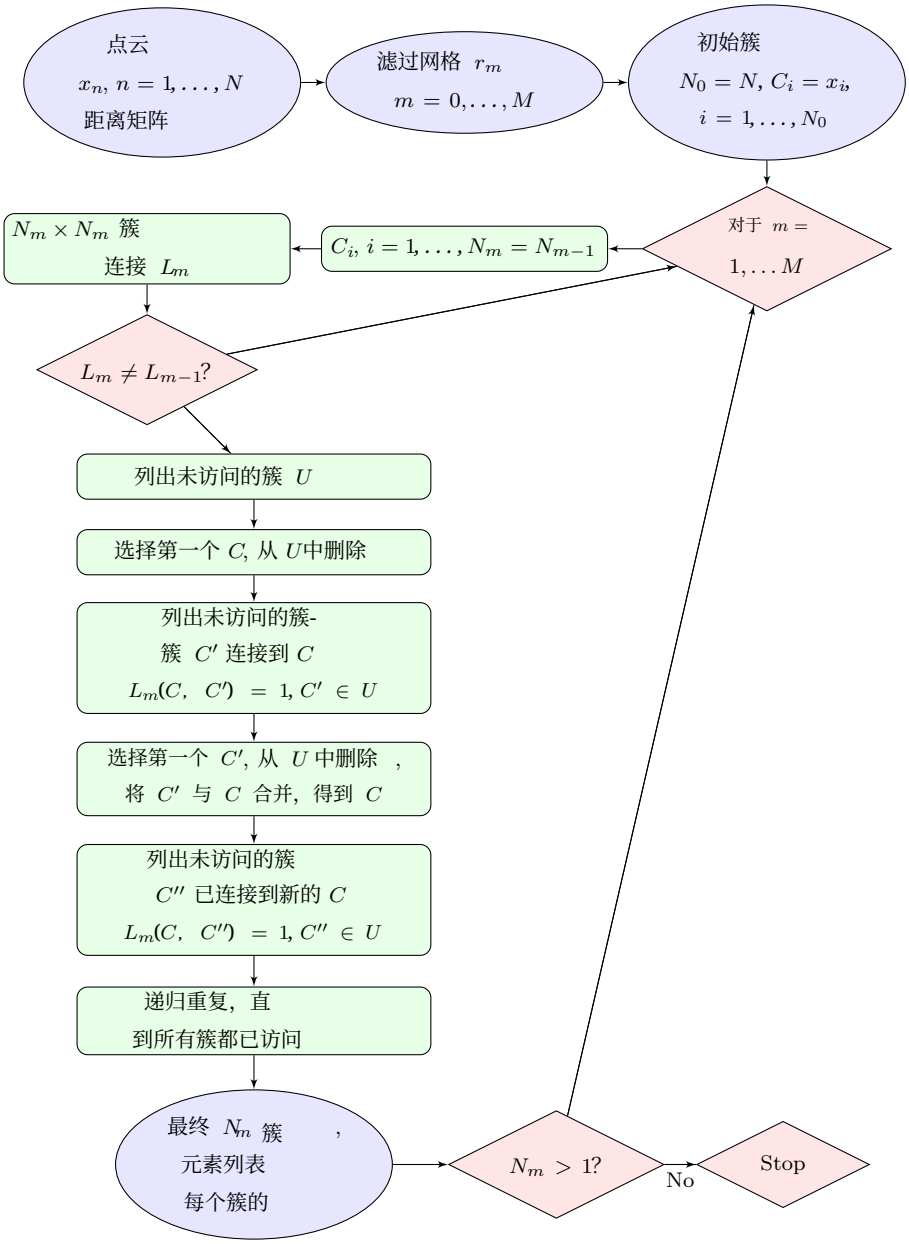

图1：**HTC算法示意图。**

III. 结果与讨论

我们在多个数据集上展示了该算法的性能，这些数据集分别取自几何研究、图像分析、贸易和基因数据集next。

a. 碎片化前沿聚类。 考虑一个采用欧几里得距离的二维示例。图2(f)中的点定义了两种细胞（健康细胞和恶性细胞）之间的界面，该界面从一幅上皮组织图像中提取而来 [13, 21]。我们将上一节所述的HTC策略应用于由这些点构成的数据集。面板(a)–(e)展示了随着过滤参数增大，相应簇层次结构的演化阶段。起初，每个点都是一个簇（未显示）。随着过滤参数 r 增大，我们识别出一个代表两种不同细胞之间界面的主簇（绿色），以及其他代表分离岛的簇，因为其中一种细胞群正在侵入另一种细胞群，见面板(a)。进一步增大 r 后，距离最近的碎片会加入主簇，直到只剩下一个簇。最后加入的簇由已深入健康细胞层的恶性细胞构成。图3展示了随着过滤参数 r增大，簇层次结构的演化，以及主导簇的持久性和大小。根据这些图，我们可以选择感兴趣的过滤值，在这些值下能够观察到若干相关簇，或相关的孤立离群点仍然存在。以这种方式获得的簇可以解释为健康细胞或恶性细胞的群组，以及它们之间的界面。它们与主导簇合并时所处的层次级别（主导簇最初代表两种细胞之间的主界面）提供了有关恶性细胞侵入健康细胞层深度的信息。

在此情境下，HTC 得到的簇具有明确的含义，而其他标准聚类方法则不然。图4(a)-(b)比较了K均值得到的簇，以及采用平均连接法的层次聚类结果；这些方法都无法捕捉上文所述的几何解释。与单连接、完全连接和加权连接等其他标准连接策略相比，对于所研究的数据，平均连接法的共表型相关系数最高。图4(c)展示了DBSCAN算法在特定最小距离 ε（用于连接点）和最小点数 M_p （用于形成一个簇）的设定下得到的结果。在这组超参数下，红色点未被分配到任何簇中。经过反复试验，我们发现了其他超参数设定，可得到与HTC自动生成并在图2中展示的簇排列相似的簇

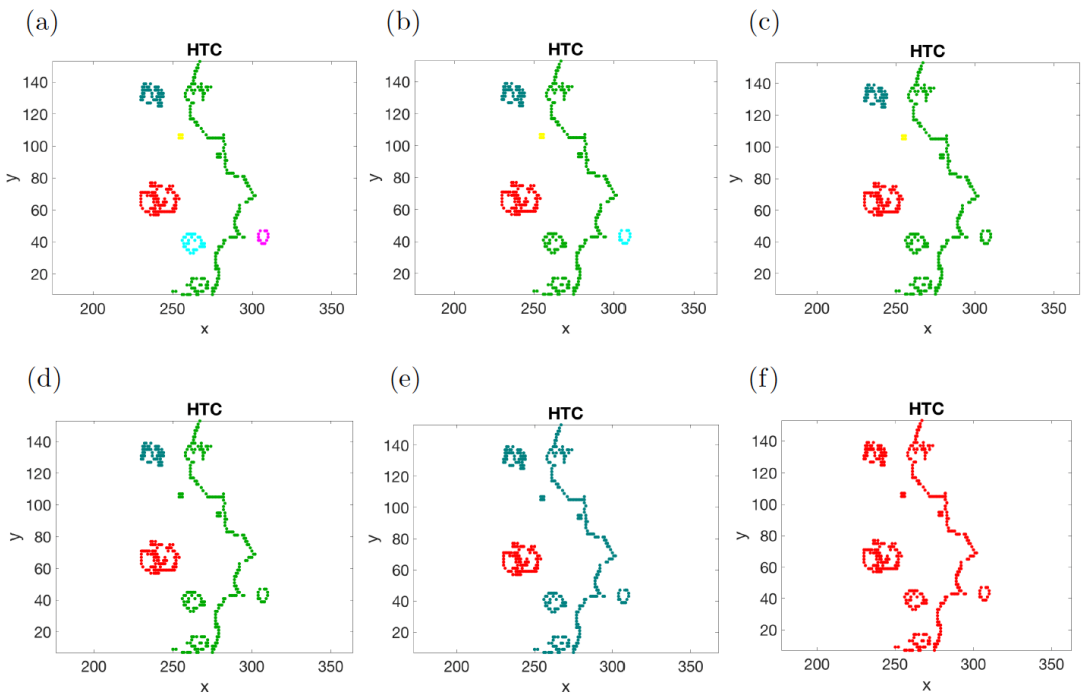

图 2：**应用于碎片化前沿的分层拓扑聚类**

用于区分恶性细胞和健康细胞。在 (a) $r = 5.2$、(b) $r = 8.8$、(c) $r = 9.6$、(d) $r = 11$、(e) $r = 14.7$ 和 (f) $r = 16.9$ 中得到的簇。最后与主簇合并的簇可视为集体离群点。它们代表侵入健康细胞的恶性细胞岛，而主簇则代表这两类细胞群体之间的界面。

每个参数提供的簇分布与 HTC 自动获得并在图2中表示的簇排列相似。

然而，我们很难预先判断哪些参数选择能使 DBSCAN 得到描述健康细胞与恶性细胞之间界面，以及侵入健康组织的恶性细胞脱落碎片的簇和离群点。相比之下，图 5 中的条形码清晰地标示了可实现的簇数量，以及预期得到特定数量簇时的过滤值范围。我们的算法将条形码所提供的数据拓扑结构信息，与各过滤值下构成每个簇的点的信息结合起来。

b. 图像处理中的质量评估 现在，我们来测试这一图像处理质量评估流程。我们考虑一系列通过缓慢变化图像压缩程度生成的图像

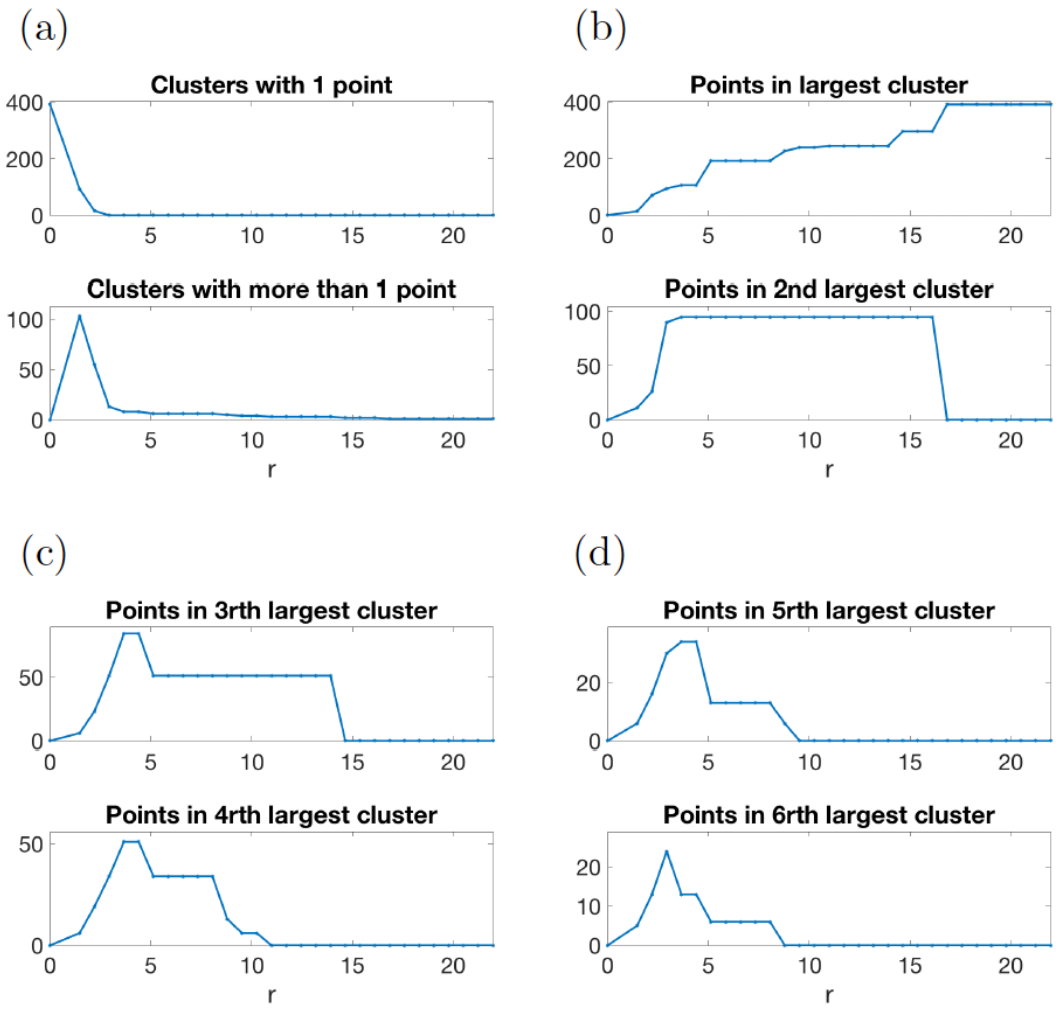

图 3：**拓扑簇随过滤参数的演变** r。我们可以识别出哪些类型的簇在较大的 r 范围内持续存在。

。图 6(a)-(l) 展示了图像序列中的一些元素。图 (a) 为原始图像（图像 1），而图 (b) 则包含一个额外物体：一条黑线（图像 32）。图 (c)-(l) 对应于原始图像，其压缩程度由参数 k 指定，即压缩图像中保留的奇异值数量 [22]。对于图像 2 至 30，该参数取值为 $k = 10 + 5q$, $q = 0,$ 1.....28。图 (l) 在图像 30 中添加了一条黑线，从而得到图像 31。我们使用 Wasserstein距离 $W^{1,\,2}$ 来比较这 32 幅图像；关于图像 [23], 的详情，请参见附录。该距离尤其适用于图像比较，因为它比较的是模式，而欧几里得距离比较的是局部偏差。

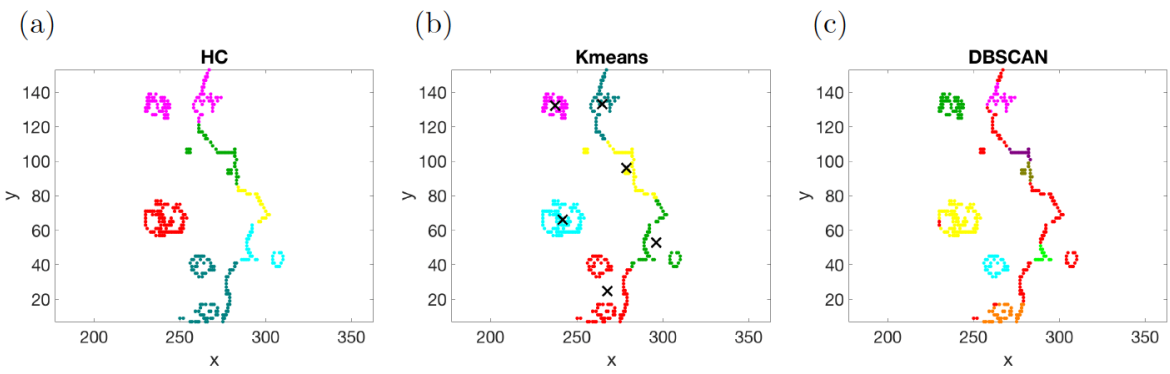

图 4: **层次聚类、K均值和 DBSCAN 应用于**

碎片化前沿。 颜色表示：(a) 通过平均连接法进行层次聚类得到的六个簇，以及 (b) 通过 K 均值得到的六种簇排列之一（可能存在不同的簇排列）。面板 (c) 展示了 DBSCAN 在 $M_p = 10$ 和 $\varepsilon = 5$下的结果；红色点未被分配到任何簇，因为它们附近没有足够多的邻居。请与图 1(a) 所示的簇分布进行比较。HTC 提供的簇的几何解释，即将其视为主界面和分离岛，已不复存在。

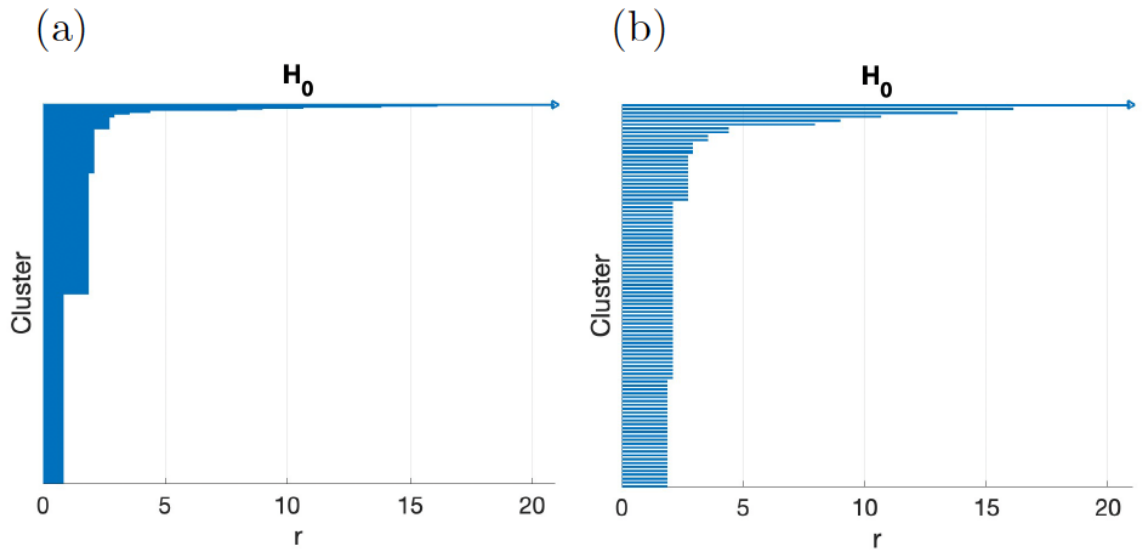

图 5: **碎片化前沿的 H_0 同调持久条形码**

数据集。完整条形码 (a) 与持久性较弱的簇的放大视图 (b)。这种表示以一种盲目的方式展示了簇随过滤参数变化的动态过程：我们不知道哪些元素属于每个簇。

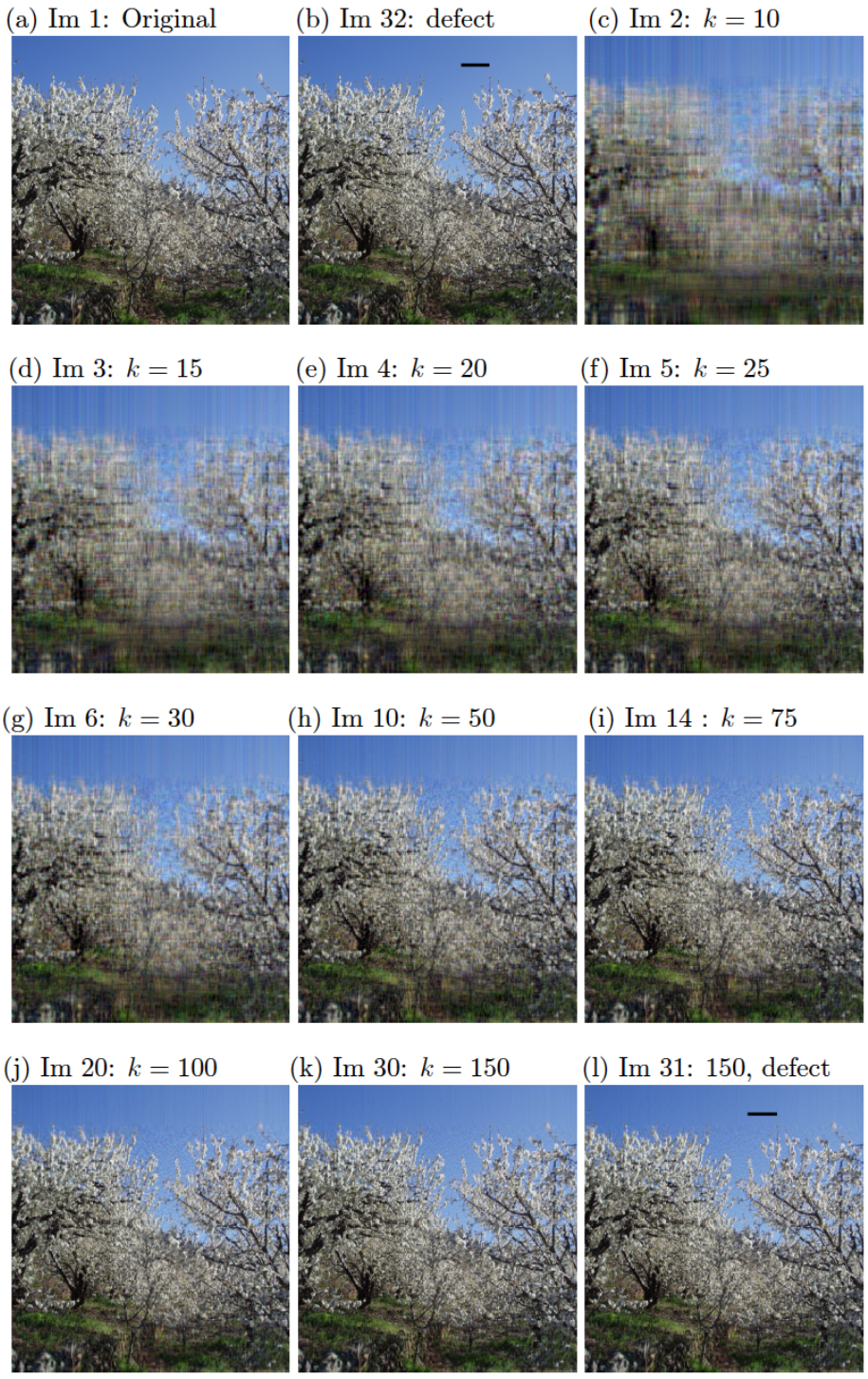

图6：**图像测试中压缩程度逐渐降低并最终出现缺陷的一系列图像**。参数 k 表示压缩图像中纳入考量的奇异值数量。

图 7(a) 中的热图表示 Wasserstein距离矩阵。采用标准层次聚类算法和平均连接法，我们得到图 7(b) 所示的树状图。图 8(f)-(j) 展示了图像分别分为 4、5、7、9 个簇时的分组情况。拓扑层次聚类产生了不同的簇和离群点排列，如图 7(c) 的树状图及图 8(a)-(e) 各子图所示。图 9 展示了随着过滤值增加，簇的动态变化。通过选择合适的过滤参数，我们发现了子图 8(a) 中标示的 9 组。原始图片（图像 1）和两张含有黑线的图片（图像 31 和 32）是离群点，压缩程度很高的图像 2-6 也是如此。其余图像构成一个簇，其中的图像清晰可辨、压缩程度适当且没有线状缺陷。随着过滤参数增大，一些压缩图像加入主簇，见子图 8(b)-(c)。接着，两张带黑线的图像在图 8(d) 中分别并入不同的簇。未压缩的那张与原始图像组成一组，而压缩的那张则加入压缩图像的主簇。

图 7(c) 显示，这两个簇——由未压缩图像 1 和 32 组成的簇，以及由中度压缩图像 5-31 组成的簇——在一系列过滤值范围内持续存在。图中还显示，带有线条缺陷的图像与压缩程度相似的其他图像有所区别，因为当过滤值较高时，它们才会与这些图像合并。这两个簇在图 8(e) 中合并，只留下较模糊的图像作为持久离群点。这一层次结构使我们能够自动找出因添加额外元素（线条）而受损的图像，以及分辨率较低（压缩过度）的图像。HTC 最初将压缩图像 31 与不带线条的压缩图像归为一类，并将未压缩图像 32 与未压缩图像 1 归为一类，这反映出：所选压缩级别造成的整体图像特征失真，对数据集中图像间距离的影响，比线条的存在更大。花朵、树枝、土壤和天空的许多细节都在大片区域内变得模糊。相比之下，图 7(b) 和图 8(f)-(j) 所示的采用完全连接法进行层次聚类得到的簇，更侧重于压缩级别。对于此数据集，完全连接法得到的共表型相关系数最高。图 8(f) 中，包含线条缺陷的压缩图像 31 和未压缩图像 32 已与原始图像 1 归为一类。拓扑层次聚类有望成为图像质量控制的实用工具，

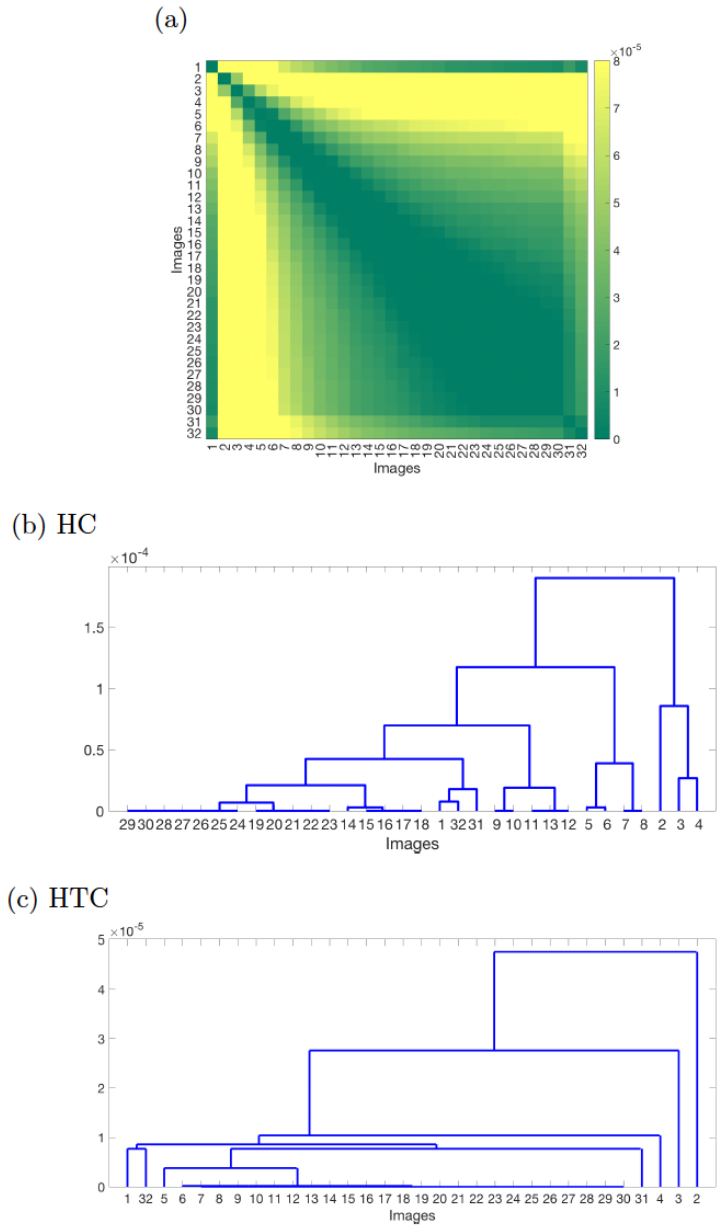

图7：**图像测试中，具有最高共表型相关系数的层次聚类与拓扑层次聚类的比较**。

Wasserstein距离矩阵 (a)，以及采用完全连接法进行层次聚类 (b) 和拓扑层次聚类 (c) 得到的树状图。虽然 (c) 区分了由未压缩图像 1 （原始图像）和 32 （带线条缺陷的原始图像）组成的簇，以及由压缩图像 5-30（无缺陷）和 31（带线条缺陷）组成的簇，但 (b) 将图像 1、31 和 32 归入同一个簇。

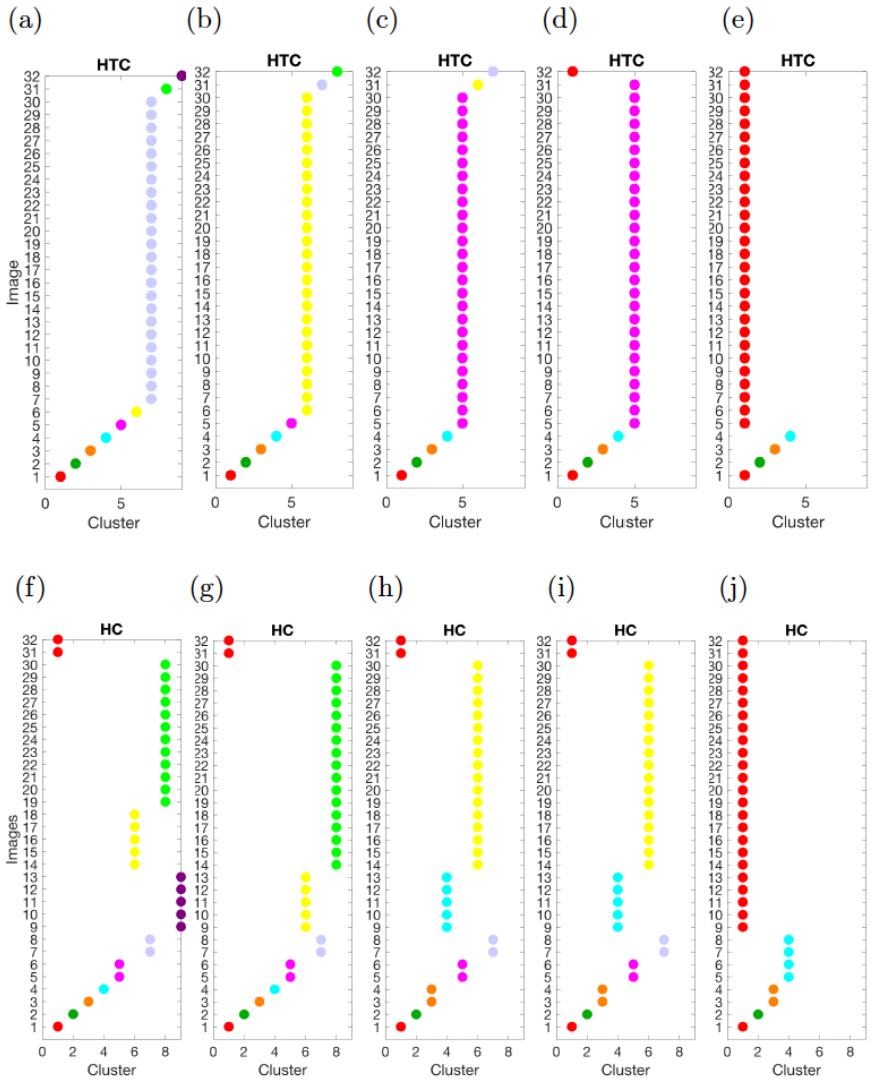

图 8：**完全连接法层次聚类与拓扑系数的比较**

用于图像测试的层次聚类。通过拓扑层次聚类（a）-（e）和标准层次聚类（f）-（j）得到的图像分组，其中簇的数量各不相同。HTC 将压缩程度相似的图像归为一个主簇，同时还能识别出由（d）中的两张未压缩图像组成的次级簇（红色）。含有线状缺陷的未压缩图像（32）和压缩图像（31），分别与（d）中其他未压缩图像（红色）和压缩图像（洋红色）归为一类。相比之下，采用余弦相关性最高的连接方法时，HC 会将这两张图像与原始图像 1 归为一类，见（f）-（i）中的红色簇。

对于此数据集。

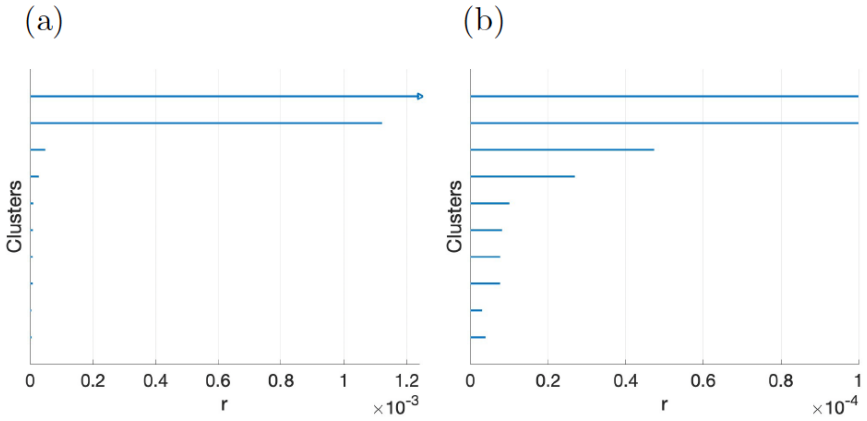

图 9：**的碎片化前沿数据集的同调**持久条形码。完整条形码 (a) 与初始簇的放大视图 (b)。

在监测图像中突发的局部变化时。

c. 经济数据集中的离群点 离群点在社会科学和经济学中尤为重要。例如，我们可以利用这些技术来筛选最适合某个职位的人选。这里我们考察国际商务发展贸易统计 [24] 的数据。这个开放获取数据库包含国际商务发展贸易统计，涵盖所有国家按月、按季度和按年的贸易数据：进口额和出口额、贸易量、增长率、市场份额等。我们特别考察固定国家西班牙与欧洲地区所有国家之间贸易中所有产品（约 100 个类别）的进出口额。以 2019年 43 个国家的数据为例。

首先，我们对每种产品 m, $m = 1, \ldots, M$ 的数值 $g_{n,m}$ 进行标准化：减去 N 个国家的均值 μ_m ，再除以标准差 σ_m：$\tilde{g}_{n,m} = (g_{n,m} - \mu_m)/(3\sigma_m)$, $n = 1, \ldots, N$, $m = 1, \ldots, M$。接下来，我们使用出口额或进口额的标准化数值 $\tilde{g}_{n,m}$，计算国家 c_n 和 c_k 之间的欧几里得距离：$d(c_n, c_k) = (\sum_{m=1}^{M} |\tilde{g}_{n,m} - \tilde{g}_{k,m}|^2)^{1/2}$ $n, k = 1, \ldots, N$。应用 HTC算法后，我们得到图 10 所示的离群点。在这两种情况下，大多数国家都形成一个大簇，而少数主导离群点则持续存在，并随着控制层次结构的过滤参数 r 增大而依次与其合并。

关于2019年西班牙对其他欧洲国家的出口，首先出现了一个几乎没有接收出口的国家簇，其中有18个国家的 $r \sim 10^{-1}$较低：

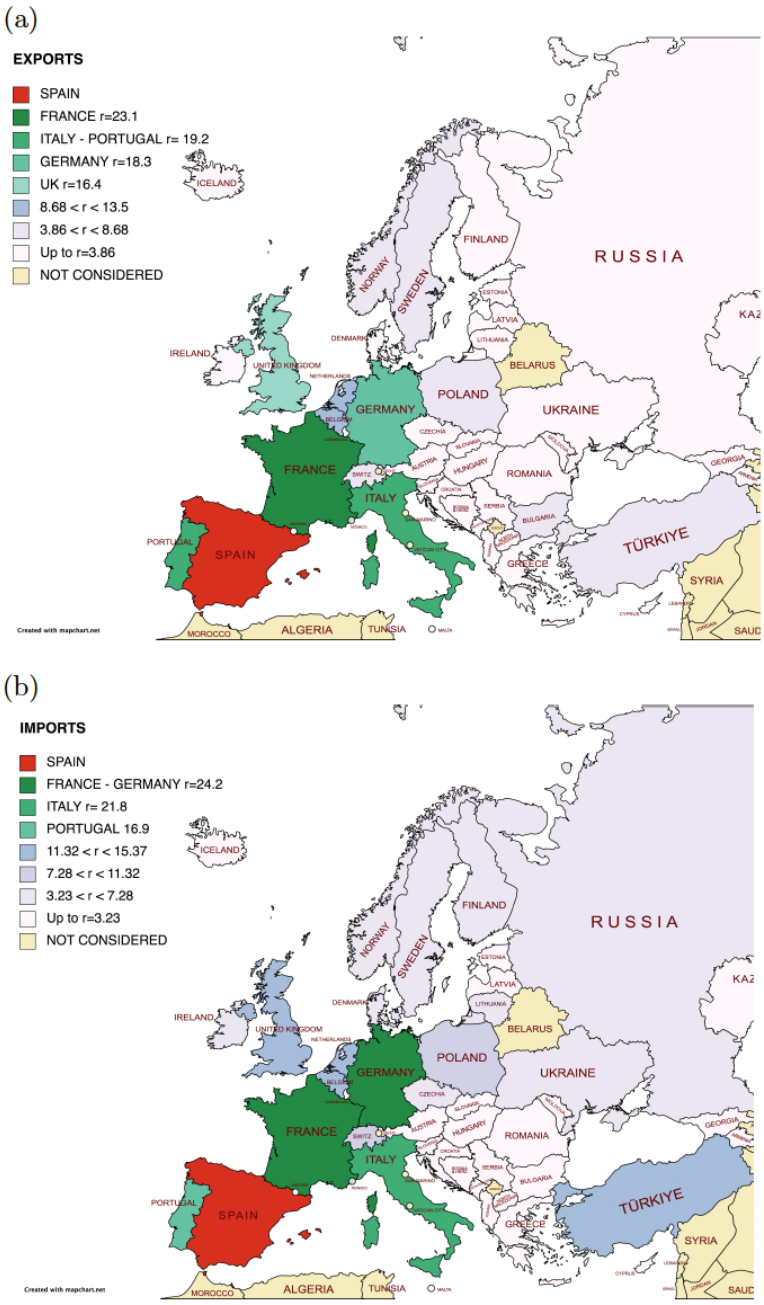

图10：**2019年西班牙贸易中的相关与无关伙伴**。欧洲层面西班牙（a）出口和（b）进口中低交互度及占主导地位的伙伴（离群点）簇，代表西班牙经济的主要供应方和接收方。规模和邻近程度似乎是相关因素。由 https://www.mapchart.net/ 可视化。颜色表示过滤参数 r 的范围，在该范围内，有色国家与主簇合并。 r 随着出口（相应地，进口）水平变得更加重要而增加

阿尔巴尼亚、亚美尼亚、波斯尼亚、格鲁吉亚、冰岛、哈萨克斯坦、马其顿、摩尔多瓦、黑山、塞浦路斯、爱沙尼亚、拉脱维亚、立陶宛、马耳他、塞尔维亚、斯洛文尼亚、卢森堡、乌克兰。在$r = 3.8$时，这个低水平进口国簇由 30 个国家组成，在图10(a)中以浅蓝色标示。西班牙产品的主要进口国是持久离群点，它们在 r的高值处与主簇合并。法国成为主要进口国，意大利和葡萄牙基本并列第二，德国位居第三。

观察西班牙进口时，在较低的 $r \sim 10^{-1}$下，几乎没有进口的初始簇由 14 个国家组成：阿尔巴尼亚、安道尔、波斯尼亚、黑山、马其顿、克罗地亚、塞浦路斯、格鲁吉亚、爱沙尼亚、拉脱维亚、马耳他、摩尔多瓦、塞尔维亚、斯洛文尼亚。在 $r = 3.2$时，这个低水平出口国簇包含 24 个国家，在图10(b)中以浅蓝色标示。德国和法国成为对西班牙的主要出口国，意大利紧随其后。在前一种情形中，意大利和葡萄牙形成一个由两个成员组成的簇；在后一种情形中，德国和法国也形成这样的簇，然后再与主簇合并。

使用K均值时，簇数量的选择可能并不自然，并会以任意方式强行将国家归为一组。具有最高共表型相关系数的层次聚类过程对应于平均连接法，该方法将簇间距离定义为簇中各点之间距离的平均值。各国加入主簇的顺序有所不同。例如，尽管HTC明确区分了德国和法国这两个主要进口伙伴，并使它们在最高过滤值之前始终与其余国家隔离，但使用平均连接法进行层次聚类时，意大利最后才并入主簇。使用平均距离对这一事实进行解释较为晦涩，不像对相互独立的连通分支进行解释那样直观。

d. 基因数据分析 基因分析提供了具有挑战性的数据集，可用于检验聚类过程。我们在此考察来自cBio癌症基因组学开放访问门户 [26],的数据，更确切地说，是来自乳腺浸润性癌 TGCA泛癌图谱数据集 [27, 28]的mRNA基因表达数据。该数据集包含1082个癌症样本中20531个基因的mRNA基因表达测量值，同时还包含114个健康样本的数据。

我们希望分析癌症样本中的基因表达。由于每个基因的测量值量级可能存在很大差异，因此，使用某些参考值对癌症样本的数据进行归一化是很有必要的。一种标准做法是利用所有可用健康样本的数据，确定健康组织中每个基因表达的均值和标准差，然后利用这些数据对数据进行归一化

癌组织中的数据。对于每个基因 m，我们根据现有的114个健康样本计算mRNA基因表达值的均值 μ_m 和标准差σ_m。然后，对于每个癌症样本 n，通过减去均值 μ_m 并除以标准差 σ_m，对每个基因 m的表达值$g_{n,m}$ 进行归一化：$\tilde{g}_{n,m} = (g_{n,m} - \mu_m)/(3\sigma_m)$， $n = 1, \ldots, N$，$m = 1, \ldots, M$。接下来，我们计算归一化基因表达$\tilde{g}_{n,m}$之间的欧几里得距离和费马距离 [25]，详细信息见附录。我们提取了参与细胞周期基因/蛋白质相互作用网络 [29]的 $M = 71$ 个基因。将HTC算法应用于癌症样本后，我们发现了一组显著的离群点，例如CCNE1、SMC1B、CDKN2A、CDC6、PKMYT1和CDK1。事实证明，它们都与乳腺癌预后相关，或已被单独确定为抗癌治疗靶点 [30–35]。图 11按照我们所研究基因与一个主簇合并的顺序，列出了这些基因；随着过滤参数 r 在对数尺度上增大，它们依次合并。图12展示了标准H_0 同调图。请注意，这些基因几乎是按顺序合并到主簇中的。最后合并的基因是显著的离群点。

共表型相关系数最高的层次聚类过程对应于加权连接法。该方法以递归方式定义簇间距离，即对先前合并形成该簇的各簇之间的距离取均值，见附录。该方法生成的树状图以对数尺度绘于图 13 中。与图 11 相比，它突出了在此语境下意义不明的中间子簇；在 HTC 中，这些子簇表现为在相似滤过水平上聚合的基因群，但同时也淡化了占主导地位且有意义的离群点。

IV. 结论

我们提出了一种基于初等持久同调计算的层次聚类算法。选定距离函数后，距离阈值控制点之间的簇形成。在一个簇中，所有元素都可以通过一条元素路径连接，且路径上相邻元素之间的距离小于阈值。由此得到的簇可以呈现任意形状。随着阈值增大，只要现有簇中有两个元素变得足够接近，这些簇就会合并。这种基于形状的方法将持久离群点识别为远离其他所有元素的元素。对于这些元素，需要取较大的阈值才能使其并入现有簇。我们的方法不同于其他拓扑学研究

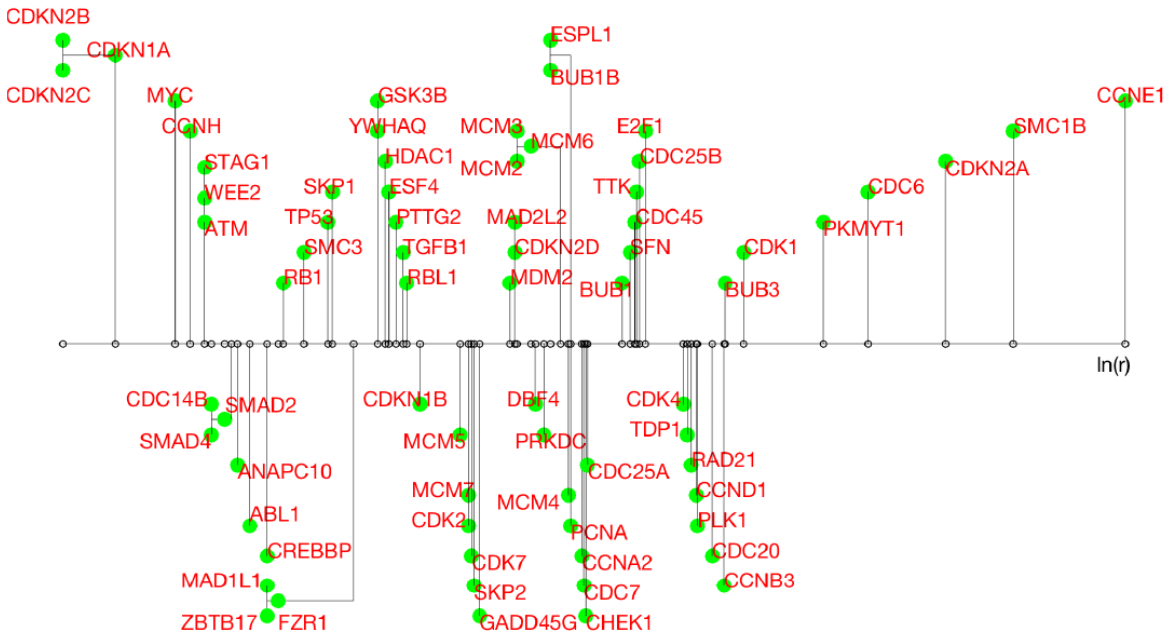

图11：**使用 TGCA 乳腺癌数据对参与细胞周期的基因进行 mRNA 基因表达 HTC 分析。**该示意图列出了细胞周期中的基因，展示它们随着过滤参数 r 增大而并入主 HTC 簇的过程。最后并入的基因是持久离群点，人们认为它们在癌症过程中发挥着相关作用。请注意对数尺度：大多数基因聚为一个簇，而少数主导离群点则持续存在。

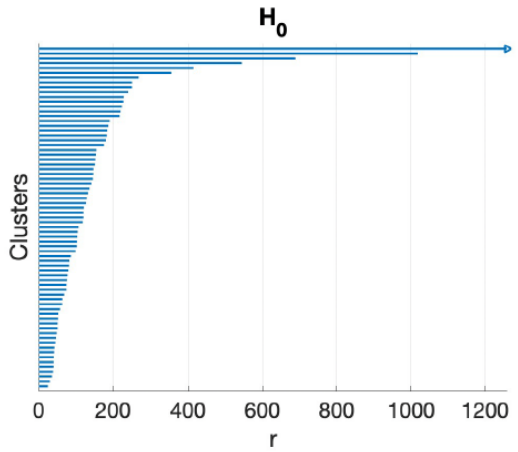

图12：**参与细胞周期的基因的 mRNA 基因表达数据的 H_0 同调持久条形码。**

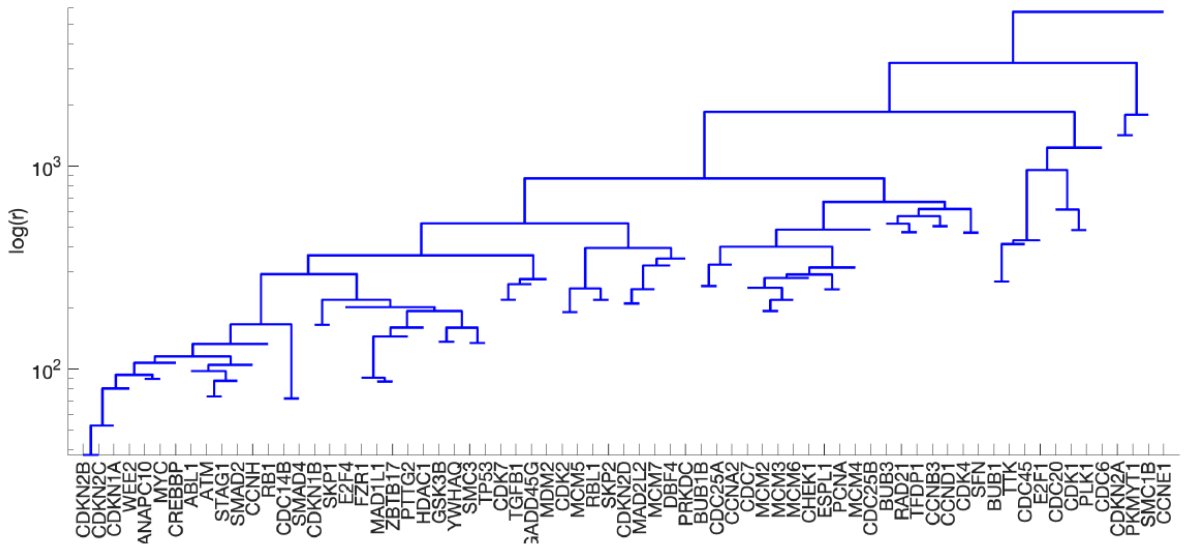

图1：对参与细胞周期的基因的mRNA基因表达数据进行加权连接法层次聚类与HTC中观察到的顺序合并相比，这里形成了中间簇。

通过瓶颈距离或其他距离计算并比较持久图，然后采用标准层次聚类算法。

我们提出的方法具有其他标准聚类过程所不具备的几何解释。通过研究分隔两类细胞的碎片化界面，并将不同聚类策略给出的分组结果进行比较，我们证实了这一点。我们提出的分层拓扑聚类方法能够自动区分主界面和分离岛，而 K均值、标准层次聚类，甚至盲目使用的 DBSCAN 都无法做到这一点。对一组图像的进一步研究表明，在 HC 无法区分的情况下，HTC 能够区分压缩图像与未压缩图像、有缺陷图像与无缺陷图像。我们的方法能够检测有意义的集体离群点（碎片化界面测试中的分离岛），也能检测单个离群点（过度压缩的图像）。我们展示了该方法在贸易数据研究中的用途（构建进口商和出口商的层次结构，以呈现关键合作伙伴），以及在癌症基因数据分析中的用途（识别值得进一步专门研究的持久离群点）。

致谢。研究工作得到西班牙 MICINN 资助，项目编号为 PID2020-112796RB-C21 和 PID2024-155528OB-C21。

作者贡献。 AC 设计了研究、算法和代码。AC 和 GD 开展了数据研究。AC 撰写了论文。所有作者均已阅读并批准最终稿。

数据和材料的可获取性。 经济学研究的开放获取数据可从 ITC 市场分析贸易地图 https://www.trademap.org 获取。乳腺癌基因研究的开放获取数据可从癌症基因组学 cBioPortal 获取，网址为 http://www.cbioportal.org/study/summary?id=brca tcga pan can atlas 2018。图像和正面的数据可从手稿图像中提取。代码使用 Matlab 编写，如有合理请求，可提供。

利益冲突。 作者声明不存在任何竞争性利益。

附录 A：聚类算法

简要概述聚类算法比较所依据的基础 .

K均值将数据分组，以最小化簇内变异总量，从而衡量簇紧凑度 [3]。给定维空间中的一个数据云 $\mathbf{x}_i = (x_i,1, \ldots, x_{i,M})$，簇内总变异为 $\sum_{j=1}^{J} W(C_j) = \sum_{j=1}^{J} \sum_{\mathbf{x}_i \in C_j} d(\mathbf{x}_i, \boldsymbol{\mu}_j)$，其中 C_j 是这类点构成的一个簇， $\boldsymbol{\mu}_j$ 是簇质心。每一项 $W(C_j)$ 表示一个簇内的变异。这里，距离 d 表示欧几里得距离 $d(\mathbf{x}_i, \boldsymbol{\mu}_j)^2 = \sum_{m=1}^{M} (x_{i,m} - \mu_{j,m})^2$.。K均值算法按以下步骤进行。我们先确定要形成的簇数 k，并通过随机生成 k 个点来初始化质心$\boldsymbol{\mu}_j$。接着，将每个数据点 $\mathbf{x}_i$ 分配给欧几里得距离最小的质心。在每个簇中，将簇内各点的平均值设为新的质心。重复这些步骤，直到簇不再变化。K均值需要一个超参数才能运行：簇数。肘部法或轮廓系数法等特定准则可用于提出一个暂定的簇数。

层次聚类 会生成多层次结构，其中某一层的簇会在下一层合并 [1]。凝聚算法从与数据点数量相同的簇开始。相邻的簇会合并成更大的簇，直到所有数据点构成一个簇。这个过程用树状图表示，该图展示了簇如何

不断合并，直至形成一棵包含所有数据点的树，见图 5(b)。为评估簇之间的接近程度并将其合并，算法会使用“连接函数”：“单连接”“平均连接”“完全连接”“加权连接”“质心连接”“中位数连接”和“Ward连接”，不过“质心连接”“中位数连接”和“Ward连接”通常与欧几里得距离搭配使用。连接函数和距离都是需要选择的超参数。本文根据测试用例选取不同的距离，并针对这些数据集采用能使共表型相关系数最大的连接方法；该系数表示原始距离与共生距离（簇合并时的高度）之间的相关性。相关性越高，说明树越能代表数据集。剩余的超参数——高度——决定簇的数量。在不同高度处剪切树状图，即可选取特定数量的簇。在所讨论的示例中，视数据集而定，共表型相关系数最高的连接方法为“完全连接”“平均连接”或“加权连接”。给定两个由元素x_{j_1}、……、x_{j_J}和x_{i_1}、……、x_{i_I}构成的簇，完全连接法将簇间距离定义为元素之间的最大距离，即 $\max\{d(x_{j_k}, x_{i_\ell})\ k = 1, \ldots, J, \ell = 1, \ldots, I\}$。平均连接法将簇间距离定义为 $\frac{1}{IJ}\sum_{k=1}^{J}\sum_{\ell=1}^{I} d(x_{j_k}, x_{i_\ell})$。加权连接法以递归方式定义簇间距离。假设我们通过合并最小的簇 C' 和C''创建簇 C。那么，另一个簇 Q 与 C 之间的距离，是 Q 与 C' 之间的距离和 Q 与 C''之间的距离的平均值：$d(C, Q) = (d(C', Q) + d(C'', Q))/2.$

—

基于密度的含噪声应用空间聚类(DBSCAN) 算法 [5] 会定义高密度区域中的簇，并将低密度区域中的观测值排除在外，这些观测值最终会成为异常。该过程从一个尚未分类的任意数据点开始。我们找出距离小于 ε （ ε邻域）的点。如果其中包含的点数超过最小值 M_p，就用这些点创建一个新簇。否则，我们将该点视为噪声。不过，该点之后可能会落入另一个包含足够多点的点的 ε邻域中，从而归属于该簇。如果没有，它就仍然是一个离群点。因此，该算法能够识别非凸簇和离群点。然而，为两个超参数 ε 和 M_p找到合适的取值并非易事，因为这在很大程度上取决于数据集结构。原则上，也需要选择距离度量，但我们固定使用欧几里得距离。

附录 B：距离

这里介绍一些用于处理图像和数据的距离度量。

Wasserstein 距离。 [23]提出了用于计算定义在网格上的分布之间 Wasserstein-1 距离的快速算法。该计算被表述为最优传输问题。最优传输在许多领域都发挥着重要作用，包括图像处理和机器学习。给定两个二维概率分布，或两幅图像 ρ^0和 ρ^1，其目标是寻找从一个到另一个的传输方案 $f(\mathbf{x})$，使成本最小

$$\min_f \int_{\mathbf{x}} ||f(\mathbf{x})||_p \, d\mathbf{x}$$

使得在零通量边界条件下，散度$_h(f) = \rho^0 - \rho^1$。此处， p可以取 1、2 或无穷大，因此，$||f(\mathbf{x})||_p$分别是 $f(\mathbf{x})$ 的 1、2 或无穷大范数；h 是网格步长，散度$_h$ 是定义在网格上的散度算子，详情请参见 [23] 。

费马距离。考虑一个由列（或行）组成的集合 S $\mathbf{m}^1, \mathbf{m}^2, ..., \mathbf{m}^L$，其中任意两列（或行）之间相对于该集合的$Fermat\ \alpha$距离为 [25]

$$d_{S,\alpha}(\mathbf{m}^1, \mathbf{m}^2) = \min \left\{ \sum_{\ell=1}^{k-1} \|\mathbf{y}^{\ell+1} - \mathbf{y}^{\ell}\|_2^{\alpha} \middle| (\mathbf{y}^1, \ldots, \mathbf{y}^k) \text{ path from } \mathbf{m}^1 \text{ to } \mathbf{m}^2 \text{ in S} \right\},$$

对于任意 $\alpha > 1$。当 $\alpha = 1$时，我们会得到欧几里得距离。费马距离比较集合中的各项，并对同一集合中所有其他项的信息进行加权；当我们希望比较基因谱，同时对患者队列中的信息进行加权时，这一点很有意义。

附录 C：图像压缩

这里详细介绍用于生成我们所比较图像的压缩过程。给定一个 $m \times n$ 矩阵 $\mathbf{M}$，其奇异值分解形式为 $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^t$，其中 $\mathbf{\Sigma}$是一个对角线上包含非负实数的 $m \times n$矩形对角矩阵，$\mathbf{U}$是一个 $m \times m$酉矩阵，而$\mathbf{V}$是一个 $n \times n$酉矩阵。 $\mathbf{\Sigma}$的对角元素 σ_{ii}是矩阵 $\mathbf{M}$ 的奇异值，而$\mathbf{U}$ 的列和$\mathbf{V}$ 的列分别是 $\mathbf{M}$ 的左奇异向量和右奇异向量。压缩策略按奇异值从大到小排序，并在分解中仅保留其中的 k 个，也就是说，我们令 $\mathbf{M}_k = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^t$，其中 $\mathbf{\Sigma}_k$ 仅包含最大的 k 个奇异值，而

$\mathbf{U}_k$ 和$\mathbf{V}_k$ 仅保留与这些奇异值对应的左奇异向量和右奇异向量[22]。

---

[1]L. Kaufman, PJ. Rousseeuw,《数据中的群组发现：聚类分析导论》，威利概率与统计丛书，John Wiley & Sons，纽约，2005。http://dx.doi.org/10.1002/9780470316801

[2] D. Xu, Y. Tian，《聚类算法综合综述》，《数据科学年刊》，**2**(2015),165-193。http://dx.doi.org/10.1007/s40745-015-0040-1

[3] J. MacQueen，《多变量观测分类与分析的一些方法》，载于：第五届伯克利数学统计与概率研讨会论文集，第1卷：统计学，281-297，加利福尼亚大学出版社，伯克利，CA，1967。可从 http://projecteuclid.org/euclid.bsmsp/1200512992 获取。

[4]G.J. McLachlan、T. Krishnan，期望最大化算法及其扩展，《概率与统计Wiley丛书》**382**，John Wiley & Sons，纽约，2008年。http://dx.doi.org/10.1002/9780470191613[5] M. Ester、H.P. Kriegel、J. Sander、X. Xu，一种基于密度的算法，用于发现含噪声的大型空间数据集中的簇，载于：Simoudis E、Han J、Fayyad U（编），第二届知识发现与数据挖掘国际会议论文集，AAAI出版社，门洛帕克，加利福尼亚州，第226-231页，1996年。可从 https://aaai.org/papers/kdd96-037 获取[6] A. Rodriguez、A. Laio，通过快速搜索和查找密度峰进行聚类，《科学》，**344**（2014），1492-1496。http://dx.doi.org/10.1126/science.12420[7]R. Chen，用于时间序列聚类与分类的拓扑数据分析，博士论文，康涅狄格大学，2365，2019年。可从 https://opencommons.uconn.edu/dissertations/2365 获取[8]M. Yuvaraja、A.K. Deya、V. Lyubchichc、Y.R. Gela、H. Vincent Poor，多层网络的拓扑聚类，《美国国家科学院院刊》，**118**（2021），e2019994118。https://doi.org/10.1073/pnas.2019994118[9] M.G. Bergomi、P. Frosini、D. Giorgi、N. Quercioli，迈向用于数据分析和机器学习的群等变非扩张算子的拓扑—几何理论，《自然·机器智能》，**1** （2019），423-433。https://doi.org/10.1038/s42256-019-0087-3

[10] A. Zomorodian、G. Carlsson，持久同调的计算，《离散与计算几何》，**33**（2005），249-274。https://doi.org/10.1007/s00454-004-1146-y[11] T. Songdechakraiwut、B.M. Krause、M.I. Banks、K.V. Nourski、B.D. Van Veen，采用*Wasserstein*距离的快速拓扑聚类，ICLR 2022会议论文。可从 https://openreview.net/forum?id=0kPL3xO4R5[12]C.M. Pereira、R.F. de Mello，用于时间序列和空间数据聚类的持久同调，《专家系统与应用》，**42** （2015），6026-6038。https://doi.org/10.1016/j.eswa.2015.04.010[13] L.L. Bonilla、A. Carpio、C. Trenado，通过拓扑数据分析追踪集体细胞运动，《*PLOS*计算生物学》，**16**（2020），e1008407。https://doi.org/10.1371/journal.pcbi.1008407[14] C. Topaz、L. Ziegelmeier、T. Halverson，生物聚集模型的拓扑数据分析，《*PLoS ONE*》，**10** （2015），e0126383。https://doi.org/10.1371/journal.pone.0126383[15] M.R. McGuirl、A. Volkening、B. Sandstede，斑马鱼图案的拓扑数据分析，《美国国家科学院院刊》，**117**（2020），5113-5124。https://doi.org/10.1073/pnas.1917763117[16] M. Sinhuber、N.T. Ouellette，昆虫群中的相共存，《物理评论快报》，**119** （2017），178003。https://doi.org/10.1103/PhysRevLett.119.178003[17] R. Gonz´alez-Albaladejo、A. Carpio、L.L. Bonilla，受限Vicsek集群模型中的无标度混沌，《物理评论*E*辑》，**107**（2023），014209。https://doi.org/10.1103/PhysRevE.107.014209[18]S.Y. Oudot，持久性理论：从箭图表示到数据分析，《数学调查与专著》**209**，美国数学会，2015年。https://doi.org/10.1090/surv/209[19] H. Edelsbrunner、D. Letscher、A. Zomorodian，拓扑持久性与简化，《离散与计算几何》，**28** （2002），511-533。https://doi.org/10.1007/s00454-002-2885-2[20]G. Carlsson，拓扑学与数据，《美国数学学会通报》，**46** （2009），255-308。https://doi.org/10.1090/S0273-0979-09-01249-X[21] S. Moitrier、C. Blanch-Mercader、S. Garcia、K. Sliogeryte、T. Martin等，集体应力驱动正常细胞单层与Ras转化细胞单层之间的竞争，《软物质》，**15**（2019），537-545。https://doi.org/10.1039/C8SM01523F[22] R.C. Gonzalez、R.E. Woods，《数字图像处理》，培生教育，2022年。可从 https://www.pearson.com/en-us/subject-catalog/p/digital-image-processing/ 获取

[23]L. Jialin、Y. Wotao、L. Wuchen、T.C. Yat，多层最优传输：Wasserstein-1 距离的快速近似，*SIAM JSci Comput.*，**43** （2021），A193-220。https://doi.org/10.1137/18M1219813

[24]国际商务发展贸易统计，网址：https://www.trademap.org。访问日期：2022年3月30日。

[25] F. Sapienza、P. Groisman、M. Jonckheere，遵循费马原理的加权测地距离，载于第六届国际学习表征会议（ICLR）论文集，加拿大不列颠哥伦比亚省温哥华，2018年。网址：https://openreview.net/forum?id=BJfaMIJwG

[26] E. Cerami、J. Gao、U. Dogrusoz、B.E. Gross、S.O. Sumer等，cBio癌症基因组学门户：用于探索多维癌症基因组学数据的开放平台，*Cancer Dis-covery*，**2**（2012），401-404。https://doi.org/10.1158/2159-8290.CD-12-0095

[27] ICGC/TCGA全基因组泛癌分析联盟，《全基因组泛癌分析》，*Nature*，**578** （2020），82-93。https://doi.org/10.1038/s41586-020-1969-6

[28]TGCA泛癌图谱（2018）乳腺浸润性癌数据，http://www.cbioportal.org/study/summary?id=brca tcga pan can atlas 2018 于2019年6月26日访问。

[29] 细胞周期的KEGG通路，获取地址：https://www.genome.jp/pathway/hsa04110。访问日期：2019年6月26日。

[30]Z.M. Zhao、S.E. Yost、K.E. Hutchinson 等，CCNE1扩增与三阴性乳腺癌患者预后不良相关，*BMC Cancer.*，**19** （2019），96。https://doi.org/10.1186/s12885-019-5290-4

[31] F. Yi, Z. Wang, J. Liu, Y. Zhang, Z. Wang, H. Xu, X. Li, N. Bai, L. Cao, X. Song，染色体结构维持蛋白1：在基因组稳定性和肿瘤发生中的作用，*Int. J.Biol. Sci.*，**13** (2017)，1092-1099。https://doi.org/10.7150/ijbs.21206[32] T. Debniak, B. Gorski, T. Huzarski 等，CDKN2A（p16）的常见变异使乳腺癌易感性增加，*J. Med. Genet.*，**42** (2005)，763-765。https://doi.org/10.1136/jmg.2005.031476[33] R. Mahadevappa, H. Neves, S.M. Yuen 等，Cdc6 和 Cdt1 在乳腺癌中的预后意义，*Sci. Rep.*，**7**(2017)，985。https://doi.org/10.1038/s41598-017-00998-9[34]S. Izadi, A. Nikkhoo, M. Hojjat-Farsangi, A. Namdar, G. Azizi, H. Moham- madi, M. Youse, F. Jadidi- Niaragh，CDK1 在乳腺癌中的作用：诊疗一体化潜力，*AnticancerAgentsMed.Chem.*，**20** (2020)，758-767。

https://doi.org/10.2174/1871520620666200203125712

[35] Y. Liu, J. Qi, Z. Dou, J. Hu, L. Lu, H. Dai, H. Wang, W. Yang, WEE家族激酶的系统表达分析揭示了 PKMYT1 在乳腺癌发生中的重要性, 《细胞增殖》, **53** (2020), e12741。https://doi.org/10.1111/cpr.12741[36] H. Edelsbrunner, H.J. Harer, 计算拓扑学：导论。美国数学学会，普罗维登斯，2010。https://doi.org/10.1007/978-3-540-33259-6 7。[37] M. Kerber, D. Morozov, A. Nigmetov, 几何学有助于比较持久图，*ACM*实验算法学期刊，**22**(2017), 1.4。https://doi.org/10.1145/3064175