arXiv:2601.00508v1 [stat.ME] 1 Jan 2026

ballmapper：在 Stata 中应用拓扑数据分析球映射器

Stata

SimonRudkin *[1] 和 Wanling Rudkin †[2]

[1]曼彻斯特大学社会科学学院，英国[2]埃克塞特大学商学院，埃克塞特大学，英国

2026年1月5日

摘要

拓扑数据分析球映射器（TDABM）提供了一种无需模型的多变量数据可视化方法，无须承受降维所带来的信息损失。D-lotko（2019年）提出的 TDABM 使用大小相同的球覆盖多维点云，球的半径是唯一的参数。TDABM 可视化能够保留数据的完整结构。TDABM 生成的图形可以根据其他变量、模型残差或多变量数据中的变量进行着色。越来越多的文献将 TDABM 的强大功能应用于金融学、经济学、地理学、医学和化学等领域。本文介绍 TDABM 以及 Stata 的 ballmapper 软件包。

关键词：拓扑数据分析、Ball Mapper、ballmapper、Stata

1 引言

D-lotko（2019）的拓扑数据分析球映射器是一种用于可视化多维数据集的算法。与 t分布随机邻域嵌入或 UMAP 等方法不同，它不进行降维，因此不会丢失数据集中的信息。有关 TDABM、t分布随机邻域嵌入、UMAP 与 Singh 等人（2007）的原始 Mapper 算法之间比较的介绍，请参阅 D-lotko（2019）和 Dlotko 等人（2022）。

*通讯作者 完整地址：社会统计系，社会科学学院，曼彻斯特大学，牛津路，曼彻斯特，M13 9PL，英国。电话：+44 (0)7955 109334 电子邮箱：simon.rudkin@manchester.ac.uk

†完整地址：英国埃克塞特大学商学院，埃克塞特大学，斯特里瑟姆庭院，雷恩斯大道，埃克塞特，EX4 4PU。电子邮箱：w.rudkin@exeter.ac.uk

查看数据中的结构可以了解潜在的相关性和关系。利用数据可视化来探索结构的优点已得到充分认识（Anscombe，1973；Matejka 和 Fitzmaurice，2017）。Anscombe（1973）进一步指出，建模后的数据可视化有助于更有效地进行模型评估[1]。实证研究者面临的挑战是，除非使用高级算法，否则数据可视化仅限于 2 或 3 个维度。主成分分析（PCA）、UMAP、TSNE 和面板图等工具可以提供可视化，而 TDABM 则提供了一种直观且易于实现的替代方案。

我们将一个数据集 X 定义为由 K 个变量构成。每个变量 $k \in [1, K]$ 都应有足够多的取值，使得绘制散点图成为合理的选择[2]。在 X中，每个数据点 x_i在变量 k上的取值为 x_{ik} 。由于 TDABM 是一种基于距离的方法，该算法无法处理缺失值。因此，要求 x_{ik} 在 $i \in [1, N]$ 和 $k \in [1, K]$中均无缺失值。此外，必须对数据集进行缩放，以确保所有 x_k 具有相似的取值范围。缩放是机器学习算法和其他基于距离的统计建模中常见的做法。本文中的所有讨论均以 X_k 作为经过适当缩放的数据集。

使用球的直观依据是：位于给定半径范围内的点“相似”。将 X 表示为一个 K维点云后，距离某个点不超过半径 ε的任何点，都可以视为与该点相似。TDABM 算法用球覆盖数据，使每个点至少位于一个球中。讨论球时，我们便可以谈论相似点集。这与聚类算法有相似之处，但组的大小不会发生变化。球处理离群点的方式也有所不同。像 k均值（Hartigan 和 Wong，1979）这样的聚类算法会将离群点归入最近的聚类，而 TDABM 会将离群点留在它们自己的球中，除非半径足够大。在优化聚类数量时，k均值生成的聚类数少于 TDABM 使用的球数。目前尚无针对球半径 ε[3]的优化方法。Otway 和 Rudkin（2024）、D-lotko 等人（2024）以及 Rudkin 等人（2024a）均对 TDABM 结果与聚类算法进行了比较。

TDABM 的目标是生成 X 的拓扑忠实表示，使其能够以二维方式进行可视化。TDABM 通过用球覆盖空间来实现这一目标。第一个球（球 1）的中心位于数据集中的第一个观测值处。绘制一个半径为 ε 的球。球内的点即被覆盖。算法继续从未覆盖集合中选取一个点，并以该点为中心构建一个新球

[1]Anscombe（1973）的示例涉及 4 个数据集，它们之间的 X 和 Y 具有相同的相关性，并且 X 和 Y 具有相同的一阶和二阶矩。此外，这 4 个数据集具有相同的普通最小二乘回归线和模型 R平方。只有查看数据后，你才能发现其中只有一个数据集中的关系是线性的。[2]我们不规定用于定义“足够”的数值，但建议在处理可能取值少于 10 个的变量时保持谨慎。[3]之所以不提供最优半径，是因为我们希望在多个层次上理解数据的结构。对于局部结构，需要较小的半径。同时，要观察数据的完整结构，则需要较大的半径。一个自然的类比是绘制地球表面。针对不同目的，需要使用不同比例尺地图。

。当所有点都被覆盖后，算法便停止。由于球的中心选取方式，球之间可能发生重叠。在二维空间中，球 j（写作 $B_j(X,\ \epsilon)$）呈现为圆盘。数据由一系列 L 球覆盖，因此 TDABM 图上有 L 个圆盘。TDABM 图中球与球之间的边表示球之间存在重叠。为体现联合分布某一部分中数据的密度，球的大小根据球内点的数量确定。第 2 节将进一步阐述实现这种抽象表示的机制。

由于 TDABM 算法保留了每个点位于哪个球内的信息，用户可以根据数据中包含的任意函数，对图中的圆盘进行着色。一种常见的着色方式是使用另一个变量，即 Y。将球内 Y 的平均值作为着色依据后，TDABM 图可以展示 Y 如何随着 X 的分布而变化。使用着色还能够让用户看到每个 x_k 如何在空间中变化。此外，成员信息使用户能够追溯到其数据。为了评估模型，可以将残差或预测值用作着色变量，从而实现模型评估。当数据点具有标识符（例如地区名称或公司名称）时，用户便可以讨论每个球中包含哪些标识，以及该球的着色所传达的信息。Stata 实现提供了更改着色方式并链接回底层数据集的选项。

TDABM 能够生成具有可解释性的多变量数据可视化结果，因此该方法已广泛应用于经济学和金融学领域。金融学论文主要利用 TDABM 重新评估现有模型。Qiu 等人（2020）表明，公司违约仅存在于由 Altman（1968）Z-score 违约区域定义的空间中的一个小子集内。Charmpi 等人（2023）在 Qiu 等人（2020）的基础上进一步指出，可以根据邻域内历史上确实发生违约的公司比例，利用 TDABM 有效预测公司违约。在 Fama 和 MacBeth（1973）的有效市场假说下，金融收益的方向不应具有可预测性。Rudkin 等人（2024b）说明，未来比特币收益方向的预测能力取决于预测是在过去轨迹空间的何处进行的。在这两种情况下，价值都源于观察过去的模型在哪些地方表现不佳，并能够将这种表现不佳与数据联系起来。D-lotko 等人（2024）展示了如何通过可视化解释变量空间中的模型拟合情况，揭示机器学习模型在哪些地方优于既有的普通最小二乘模型。通过表明机器学习模型的残差仅在该空间的极端区域内显著低于其他模型（按绝对值计），人们重新审视了放弃普通最小二乘法的可解释性以换取更优机器学习拟合效果的价值。

经济学研究更加注重运用 TDABM 将数据中的故事进行可视化。例如，Rudkin 等人（2024a）表明，英国在2016年脱离欧盟的投票集中于一组高度同质化的选区（具有许多连接的大球），而投票支持留在欧盟的选区则更加分散。Otway 和 Rudkin（2024）随后

展示了这些模式如何随时间持续存在，并与大选结果相关联。其他应用包括研究欧洲地区的数字鸿沟（Rudkin and Dlotko，2024），以及 Tubadji and Rudkin（2025）对迁移模式的研究。Benites 等人（2025）近期的一篇论文进一步研究了小规模地方地理区域的社会人口空间中教育成果的变异性。Rudkin and Webber（2023）研究了各地区的发展轨迹及其对全球金融危机的韧性。每个球内部韧性的极端变异性表明，发展轨迹与韧性并不一致。在经济这一复杂生态系统中，变量的选择对于解释 TDABM 十分重要。TDABM 通常表明还需要更多变量。

在自然科学领域，TDABM 也日益受到关注。例如，Madukpe 等人（2025年）将 TD-ABM 应用于环境监测，展示了这种方法如何助力实现更高效的污染监测；作者指出，与 Singh 等人（2007年）的传统映射器相比，TDABM 只需一个参数，这是一项优势。Koljanˇci´c 等人（2025年）进一步表明，在单个图中查看多维数据集有助于理解《分析化学》中的反应。与此相关的是，生物学中的多维数据集使可视化变得复杂。以从鱼类整体标本中学习为例，TDABM 展示了降维对理解衰变的影响（Valerio 等人，2025年）。Han 等人（2025年）使用 TDABM 查看颅内压轨迹，利用该方法既能将数据的维度视为滞后值，也能将其视为在同一时间段观测到的不同变量。轨迹方法与 Rudkin and Webber（2023年）及 Rudkin 等人（2024b）采用的方法类似。

本文其余部分安排如下。第 2 节将深入介绍 TDABM 方法论。第 3 节通过一个双变量示例，帮助读者直观理解 TDABM 方法论。第 4 节讨论示例构建中使用的人工数据，并基于这些数据构建 TDABM 图。第 5 节介绍使用 TDABM 可开展的其他分析阶段。第 6 节展示一个示例分析，所用数据来自 Stata 内置的 1978 年汽车数据集及其价格。第 7 节为结论。

2 方法论

本指南中有一个包含 K 个变量的数据集 X。我们将数据视为点云 P。将每个数据点编为索引 i，$i \in \{1, ..., N\}$ 后，我们可以通过数值 x_{ik}，$k \in \{1, ..., K\}$ 定义 x_i 在 P 中的位置。直观来说，这就是将双变量数据集绘制到散点图的二维空间中的方式。点云是对散点图概念的推广。TDABM 算法要求另有一个变量 Y，用于为数据点着色。着色变量 Y 可以

是 X 个变量之一，也可以是整个数据集中的一个独立变量。TDABM 算法的最后一个输入是覆盖 ε 中所用球的半径。用于确定最优 ε 的算法仍是一个持续推进的研究议程。不过，我们强烈建议了解数据在 ε 的多个取值下是如何构成的。本节将介绍 TDABM 如何为 $K-$ 维数据创建抽象的二维表示。

算法首先从 P 中随机选择一个点。第一个被选中的点成为第一个地标，即 l_1。以 l_1 为中心绘制一个半径为 ε 的球。绘制的第一个球称为球 1，$B_1(X, \varepsilon)$。请注意，对球进行编号只是为了便于后续讨论 TDABM 图，数字 1 本身没有其他含义。P 中包含在 $B_1(X, \varepsilon)$ 内的点被视为已覆盖，并成为已覆盖集合的首批成员。随后从 P 中未覆盖的点里随机选择第二个地标。新的地标是 l_2，并作为球 2，$B_2(X, \varepsilon)$ 的中心。$B_2(X, \varepsilon)$ 内任何尚未被球 1 覆盖的点都会被加入已覆盖集合。$B_1(X, \varepsilon)$ 与 $B_2(X, \varepsilon)$ 的组合成为总体覆盖 $B(X, \varepsilon)$ 的起始部分。如果仍有未覆盖的点，则从未覆盖集合 $B'(X, \varepsilon)$ 中选择另一个地标 l_3。绘制一个球，即 $B_3(X, \varepsilon)$，将更多点加入 $B(X, \varepsilon)$。选择地标并绘制球的过程会持续进行，直到所有点至少被一个球覆盖，也就是说 $B'(X, \varepsilon) = \emptyset$。地标总数为 L。

每个球 $B'_b(X, \varepsilon)$ $b \in [1, L]$ 都会保留其所包含点的信息。球 b, n_b中的点数反映了该球所覆盖区域内数据集的密度。球 $\bar{y}_b$内各点的Y 平均值，可提供有关数据空间中由球 $B_b(X, \varepsilon)$ 覆盖部分的 Y 信息。点的信息也使算法能够识别同时属于多个球的点。当两个球 q 和 s的交集非空时 $B_q(X, \varepsilon) \cap B_s(X, \varepsilon) \neq \emptyset$, 就在 $B_q(X, \varepsilon)$ 与 $B_s(X, \varepsilon)$ 之间绘制一条边。绘制 TDABM 图时，球以圆盘表示，其大小与 n_b 成正比，并在识别出连接关系时用边将其连接起来。由于该算法将一个 K 维数据集转换为二维数据，因此生成的图在页面上水平方向和垂直方向没有任何解释意义。该图实际上是数据的一种抽象二维表示。

要在 Stata 中应用 TDABM，需要从 Simon Rudkin 的 GitHub 网站安装 ballmapper 软件包。安装命令见方框 2.1。

方框 2.1：安装 `ballmapper`

目前可以从 Simon Rudkin 的 GitHub 仓库获取 ballmapper

```
net install ballmapper ,
    来自（" https :// raw. githubusercontent . com / srudkin12 / statabm / main ") replac e
```

3 直观理解

为了在实践中理解 TDABM 图的构建，本简短章节将介绍一个双变量示例。随 ballmapper 软件包附带的 GitHub 上提供了用于生成分析的 .do 文件 intuition.do。数据包含 1000 个点， $N = 1000$ 和两个变量， $K = 2$。这两个变量均从标准正态分布中随机抽取，且相互独立。因此，我们有$X_1 \sim N(0,1)$ 和 $X_2 \sim N(0,1)$。使用这样的高斯云，能使示例符合经济学和金融学中许多假设的数据生成过程。高斯云还具有一些结构特征，我们可以通过 TDABM 加以展示。首先，点云中心的密度较高，对应于两个变量 X_1 和 X_2都处于距均值较短距离内的点。离均值越远，点云就越稀疏。让一个变量保持在较接近其均值的位置，同时让另一个变量变化范围更大，点的数量就会减少。然而，若两个变量都偏离均值，观测到一个点的联合概率就会低得多。具体而言， X_1 和 X_2 的散点图中心密集，沿 $X_1 = 0$ 和 $X_2 = 0$ 坐标轴的区域较为稀疏，而图的角落几乎没有点。高斯云密度的等高线呈圆形。因此，点云的 TDABM 图必须呈现这些特征。框 3.1 提供了生成并绘制数据所需的 Stata 代码。

框 3.1：数据集的 Stata 代码

首先生成 $N \sim (0,1)$ 中的初始值 X_1 和 X_2，从而构建一个初始的双变量数据集。第一步是清空环境，确保 Stata 使用的是新数据。

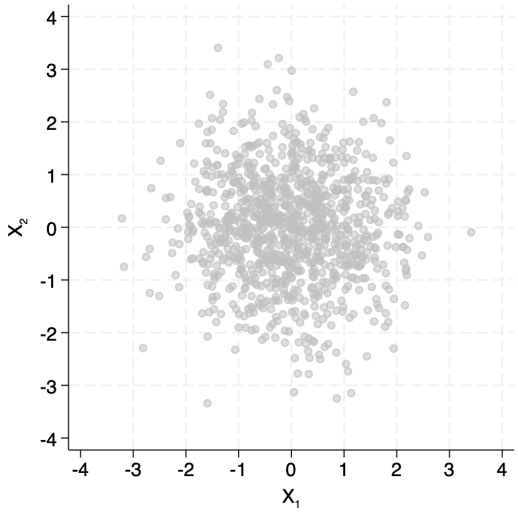
```
clear
设置随机种子1

设置观测值 1000


生成 x1 =正态随机数(0,1)
生成 x2 =正态随机数(0,1)
```

生成数据图以展示结果

```
twoway (scatter x2 x1, mcolor(gs12%50) msize(small)), xtitle("X{sub:1}")
    ytitle("X{sub:2}") xlabel(-4(1)4) ylabel(-4(1)4) aspect(1)
    graphregion(color(white)) name(scatter_orig, replace)
```

在此示例中，选择了 $\varepsilon = 1$，对应于变量的 1 个标准差。使用 1 也能确保代码易于理解。在此 $\varepsilon = 1$ 下，「球」足够大，可以跨越各个观测值之间的间隙，从而确保形成连通图。 $\varepsilon = 1$ 也足够小，能够捕捉

图 1：二元正态示例数据



注：人工双变量示例中使用的数据点。每个变量均独立地从标准正态分布中随机抽取，$X_1 \sim N(0,1)$ 和 $X_2 \sim N(0,1)$。$N = 1000$。

当我们远离原点时，分布的密度逐渐衰变。虽然较小的 ε 值能更细致地呈现点云，但 $\epsilon = 1$ 成功地将 1,000 个点的样本简化为精简的「骨架」。我们将展示该示例实际上有 $L = 21$ 个地标。

图 1 展示了本示例中用于说明 TDABM 方法的数据集。高斯云清晰可见。点云中心密集，而外围区域的点较少。一些点远离中心，孤立地分布在图的边缘。因此，要准确表示这些数据，就需要捕捉其密度以及大多数点之间的整体邻近关系。要开始构建 TDABM 表示，我们先在数据集中的点 1 周围画一个球。绘制第一个球的代码见框 3.1。

框 3.2：数据集的 Stata 代码

为了演示如何围绕单个数据点画球，取数据集中的第一个数据点，并计算所有其他点到该点的距离。这个单个数据点就是地标 1。

```
局部 x11 = x1[1]
局部 x21 = x2[1]
gen dist_to_p1 = 平方根((x1 - `x11')^2 + (x2 - `x21')^2)
```

该示例包含 $\epsilon = 1$，因此找出所有与地标距离不超过 1 的点 k.

```
生成 位于_球 = (距离_到_p1 <= 1)
```

绘图时，我们需要用一个圆来表示球

```
范围 phi 0 2*_pi 100
生成 圆_x = `x11' + cos(phi)
生成圆_y = `x21' + sin(phi)
```

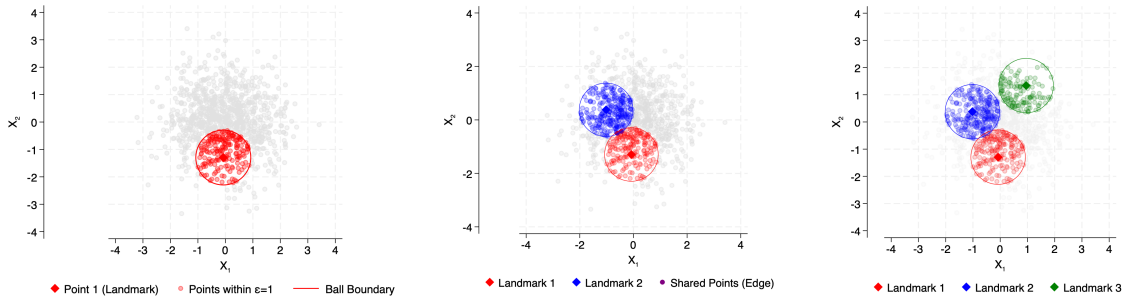
最后生成了散点图。请注意，现在命令中包含的元素多得多。

```
twoway (scatter x2 x1 if in_球 ==0, mcolor(gs14 %40) msize(small)) (scatter x2 x1 if in_球 ==1 & _n > 1, mcolor(red %30) msize(small)) (scatter x2 x1 if _n ==1, mcolor(red) msize(medium) msymbol(D)) (line circle_y circle_x, lcolor(red) lwidth(medium)), xtitle("X{sub:1}") ytitle("X{sub:2}") xlabel(-4(1)4) ylabel(-4(1)4) aspect(1) legend(order(3 "点 1(地标)" 2 "{&epsilon}=1 内的点" 4 "球边界") pos(6) rows(1)) graphregion(color(white)) name(ball_demo, replace)
```

运行框 3.2 中的代码后，我们得到图 2 的子图 (a)。单个球位于整个点云的下方中心位置。由于这里是空间中较为密集的部分，因此覆盖的点数量很多。然而，未覆盖的点集合仍然很大。为了展示重叠，在一个靠近球 1 的点周围绘制第二个球。在 TDABM 算法的随机实现中，所示点有可能被选中。在大多数 TDABM 软件实现中，算法会按照点在数据集中的出现顺序选择点。对于此示例，第二个点就是图 2 的子图 (c) 中所示的点。它与此前选定的两个球都没有重叠。

图 2 中选作地标的具体点取决于随机选择过程。不过，最终生成的 TDABM 图对于地标选择顺序是稳定的。在截至目前发表的 TDABM 论文中，通常使用图来说明经过数千次选择迭代后 TDABM 图的一致性（例如，Rudkin 等，2024a）。如需查看地标选择的不同扰动，您可以

图 2：构建拓扑数据分析球映射器图



(a) 球 1　　(b) 重叠示例球　　(c) 添加球 2

注：使用 Stata 软件包 `ballmapper` 中实现的拓扑数据分析球映射器（TDABM）图构建球。子图 (a) 展示了以第一个数据点（标记为地标 1）为中心、半径为 1 的球的构建过程。该球标记为球 1。子图 (b) 展示了以另一个数据点为中心的第二个球，该数据点经过选择，与球 1 存在重叠。此处该地标点标记为地标 2。子图 (c) 展示了以数据框中的第二个数据点（此处标记为地标 3）为中心的球。数据具有 $X_1 \sim N(0,\ 1)$ 和 $X_2 \sim N(0,\ 1)$。 $N = 1000$。

在将数据框提供给 ballmapper() 函数之前重新排列数据。为进一步展示该双变量数据集的 TDABM 图构建过程，本文提供了一个附录，其中使用了一个在构建 X_1 和 X_2时采用不同随机种子生成的数据集。随附的 GitHub 中提供了用于创建该附录的完整 Stata 代码。

4 人工数据

为了说明同时考虑局部与全局结构的必要性，我们考虑一个包含 900 个观测值的双变量数据集 X。变量 X_1 和 X_2 独立地从标准正态分布中抽取。抽取后，对部分子集进行平移，以构造一个大型 X 形。100 个观测值在 X_1 上平移 -6，并在 X_2上增加 6。一个包含 100 个观测值的组在 X_1 上平移 -6，并在 X_2上平移 -6。第三个组在 X_1 上平移 6，并在 X_2 上平移 6；而第四个组在 X_1 上平移 6，并在 X_2 上平移 -6。这四个组构成 X 形的末端。构成全局 X 结构的还有 4 个按与平移 6 相同的模式平移 3 的组，以及最后一个保留在 X 形中心的组。由于该数据集呈 X 形，因此下文将其称为 X 数据集。

经过 $X_1 = -0.032$ 和 $X_2 = 0.024$的平移后，均值仍接近 0。平移使 X_1 和 X_2 的标准差增大，因此 $\sigma_1 = 4.592$ 和 $\sigma_2 = 4.601$。两个变量之间的皮尔逊相关系数为 $\rho_{12} = 0.003$。汇总统计量与从标准差更大的分布中抽取的高斯云一致。前两个矩中没有任何信息表明存在 X 形。X 数据集也提醒我们，按照 Matejka 和 Fitzmaurice（2017年）的理念，可视化数据十分重要。

表1：X 数据集的结果

结果	简短	方程	假设
Y_1	线性	$Y_1 = X_1 + X_2 + \theta$	$\theta \sim N(0, 0.2)$
Y_2	组	-	-
Y_3	二次	$Y_3 = X_1^2 + X_2^2 + \theta$	$\theta \sim N(0, 0.2)$
Y_4	噪声	$Y_4 = \phi$	$\phi \sim N(0, 1)$
Y_5	受限范围	$Y_5 = \begin{cases} 1 \text{ 如果 } & 0 < X_1 < 3 \quad 0 < X_2 < 3 \\ 0 & \text{否则} \end{cases}$,	

注：用于分析结果 用于分析 X 数据集。 X_1 和 X_2 是数据集中的变量。 θ 和 ϕ是噪声项 a s 定义在 相关行上。 $N = 900$。

结果 Y_i 根据表1中的表达式给出。这些表达式表示 X变量的线性函数、组成员关系、 X变量的二次函数、噪声，以及空间某个子集的二元指示变量。通过设置 5 种不同的模式，我们可以看到 TDABM 如何在 X 形数据点云上表示这些结果。图形 3 和 4 展示了这一点。

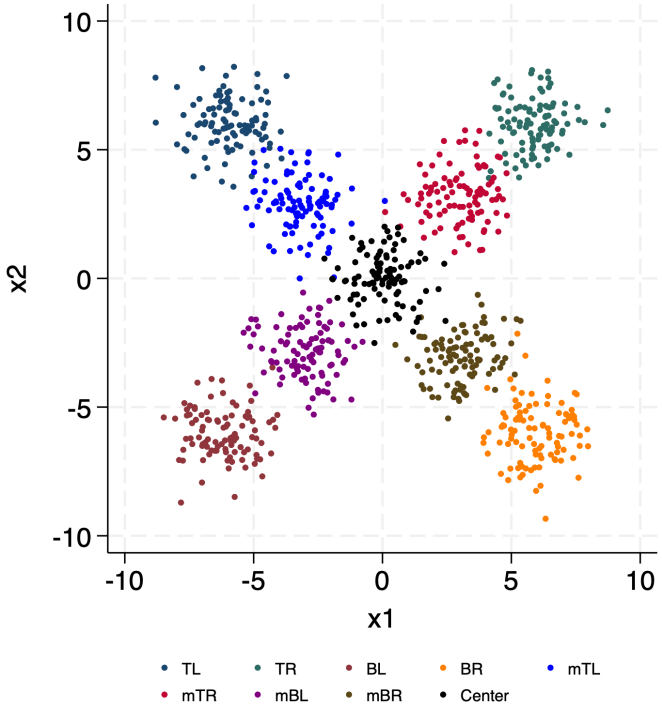
图3展示了 X 形如何自然形成。散点图中可以看到各个子云。9 个密集的中心一眼便能辨认出来。为了突出结构，图3根据组号进行着色。颜色的排列方式使相邻组之间具有足够的对比度。从标准正态分布中抽取 X_1 和 X_2 意味着，在极限情况下，95% 的观测值位于均值上下 2 个标准差的范围内，即介于 -2 和 2 之间。在 3 个标准差范围内，极限情况下可以包含 99% 的观测值。由于各点云按 3 的倍数进行偏移，因此点云之间可能发生重叠。相邻两个中心之间的距离为 $\sqrt{18} \approxeq 4.24$ ，所以中心周围 2 个标准差的范围不会重叠，但 3 个标准差的范围会重叠。仔细观察图3可以发现，确实存在一些重叠。

—

图4展示了同一个 X 数据集，并按照另外 4 种着色规则进行着色。在子图 (a) 中， Y_1 的最高值出现在右上方，此处 X_1 和 X_2 的值也最高。与此同时， Y_1 的最低值出现在左下方。沿着从左上到右下的对角线，颜色呈现渐变，这与对角线上的每个子云中最低值都位于该子云左下方这一事实相符。子图 (b) 显示，正如预期，X 形两端的值最高。子图 (c) 没有呈现任何模式，因为 Y_4 是一个独立于 X_1 和 X_2的噪声项。最后，子图 (d) 将带有$0 < X_1 < 3$ 和 $0 < X_2 < 3$ 的点显示为红色，与其他所有点明显区分开来。由于这里处理的是二维数据，散点图足以展示整个空间中的模式。

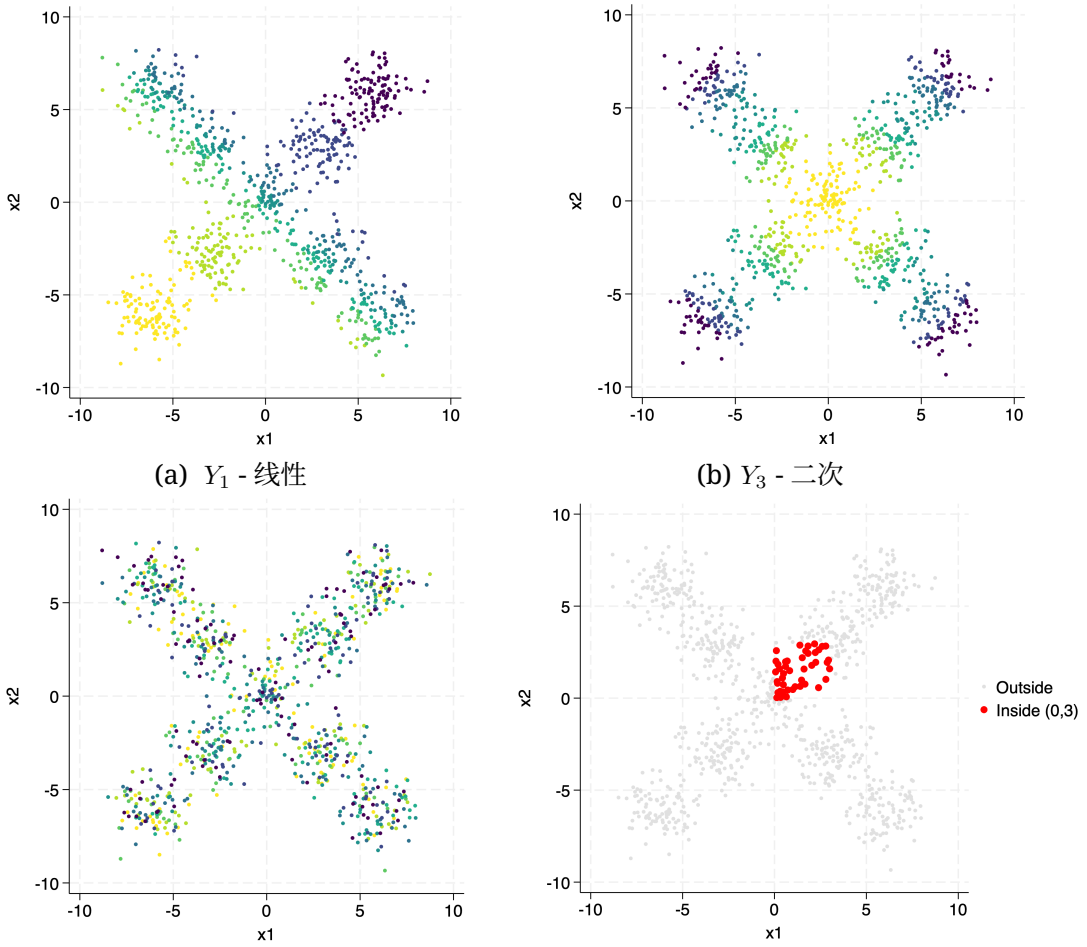
现在，我们希望展示 TDABM 如何捕捉图4各子图中观察到的模式。要在数据上实现 ballmapper，请按照框 4.1 中的代码操作。ballmapper 命令接受坐标轴变量 X、着色变量 Y 和半径 ε 作为输入。此外，还有一些可选参数：布局排斥力和吸引力，它们控制算法如何显示 TDABM 图，

图3：带结果的人工 X 形数据



注：该图展示了 X 数据集。底层数据集包含 900 个观测值，涉及 2 个变量， X_1 和 X_2。数据经过平移，形成 9 组，每组包含 100 个点，中心分别位于 (-6,6)、(-3,3)、(3,3)、(6,6)、(0,0)、(-3,-3)、(3,-3)、(-6,-6) 和 (6,-6)。着色变量 Y_2 根据数据点所属的组确定，取值范围为 1 至 9。图例中的标签仅用于识别。

图4：具有替代结果的人工 X 形数据



(a) Y_1 - 线性　(b) Y_3 - 二次

(c) Y_4 - 噪声　(d) Y_5 - 受限范围

注：X 数据集的不同着色散点图。底层数据集包含 2 个变量 X_1 和 X_2的 900 个观测值。对数据进行平移，使其分为 9 组，每组 100 个点，中心分别位于 (-6,6)、(-3,3)、(3,3)、(6,6)、(0,0)、(-3,-3)、(3,-3)、(-6,-6) 和 (6,-6)。着色方式表示为 $Y_1 = X_1 + X_2 + \theta$ where $\theta \sim N(0,\ 0。2)$，Y_2 为组号，$Y_3 = X_1^2 + X_2^2 + \theta$ where again $\theta \sim N(0,\ 0。2)$，$Y_4 = \phi$ where $\phi \sim N(0,\ 1)$，且当 $0 < X_1 < 3$ 和 $0 < X_2 < 3$ 均满足时，Y_5 取值为 1。

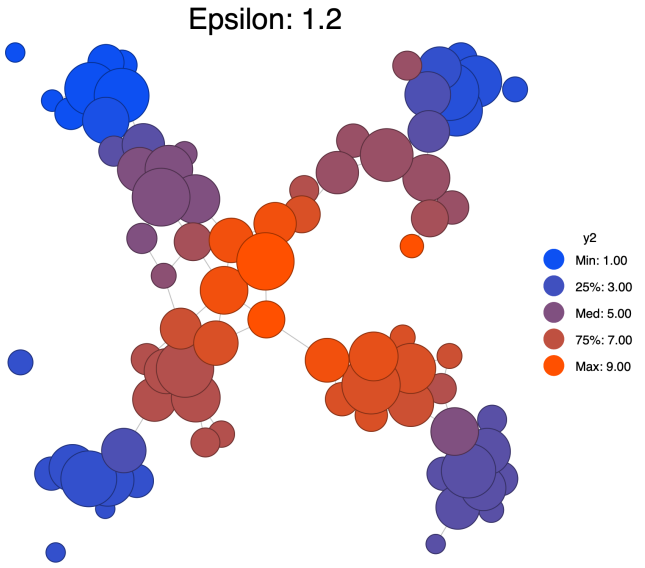
图形。如果吸引力过强，边可能会把球拉得过近。如果排斥力过强，所有东西都会被推向边缘。由于吸引力和排斥力是显示参数，因此不会改变 TDABM 图所提供的推断。用户可以根据自己的数据调整吸引力和排斥力。最后，文件名参数指定算法将创建的 .png 文件的名称。.png 文件会保存到 Stata 会话当前使用的工作目录中。完成后，边强度、地标和球大小的详细信息会存入新的 Stata 框架 BM RESULTS。每个点所属的球的信息，以及与底层数据集合并后的数据，会存入 BM MERGED。在探索阶段，只有 TDABM 图是必不可少的。

框 4.1：在`ballmapper`上应用于 X 数据集为演示如何围绕单个数据点绘制球，取数据集中的第一个数据点，并计算所有其他点到该点的距离。这个单独的数据点就是地标 1。ballmapper x1 x2 , epsilon (0.8) color (y1) layout repulsion (0.05) attraction (0.01) filename (" xy1bm08 ")该示例的 $\epsilon = 0.8$，但我们也可以使用任何其他 epsilon 进行拟合。我们还会使用其他 y 值。因此， $y4$ 半径为 $\varepsilon = 1.2$ 时，输入为 ballmapper x1 x2 , epsilon (1.2) color (y4) layout repulsion (0.05) attraction (0.01) filename (" xy4bm12 ")代码运行完成后，Stata 会显示：Ball Mapper 成功。-> 图数据存储在框架中：BM_RESULTS -> 原始数据 + 球 ID 存储在框架中：BM_MERGED

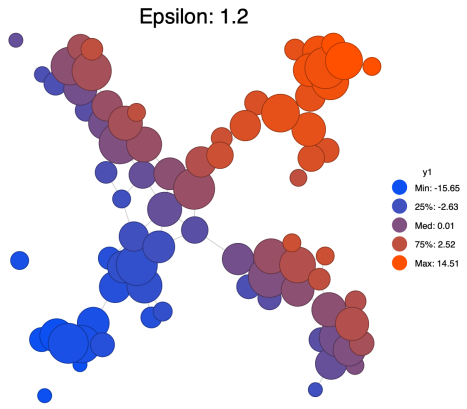
图 5 包含 5 个子图，分别对应一个结果变量。由于这些图基于相同的随机变量，因此数据的结构完全相同。这里使用半径 $\varepsilon = 1.20$来同时展示子云的局部结构和 X 的全局结构。图 6 将进一步探讨半径的选择。子图 (a) 按组成员关系着色。组成员关系包含 9 个不同的值，因此较为复杂，但其结构很清晰。当一个球的组成员关系不是整数时，这是因为球内包含了来自不同子云的点。正如预期的那样，我们可以看到各个子云之间的区域存在重叠。在 $\varepsilon = 1.20$处，仍有一些球未与任何组相连。这些球位于分布的尾部。在绘制高斯云时，观察到这类离群点是很常见的。

观察图 5 的另外 4 个子图，可以看到预期的模式确实出现了。子图 (b) 按 Y_1 着色，并呈现出从 X 左下方的最低值开始的预期梯度，延伸至

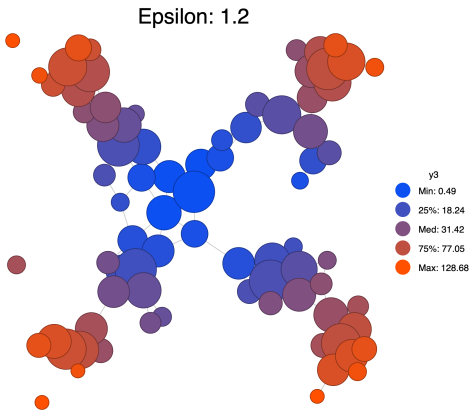
图 5：X 数据集的 TDABM 图



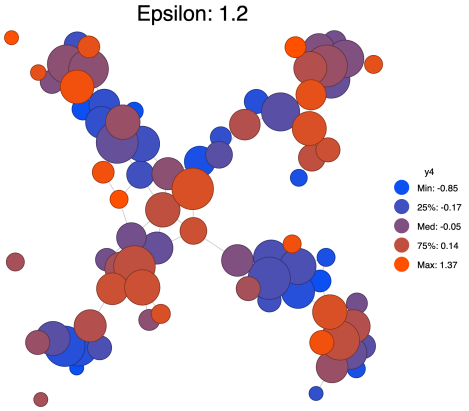
(a) 按组成员关系着色



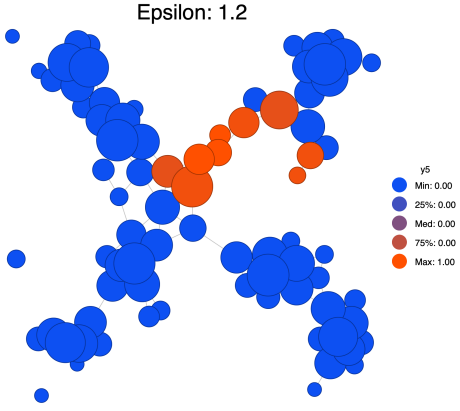
(b) Y_1 - 线性



(c) Y_3 - 二次



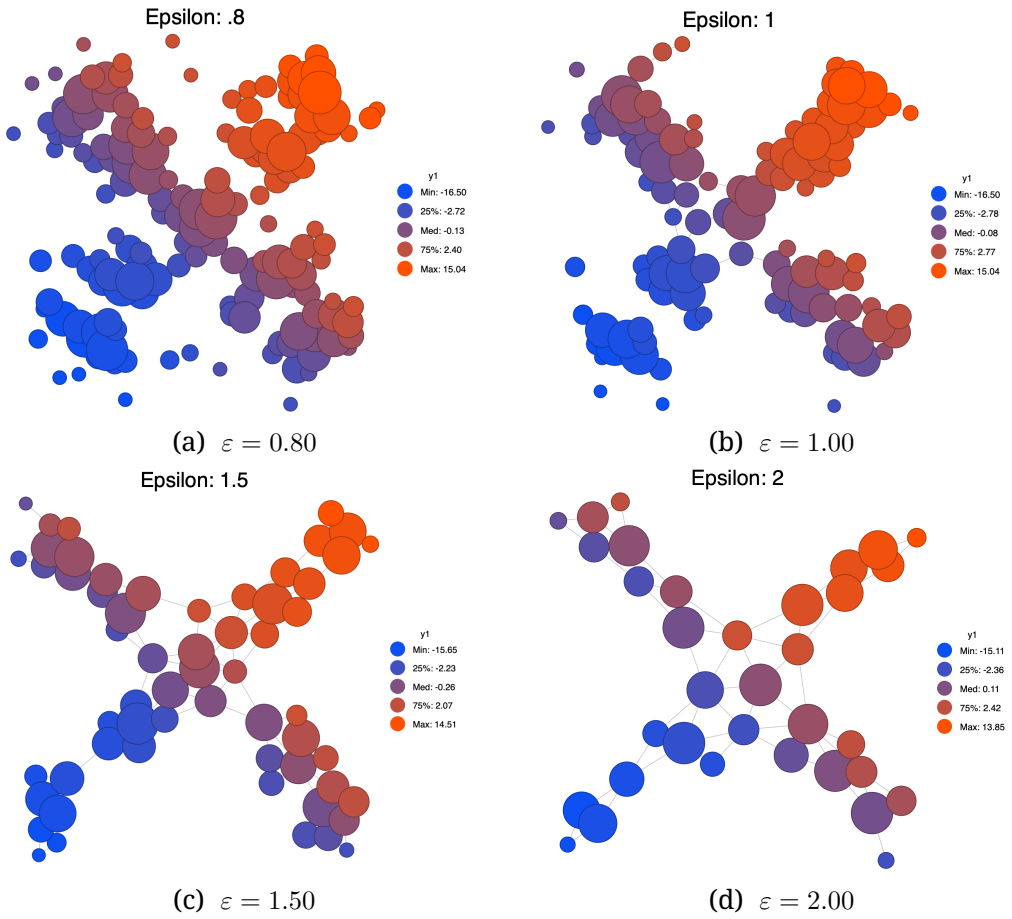
(d) Y_4 - 噪声



(e) Y_5 - 受限范围

注：底层数据集包含 2 个变量 X_1 和 X_2的 900 个观测值。数据经过平移后，形成 9 组，每组包含 100 个点，中心分别位于 (-6,6)、(-3,3)、(3,3)、(6,6)、(0,0)、(-3,-3)、(3,-3)、(-6,-6) 和 (6,-6)。着色定义为 $Y_1 = X_1 + X_2 + \theta$，其中 $\theta \sim N(0, 0。2)$， Y_2 为组号；$Y_3 = X_1^2 + X_2^2 + \theta$，其中同样有 $\theta \sim N(0, 0。2)$； $Y_4 = \phi$，其中 $\phi \sim N(0, 1)$；当 $0 < X_1 < 3$ 和 $0 < X_2 < 3$ 均满足时， Y_5 取值为 1。

图 6：半径对 X 数据集的作用



(a) $\varepsilon = 0.80$ (b) $\varepsilon = 1.00$

(c) $\varepsilon = 1.50$ (d) $\varepsilon = 2.00$

注：通过球半径 ε 的 4 个不同取值评估半径的作用。底层数据集包含 2 个变量 X_1 和 X_2 上的 900 个观测值。数据经过平移，形成 9 组，每组 100 个点，分别以 (-6,6)、(-3,3)、(3,3)、(6,6)、(0,0)、(-3,-3)、(3,-3)、(-6,-6) 和 (6,-6) 为中心。球体着色表示为 $Y_1 = X_1 + X_2 + \theta$，其中 $\theta \sim N$ (0，0。2)。

右上方的值最高。另一条对角线内部的颜色渐变更明显，沿着整条臂，左下方的值较低，右上方的值较高。请注意，即使我们旋转 X，仍然能够辨认出 X 中从左下到右上的那条对角线。子图 (c) 显示了 Y_3 完整的结果，X 的各条臂末端的值如预期般较高。子图 (d) 确实显示出噪声，因为 X 的各个部分都混杂着不同颜色。子图 (e) 只用两种颜色表示，因为所有值都是 0 或 1。所有值为 1 的点都位于中心子云的右上方，以及以 (3,3) 为中心的子云的部分区域。我们可以看到，以 (3,3) 为中心的子云中有一小条球体臂朝中心延伸。这条球体臂位于两个子云之间较稀疏的空间中，并未与其他任何点相连。实际上，从子图 (a) 可以看出，这条臂的一部分延伸到了中心子云的离群点区域。这些都是人工数据集；图 5 让我们确信，TDABM 捕捉到了预期的特征。

这 4 张图　图 6 中表示 4 个不同的 ε 值，以补充 $\varepsilon = 1.20$。增加额外的半径 i　　s

这是确保从 TDABM 图得出的推断稳健的重要环节。当 ε 较小时，点云稀疏区域中会出现更多与主体形状断开的球。子图 (a) 展示了 $\varepsilon = 0.80$。较小的球也意味着，相较于使用较大的半径，需要更多的球来覆盖空间。当半径增大到子图 (b) 中的 $\varepsilon = 1.00$ 时，断开的球数量减少，X 的主体形状也更加清晰。在 $\varepsilon = 1.00$ 时，以 (-6,-6) 为中心的子云与其余子云不相连。正如设置部分所述，每个轴上距离取 3 意味着重叠有限。只有更大的半径才能找到重叠。我们在主要结果中的 $\varepsilon = 1.20$ 处看到了重叠。继续增大半径会减少球的数量，并开始丢失子云的一些形状。在 $\varepsilon = 1.50$时，子图 (d) 显示子云仍有密集的中心，但在 $\varepsilon = 2.00$时，子图 (e) 难以展现局部结构。

本节展示的数据集有意设计了全局和局部结构。验证 TDABM 是否捕捉到了这些结构，是对该算法的概念验证。在现实场景中，结构并非已知。因此，我们建议使用多个半径。待探索的半径范围由数据的范围决定。在本例中，两点之间可能的最大距离约为 25[4]。1.2 可使对角线上分布 20 个球。

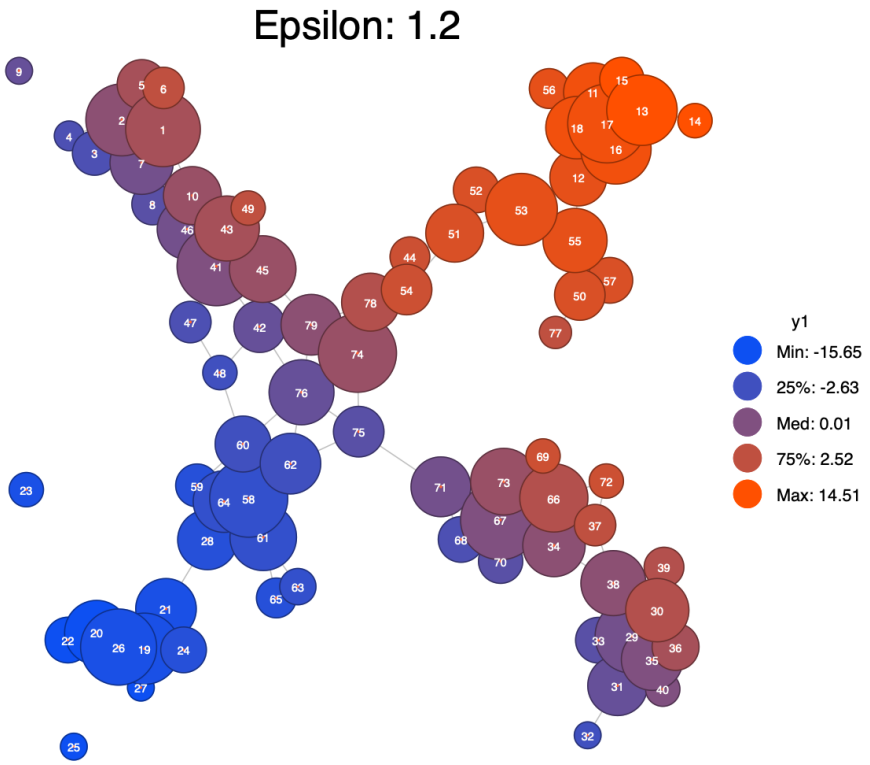
5 在 ballmapper的基础上构建

本节将重点关注 X 数据集的 $Y1$ 着色。为便于分析 TDABM 图，第一步是添加 labels 选项。labels 选项会指示 Stata 在 TDABM 图上显示球编号。添加 $\varepsilon = 1.20$ 后的图示见图7。它与图 5 的子图 (b) 唯一的区别是，现在每个球上都显示了球编号。

借助标签，我们现在可以讨论出现在球11到17中的较高数值 $Y1$ 。请注意，这里的连续编号是因为我们手动按块调整了数据。因此，下一个未覆盖的点很可能位于同一个子云中。点云的重叠产生了一个有趣的结果：来自以 (3,3) 为中心的子云中的一个点，最终成了附属于以 (6,6) 为中心的子云的球的地标。球56的编号也属于以 (3,3) 为中心的子云所对应的编号集合。所有子云中都能发现类似的模式。能够描述 TDABM 图是第一步，但我们希望充分利用算法确切知道每个球中包含哪些点这一事实。

[4] 3 18 $\sqrt{18^2 + 18^2} \approx$ 个标准差（从相距最远的中心算起），在每个轴上得 25.46 到的距离为，总距离为。

图7：带标签的 X 数据集



注：底层数据集包含 900 个观测值，涉及 2 个变量： X_1 和 X_2。数据经过平移，形成 9 组，每组包含 100 个点，中心分别位于 (-6,6)、(-3,3)、(3,3)、(6,6)、(0,0)、(-3,-3)、(3,-3)、(-6,-6) 和 (6,-6)。

5.1 `ballmapper()`的输出

ballmapper() 函数会创建两个新的框架：BM RESULTS 和 BM MERGED。BM RESULTS 框架包含 TDABM 图的信息。框架中包括球编号、TDABM 图中各点的坐标、球内的点数，以及绘图时使用的大小值。由于 Stata 使用分箱着色，BM RESULTS 数据框还将颜色箱作为一列列出。在球的列表下方，还有一个边的列表。边没有编号，但提供了其起点和终点坐标。类型列用于说明一条线在 TDABM 图中表示的是节点还是边。

BM MERGED 框架包含一个合并数据集，其中每个球的成员关系都与底层数据集相关联。这个合并后的数据框可用于分析底层数据集中未纳入 TDABM 图的变量。例如，如果各点带有文本标识符，BM MERGED 中的数据框会列出哪些点属于哪个球。若要计算每个球的汇总统计量，可以将 BM MERGED 数据框与按球分组的操作结合使用。

可以通过 Stata 中的菜单或 frame change 命令在不同框架之间切换。若要查看任一框架中的数据，可在该框架获得焦点后使用 data editor 的浏览选项。需要注意的是，每次调用 ballmapper() 函数时，这些框架都会被覆盖。因此，在开始本节中的任何分析之前，请确保最近一次运行的确是你想要分析的结果。

5.2 ballmapper() 输出摘要

由于我们已将底层数据集与 TDABM 图中球的成员关系合并为一个数据框，因此可以利用 BM MERGED 中的信息进一步分析这些球。用户可以基于合并后的数据创建自己的函数。基础程序包提供了两个摘要函数。首先，ballsummary() 命令可汇总多个变量在各个球中的均值。其次，variablesummary() 命令可对单个变量在各个球中的情况进行详细摘要。方框 5.1 从前一个函数开始介绍。

方框 5.1：使用 `ballsummary` 为多个变量生成摘要ballsummary 函数可对指定变量列表中的每个变量，计算其在各个球中的均值。输出结果将保存到指定的 .csv 文件中：`ballsummary y1 y2 y3 y4 y5, csvfile("ymeans12")`示例展示了 X 数据集中所有 Y 变量的摘要。

方框 5.1 中的代码会生成一个 .csv 文件，其中每个球占一行，共 79 行。表 2 展示了前

表 2：X 数据集 Y变量摘要

Ball	Y_1	Y_2	Y_3	Y_4	Y_5	Size
1	0.656	1	65.43	-0.016	0	44
2	0.136	1	82.75	-0.039	0	40
3	-2.102	1	79.77	-0.060	0	11
4	-2.329	1	101.5	0.384	0	3
5	1.110	1	90.63	-0.195	0	15

注：表中列出了用于 X 数据集着色的各个 5 Y变量的平均值。底层数据集包含 900 个观测值和 2 个变量，即 X_1 和 X_2。数据经过平移，形成 9 组，每组 100 个点，中心分别位于 (-6,6)、(-3,3)、(3,3)、(6,6)、(0,0)、(-3,-3)、(3,-3)、(-6,-6) 和 (6,-6)。着色定义为 $Y_1 = X_1 + X_2 + \theta$，其中$\theta \sim N(0,\ 0.2)$，Y_2为组号；$Y_3 = X_1^2 + X_2^2 + \theta$，其中 $\theta \sim N(0,\ 0.2)$；$Y_4 = \phi$，其中 $\phi \sim N(0,\ 1)$；以及 Y_5，当 $0 < X_1 < 3$和 $0 < X_2 < 3$同时满足时取值为 1。

以五行为例。可以看到，前 5 个球都位于以 (-6,6) 为中心的子云中，Stata 将其编号为 1。因此，所有球的 Y_2 都是 1，且没有点具有 $Y_5 = 1$。我们看到球 1 有 44 个点，而球 4 只有 3 个点。图7中可以明显看出大小上的差异。

对 TDABM 图进行的第二项常规分析，是考察球内的变异。如果 TDABM 图捕捉到了着色变量在空间中的变异，我们预期每个球内的 Y 取值范围都较小。各个球的 Y 取值之间重叠较少，也表明 Y 在空间中存在系统性变异。框5.2提供了用于生成数据集中 Y_5 摘要的代码。

框5.2：使用 variablesummary生成变量的详细摘要

若要更详细地汇总各个球中的单个变量，可使用 variablesummary() 命令。还可以选择生成箱线图，以汇总每个球中的变量。输出结果将保存到指定的 CSV 文件中。

```
variablesummary y5, boxplot boxfile("y5_12_box") csvfile("y5_12_stats")
```

示例展示了 X 数据集中 Y_5 变量的详细摘要。

使用框5.2中的代码，我们可以详细汇总各个球中的单个变量。摘要列出了每个球中该变量的均值、标准差、最小值、最大值、四分位数和中位数。由于球中的点数可能较少，因此解读这些数值时应谨慎。为便于解读，结果表格的最后一列列出了每个球中的点数。与 ballsummary() 函数一样，该代码应用于最近生成的 TDABM 图。表3列出了 Y_1 变量在球1至球5中的摘要。使用 variablesummary() 命令生成的Y_4 和 Y_5 箱线图见图8。

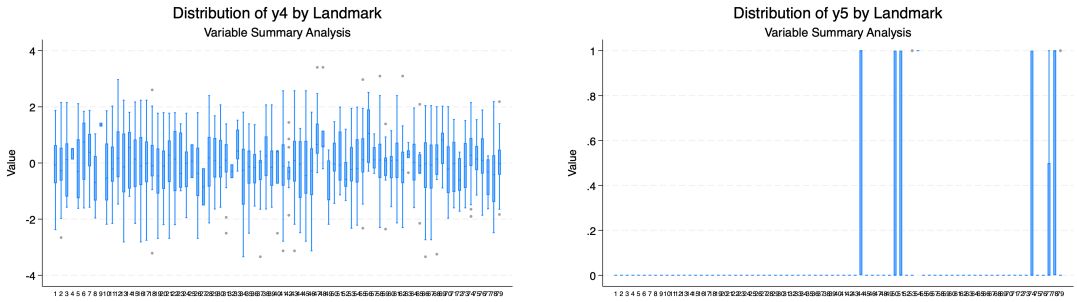
球 1 至 5 都位于以 (-6,6) 为中心的子云中。这个子云中心附近的球，其 Y_1 值会接近 0。在子云的边缘，X_1 和 X_2 的总值会受到两者差值的影响，因此我们看到较小的球，其 Y_1 离 0 更远。球 3 可以看作是一个

表3： Y_1 摘要 X 数据集中的示例表格行

Ball	Mean	标准差	Min	q25	q50	q75	Max	Size
1	0.656	0.770	-0.696	0.030	0.542	1.228	2.699	44
2	0.136	0.775	-1.652	-0.578	0.158	0.785	1.732	40
3	-2.101	0.752	-3.379	-2.608	-2.279	-1.431	-0.924	11
4	-2.329	0.593	-2.756	-2.756	-2.579	-1.652	-1.652	3
5	1.110	0.520	0.344	0.918	1.053	1.376	2.456	15

注：数值给出了 TDABM 图中每个球对应的 Y_1 均值、标准差、最小值、最大值、四分位数和中位数。 $Y_1 = X_1 + X_2 + \theta$，其中 $\theta \sim N(0,0.2)$ 是随机噪声分量。底层数据集包含 2 个变量的 900 个观测值， X_1 和 X_2。数据经过平移，形成 9 组，每组包含 100 个点，中心分别位于 (-6,6)、(-3,3)、(3,3)、(6,6)、(0,0)、(-3,-3)、(3,-3)、(-6,-6) 和 (6,-6)。

图8：X 数据中各球的 Y 箱线图 set



(a) Y_4 - 噪声 (b) Y_5 - 受限范围

注：图形绘出了 TDABM 图中每个球对应变量的最小值、最大值、四分位数和中位数。底层数据集包含 2 个变量 X_1 和 X_2 的 900 个观测值。数据经过平移后，分成了 9 组，每组 100 个点，中心分别位于 (-6,6)、(-3,3)、(3,3)、(6,6)、(0,0)、(-3,-3)、(3,-3)、(-6,-6) 和 (6,-6)。 $Y_4 \sim N(0,1)$ 与球编号完全独立。对于任意一个 X_1 和 X_2 介于 0 到 3 之间的点， Y_5 的值为 1。

示例：它覆盖了点云中 Y_1 为负值的部分，而球 1 和球 2 所覆盖的部分则同时包含 Y_1 的负值和正值。球 5 则是一个示例，它完全位于 Y_1 为正值的空间区域中，但又足够靠近子云中心，因此仍有 15 个点。附录中收录了完整表格，

图8由此生成.....

6 AutoData示例

为了演示如何在 Stata 内置数据集上使用 TDABM，我们将考察汽车数据集。汽车数据集包含1978年可供销售的汽车车型观测值。该数据集包含12个变量，共有74个观测值。不过，有5个观测值缺少品牌变量 rep78。我们将 rep78 删除。品牌变量是文本类型，因此不会出现在汇总统计量表中，也不适用于 TDABM。foreign变量是一个虚拟变量，仅适合用于着色。Headroom 的取值有限，因此从分析中删除。用于 TDABM 分析的变量的汇总统计量

表4：汽车数据集汇总统计量

变量	简短	Mean	标准差	Min	Max
建议零售价	价格	6165	2950	3291	15906
燃油经济性	mpg	21.30	5.786	12	41
Headroom	Headroom	2.993	.846	1.5	5
后备箱容积	后备箱	13.76	4.277	5	23
轴重（千克）	重量	3020	777.2	1760	4840
车辆长度（英寸）	长度	187.9	22.27	142	233
转弯直径（码）	turn	39.65	4.399	31	51
排量	排量	197.3	91.84	79	425
齿轮比	齿轮比	3.015	0.456	2.19	3.89
非美国制造商	进口车	.297	.461	0	1

注：我们对1978年汽车内置汽车数据集进行分析时所用变量的汇总统计量。有关变量的进一步说明，请参阅 Stata 文档。 $N = 74$。

表5：汽车数据集相关矩阵

	价格	mpg	后备箱	重量	长度	turn	排量	齿轮比	进口车
价格	1.0000								
mpg	-0.4686	1.0000							
后备箱	0.3143	-0.5816	1.0000						
重量	0.5386	-0.8072	0.6722	1.0000					
长度	0.4318	-0.7958	0.7266	0.9460	1.0000				
turn	0.3096	-0.7192	0.6011	0.8574	0.8643	1.0000			
排量	0.4949	-0.7056	0.6086	0.8949	0.8351	0.7768	1.0000		
齿轮比	-0.3137	0.6162	-0.5087	-0.7593	-0.6964	-0.6763	-0.8289	1.0000	
进口车	0.0487	0.3934	-0.3594	-0.5928	-0.5702	-0.6311	-0.6138	0.7067	1.0000

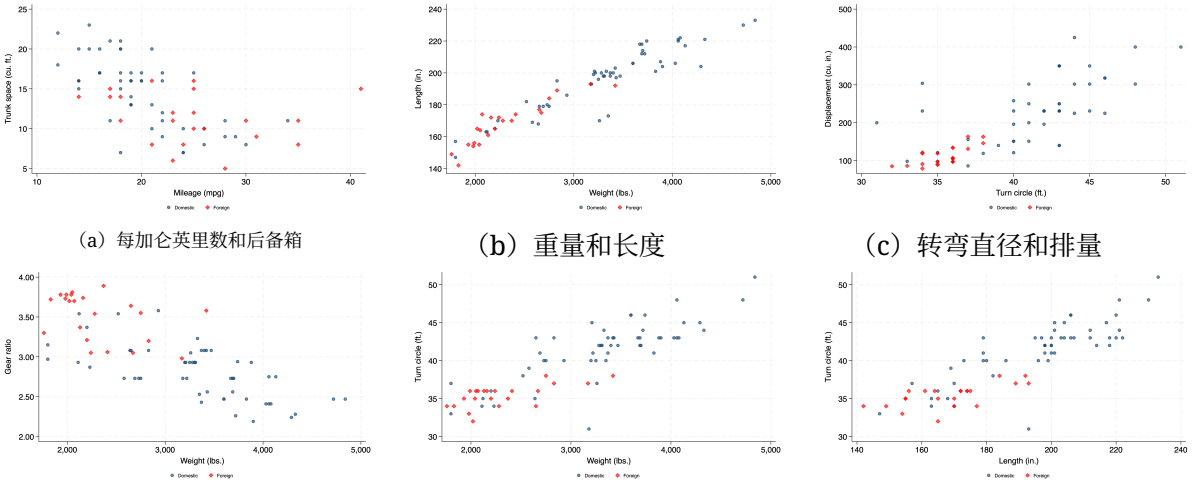
注：相关矩阵展示了我们在分析1978年汽车的内置汽车数据集时所使用的变量。有关变量的进一步说明，请参阅 Stata 文档和表4。
$N = 74$。

X 见表4。本节的代码可在 GitHub 网站上的 autodata.do 文件中找到。

表4显示，数据集中的汽车有29.7%由外国制造商生产。汽车的平均价格为6165美元。最低价格为3291美元，最高价格为15906美元。价格较低是因为数据来自1978年。TDABM 图的坐标轴变量所采用的尺度各不相同。例如，最高齿轮比为3.89，而最小重量为1760。因此，在运行 ballmapper() 函数之前，有必要对变量进行标准化。本节的 .do文件中包含了标准化代码。

表5中的相关矩阵显示，燃油经济性与长度、重量、转弯直径和排量之间存在显著负相关。长度与后备箱容积、重量、转弯直径、排量和齿轮比呈正相关。重量与转弯直径和排量呈正相关。最后，转弯直径与排量之间存在显著正相关，而齿轮比与排量及转弯直径之间则存在显著负相关。价格与

图9：选定的成对散点图



（a）每加仑英里数和后备箱　　（b）重量和长度　　（c）转弯直径和排量

(d)齿轮数和重　ght　　(e) Weig量和转弯直径　　(f) Leng径与转弯直径

注：选取了用于分析1978年汽车内置汽车数据集的变量两两散点图。有关这些变量的更多说明，请参阅 Stata 文档和表4。点的颜色表示汽车是否由外国制造商生产，红色表示进口车，蓝色表示国产车。 $N = 74$。

后备箱容积、重量、长度、转弯直径和排量呈正相关。同时，价格与燃油经济性和齿轮比之间存在负相关。外国品牌与较高的每加仑英里数、较小的后备箱容积、较低的重量、较短的长度、较小的转弯直径和较小的排量相关。进口车与齿轮比之间存在正相关。

证据，尤其是我们的 X 示例以及 Matejka 和 Fitzmaurice（2017），强调了考察 X的联合分布的重要性。Anscombe（1973）提醒我们，应通过图形考察 X 与 Y 之间关系的重要性。为了初步探索数据中的关系，我们为数据集中的所有变量绘制了两两散点图。图9展示了部分图。其余两两散点图的绘制代码包含在 GitHub 网站上的 .do文件中。

图9显示，每加仑英里数与后备箱容积之间，以及齿轮数与重量之间存在负相关。不过，相关性较弱这一点也很明显。图9的其他子图中，相关性更强。外国汽车与国产汽车的分化也显现出来。可以看到，外国汽车分布在空间中重量较轻、排量较低、转弯直径较小的区域。本节将探讨，组合这些变量是否能更充分地区分外国汽车与国产汽车，以及特征的联合分布是否能提供有关价格的信息。

框 6.1：使用ballmapper()生成 TDABM

为准备数据，先创建一个外国车虚拟变量，并对坐标轴变量进行标准化：生成 double for_nu

```
m = ( foreign == 1) foreach v in 每加仑英里数 后备箱容积 重量 长度 转弯直径 排量 齿轮数_ratio
{ egen std_‘v’ =std (‘v’) }然后生成 TDABM 图，使用 ε = 1.50。这里，foreign变量用于着色：
ballmapper std_\ast , epsilon (1.5) color ( for_num ) layout repulsion (0.05) attraction (0.01)
filename (" foreign_std15 ") labels
```

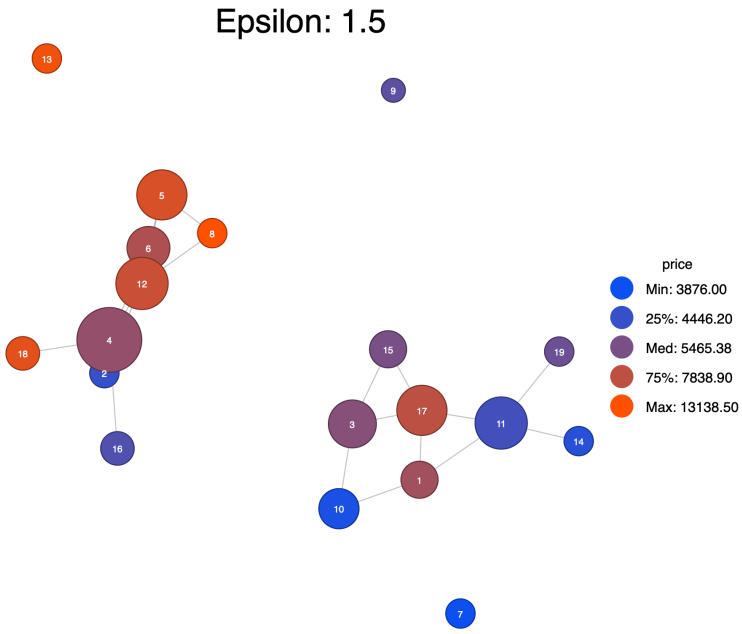
使用框 6.1 中的代码，我们为汽车数据构建 TDABM 图。我们总共生成了 10 张图，使用的半径为 $\varepsilon = 1$、1。5、2、2。5、3，并在图 10 和图 11 中展示结果。图 10 根据球内的平均价格着色。图 11 根据球内由外国制造商生产的车型比例着色。

图 10 显示，在 $\varepsilon = 1.50$，数据排列成两个连通分量。随着半径增加，这两个连通分量逐渐形成，但在 $\varepsilon = 2.00$ 时合并。随着半径增加至 $\varepsilon = 2.50$ 和 $\varepsilon = 3.00$，即子图 (d) 和 (e)，球的数量进一步减少。可以看出存在一个稠密核心，两端各有较小的球；只有到了 $\varepsilon = 3.00$，两个极端中的一个才合并进来。我们关注的是 $\varepsilon = 1.50$ 情形。子图 (a) 显示，两个较大组中，左侧的价格较高。右侧组中蓝色球较多，价格较低。在三个彼此分离的球中，球 13 的价格较高，而球 7 和 8 的价格较低。我们注意到，两个连通分量内部都存在价格变异：球 2 和球 16 的价格较低，却与价格较高的球相连；球 17 则位于价格较低的组中，其价格接近中位数。

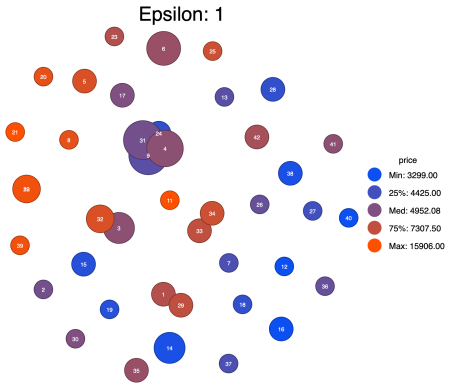
观察品牌国别可以清楚地看到，两个连通分量组之间存在明显的分隔。左侧的组几乎全部是国产车，其中包括球 2 和球 16 中价格较低的汽车。右侧的组几乎全部是外国制造的汽车，除了球 10 以及其与球 1 和球 3 的重叠部分。所有不相连的球都是国产车。尽管这种分隔并不完整，但图 11 的子图 (a) 说明了为什么分类器可能能够利用这些数据取得良好表现。将半径增大到 $\varepsilon = 2.00$ 后可以看到，外国汽车位于空间的右侧，而国产车位于左侧。因此，线性判别分类器有可能在这些数据上拟合出一个平面。子图 (d) 和 (e) 表明，这种模式在更高的 ε下仍然成立。

为了获取有关这些球的更多信息，我们可以使用 ballsummary() 命令计算每个球的平均值。代码见方框 6.2。所得摘要列于表 6 中。

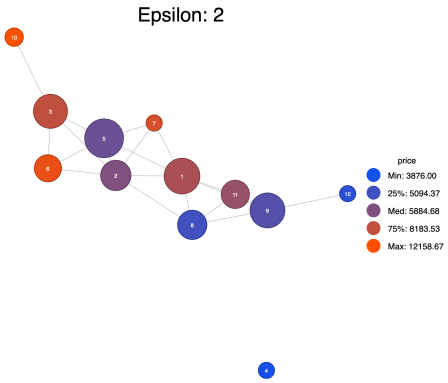
图 10：汽车数据的 TDABM 图：价格



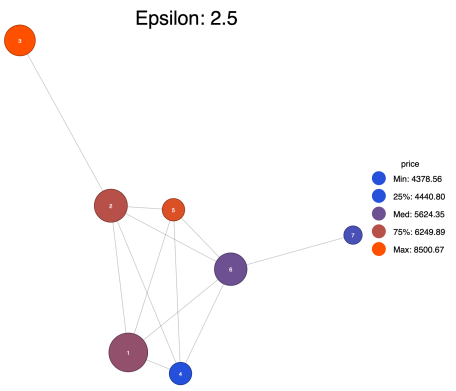
(a) $\varepsilon = 1.50$



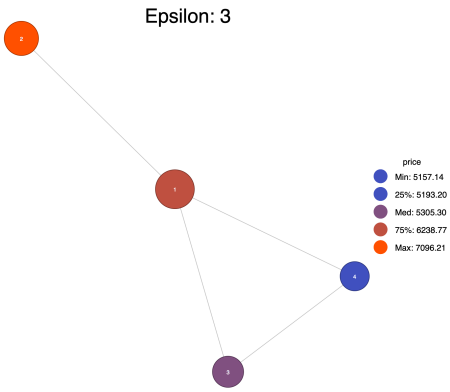
(b) $\varepsilon = 1.00$



(c) $\varepsilon = 2.00$



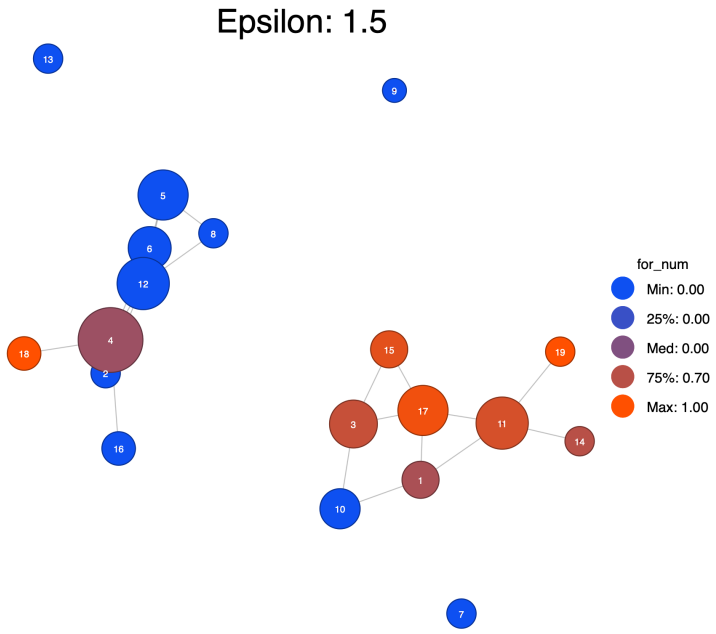
(d) $\varepsilon = 2.50$



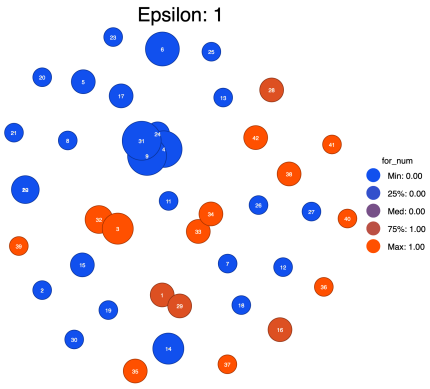
(e) $\varepsilon = 3.00$

注：使用所述球半径绘制汽车数据的 TDABM 图。所用坐标轴变量为每加仑英里数、后备箱、重量、长度、转弯直径、排量和齿轮比。应用 `ballmapper()` 函数前，所有坐标轴变量均已标准化。颜色根据以美元计的平均价格着色。分析使用 Stata 内置的 1978年汽车数据集。有关这些变量的进一步说明，请参阅 Stata 文档和表4。 $N = 79$

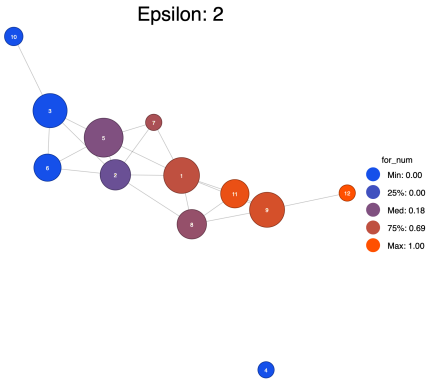
图 11：汽车数据的 TDABM 图：外国品牌



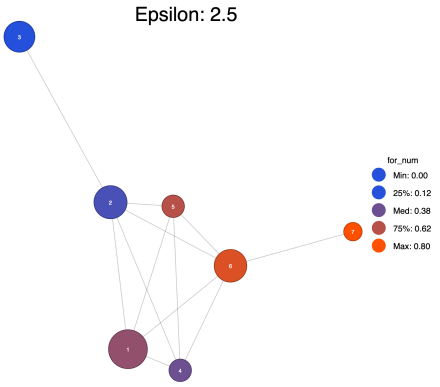
(a) $\varepsilon = 1.50$



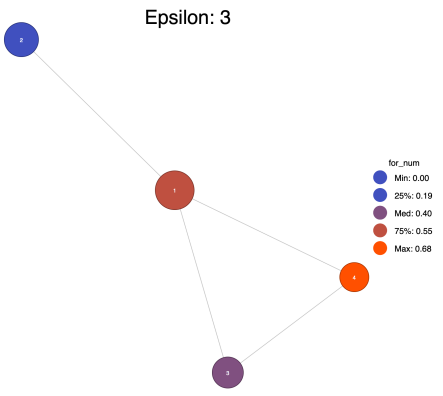
(b) $\varepsilon = 1.00$



(c) $\varepsilon = 2.00$



(d) $\varepsilon = 2.50$



(e) $\varepsilon = 3.00$

注：采用所述球半径绘制汽车数据的 TDABM 图。所用坐标轴变量为每加仑英里数、后备箱、重量、长度、转弯直径、排量和齿轮比。在应用 `ballmapper()` 函数之前，所有坐标轴变量均已标准化。着色依据球内由外国制造商生产的观测值比例确定。分析使用 Stata 内置的 1978年汽车数据集。有关这些变量的更多说明，请参阅 Stata 文档和表4。 $N = 79$

表 6：汽车数据的球体汇总统计

Ball	mpg	后备箱	重量	长度	turn	排量	齿轮比	价格	进口车	Size
1	22.50	9.250	2713	182.8	39.75	131.5	3.438	5725	0.25	4
2	18.00	12.00	3390	185.0	41.50	254.0	2.545	4352	0	2
3	22.38	11.88	2573	170.8	36.13	127.3	3.129	5465	0.625	8
4	18.82	15.29	3283	198.5	41.65	211.7	2.939	5619	0.118	17
5	15.67	18.78	3949	213.3	43.78	333.0	2.416	8551	0	9
6	17.17	19.83	3728	214.8	43.17	233.2	2.770	6483	0	6
7	27.50	9.500	2170	166.5	34.00	267.5	2.900	3876	0	2
8	18.50	15.00	4160	205.0	44.00	350.0	2.325	13139	0	2
9	22.00	17.00	3180	193.0	31.00	200.0	2.730	4504	0	1
10	23.00	8.600	2680	177.2	40.60	146.6	2.800	4010	0	5
11	29.30	8.700	2056	160.1	34.40	96.90	3.524	4446	0.700	10
12	16.20	16.70	3833	209.0	44.10	309.2	2.504	7839	0	10
13	12.00	20.00	4780	231.5	49.50	400.0	2.470	12546	0	2
14	32.50	10.00	2000	161.0	36.00	91.50	3.090	4087	0.5	2
15	24.00	16.00	2063	159.3	35.75	99.00	3.575	5081	0.75	4
16	18.67	11.00	3430	200.3	42.33	231.0	3.080	4480	0	3
17	23.89	10.67	2302	171.1	35.78	111.6	3.684	6738	0.889	9
18	16.00	14.33	3140	191.3	37.33	152.3	3.253	11558	1	3
19	38.00	13.00	2045	159.5	35.50	93.50	3.795	4598	1	2

注：使用 $\varepsilon = 1.50$ 对汽车数据进行 TDABM 分析后得到的 19 个球的汇总统计量。有关这些变量的进一步说明，请参阅 Stata 文档和表4。 $N = 74$。

方框 6.2：使用`ballsummary()`生成球摘要

运行 $\varepsilon = 1.50$ 处的 ballmapper() 后，我们可以生成摘要。请注意，该摘要将

仅适用于最近一次执行的 ballmapper()：

```
ballsummary 每加仑英里数 后备箱 重量 长度 转弯直径 排量 gear_ratio 价格 进口车，csvfile("
auto_means12 ")
```

摘要已写入文件 auto means.csv。

表 6 显示，各个球中的所有变量都存在显著变异。这是意料之中的，因为这些球覆盖了空间的不同部分。然而，表 6 证实了这种变异适用于 TDABM 的多维应用。在前面的讨论中，球 2 和球 16 被强调为与价格较高的连通分量相连的低价球。我们看到，球 2 仅包含 2 辆汽车，且均为国产。球 16 包含 3 辆汽车。通过查看 Stata 中合并后的数据框，我们看到球 2 包含 AMC Pacer 和 Chevrolet Nova。球 16 中的 3 辆汽车是 Buick Skylark、Pontiac Firebird 和 Pontiac Pheonix。对于熟悉 1978年美国汽车的人来说，还可以进一步探索。就我们的目的而言，这说明了如何从数据中轻松获取有关这些球的信息。

球 1、14 和 15 值得关注，因为每个球最多包含 4 辆汽车，且国产汽车和进口汽车混杂在一起。球 1 包含 AMC Concord、Ford Mustang、Plymouth Sapporo 和 Datsun

810。Datsun 810 是这一组中的进口车。球14包含 Plymouth Champion 和 Toyota Corolla。球15包含 Plymouth Horizon、Fiat Strada、Volkswagen Rabbit 和 Volkswagen Sirocco。在这里，观察哪些进口汽车与国产汽车具有相似特征很有意思，因为对于那些在特征空间这一部分寻找汽车的顾客而言，它们会是竞争对手。我们使用汽车数据集，可以探讨汽车之间的竞争，但由于数据有限，无法揭示因果关系。

第二个摘要针对单个变量生成，但提供了更详细的汇总统计量。我们希望进一步了解结果变量——价格以及进口车所占比例。框6.3中的代码生成表7，同时也启用了生成箱线图的选项。

框6.3：使用`variablesummary()`生成球摘要在 $\varepsilon = 1.50$ 运行 ballmapper() 后，我们可以生成单个变量的详细摘要。

```
variablesummary foreign , boxplot boxfile("foreign_15_box")csvfile("foreign
_15_stats")variablesummary price , boxplot boxfile("price_15_box")csvfile("price_15_stats")
```

摘要会保存到指定的 .csv 文件中，箱线图则保存为指定的 .png 文件。

表7分为两个子图。数据显示，球内存在变异，标准差从不到平均价格的10%到价格的33%不等。这表明，用于生成 TDABM 图的特征与汽车价格确实存在一定关系。不过，我们也应认识到，还有其他影响价格但未被纳入考量的因素。球内价格的变异表明，建模价格还需要更多解释变量。对于外国所有权，我们看到大多数球要么完全由国产车构成，要么完全由进口车构成。我们已经考察过一些混合球。球内价格的变异也体现在图 12 中相应的箱线图上。

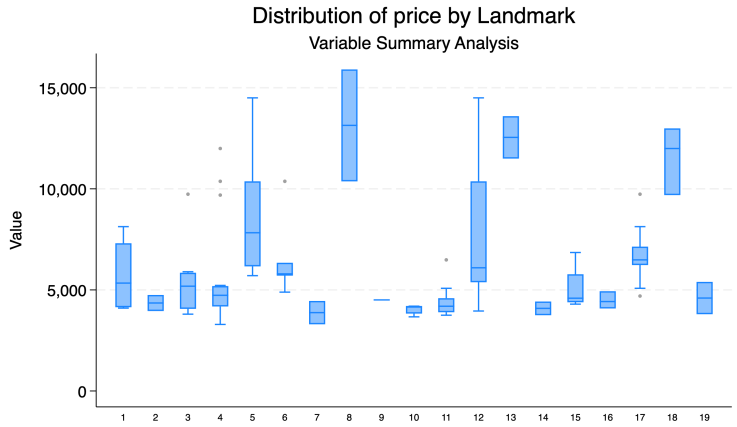
图 12 显示，有些球的价格较高，有些则较低。这与图 10 中 TDABM 图所传达的信息一致。球5 和球 12 所包含的价格范围最大。球5 有 9 辆汽车，球12 有 10 辆。两者中的汽车均为国产。因此可以认为，这些球代表了来自不同声望等级品牌的汽车，这些品牌对同款汽车收取较高价格的能力也各不相同。有几款汽车同时出现在这两个球中，包括价格处于范围低端的 Dodge Magnum，以及价格处于范围高端的 Cadillac Eldorado。TDABM 不考虑品牌因素，但汽车在空间中的位置可以启发用户讲述这些汽车为何会出现在同一部分

表7：汽车数据的结果摘要表

Y	Ball	Mean	标准差	Min	q25	q50	q75	Max	N_b
价格（美元）	1	5725	1946.6	4099	4143	5336.5	7307.5	8129	4
	2	4352	561.4	3955	3955	4352	4749	4749	2
	3	5465	1930.4	3799	4062.5	5183	5849	9735	8
	4	5619	2506.9	3291	4181	4733	5189	11995	17
	5	8551	3026.5	5705	6165	7827	10371	14500	9
	6	6483	1961.6	4890	5705	5793	6342	10372	6
	7	3876	816.0	3299	3299	3876	4453	4453	2
	8	13139	3913.8	10371	10371	13139	15906	15906	2
	9	4504		4504	4504	4504	4504	4504	1
	10	4010	246.1	3667	3829	4172	4187	4195	5
	11	4446	831.1	3748	3895	4192	4589	6486	10
	12	7839	3799.4	3955	5379	6094.5	10371	14500	10
	13	12546	1482.8	11497	11497	12546	13594	13594	2
	14	4087	478.7	3748	3748	4086.5	4425	4425	2
	15	5081	1190.5	4296	4389	4589.5	5773.5	6850	4
	16	4480	428.8	4082	4082	4424	4934	4934	3
	17	6738	1521.8	4697	6229	6486	7140	9735	9
	18	11558	1692.8	9690	9690	11995	12990	12990	3
	19	4598	1130.7	3798	3798	4597.5	5397	5397	2
进口车	1	0.25	0.5	0	0	0	0.5	1	4
	2	0	0	0	0	0	0	0	2
	3	0.625	0.518	0	0	1	1	1	8
	4	0.118	0.333	0	0	0	0	1	17
	5	0	0	0	0	0	0	0	9
	6	0	0	0	0	0	0	0	6
	7	0	0	0	0	0	0	0	2
	8	0	0	0	0	0	0	0	2
	9	0		0	0	0	0	0	1
	10	0	0	0	0	0	0	0	5
	11	0.700	0.484	0	0	1	1	1	10
	12	0	0	0	0	0	0	0	10
	13	0	0	0	0	0	0	0	2
	14	0.500	0.708	0	0	0.5	1	1	2
	15	0.750	0.500	0	0.5	1	1	1	4
	16	0	0	0	0	0	0	0	3
	17	0.889	0.333	0	1	1	1	1	9
	18	1	0	1	1	1	1	1	3
	19	1	0	1	1	1	1	1	2

注：使用 $\varepsilon = 1.50$ 对汽车数据进行 TDABM 分析后得到的 19 个球的汇总统计量。图形列出了每个球的均值、标准差、最小值、第25百分位数、中位数、第75百分位数和最大值。分析涉及两个变量，即 TDABM 图中的结果变量：价格和外资持股比例。有关这些变量的更多说明，请参阅 Stata 文档和表4。 $N = 74$。

图 12：汽车数据中各球内价格变动的箱线图



注：使用 $\varepsilon = 1.50$ 对汽车数据进行 TDABM 分析后得到的 19 个球内价格变动的箱线图。图中基于每个球的最小值、第25百分位数、中位数、第75百分位数和最大值绘制。有关变量的进一步说明，请参阅 Stata 文档和表4。 $N = 74$。

特征空间。

本节的分析展示了如何使用 ballmapper() 函数生成 TDABM 图，并利用 ballmapper 软件包的其他功能对其进行探查。随后，用户可凭借集体专业知识进一步探查和探索数据。我们的分析展示了建模特征在分类器中的潜力。我们还指出，在对价格进行建模时，需要考虑其他变量。

7 摘要

本指南介绍了 ballmapper 软件包，并展示了如何在 Stata 中应用拓扑数据分析球映射器（TDABM）（D-lotko，2019年）。TDABM 是一种无需模型的多变量数据可视化与理解方法。Anscombe（1973）和 Matejka 和 Fitzmaurice（2017年）清楚地阐述了数据可视化的重要性。TDABM 提供了一种方法，可突破页面上只能呈现 2 个维度的限制。由此生成的数据空间抽象映射可用于推动讨论性分析、为建模提供参考并评估模型。文献中已应用这些功能（Qiu 等人，2020；Rudkin 等，2024a,b；Benites 等人，2025年；Tubadji and Rudkin，2025年），此外还有许多其他研究。ballmapper 软件包使 Stata 用户也能享有这些优势。

Stata 代码是一个基础实现，支持 TDABM 应用中最常见的工作流程。对于希望考察地标选择的影响，或针对不同半径开展广泛分析的用户，则需要编写更多函数，或使用 Python 或 R 库[5]。在本

[5]Python 相关内容参见 Rudkin（2025a），R 相关内容参见 Rudkin（2025b）。

指南中，我们没有考虑统计模型的评估。不过，这很容易实现，因为你会有一列预测值和残差，可将它们用作 TDABM 中的着色变量。箱线图和汇总统计可用于直观检查残差是否确实独立于特征的联合分布。未来的工作将考虑扩展 ballmapper 软件包，以涵盖方法学创新。

参考文献

Altman, E. I.（1968年）。财务比率、判别分析与企业破产预测。《金融学杂志》，23(4):589–609。

Anscombe, F. J.（1973年）。统计分析中的图表。《美国统计学家》，27(1):17–21。Benites, A.、Hara, K.、Singh, P.、Zubier, B. 和 Rudkin, S.（2025年）。教育成果不平等的拓扑学。可通过 SSRN 5401815 获取。

Charmpi, M.、Giannouli, P. 和 Xanthopoulos, S.（2023年）。信用评分的拓扑学方法。可通过 SSRN 4629581 获取。

D-lotko, P.（2019年）。Ball Mapper：用于拓扑数据分析的形状摘要。arXiv 预印本 arXiv:1901.07410。

Dlotko, P.、Qiu, W. 和 Rudkin, S.（2022年）。用于金融的拓扑数据分析球映射器。arXiv 预印本 arXiv:2206.03622。

D-lotko, P., Qiu, W. 和 Rudkin, S.（2024）。通过拓扑数据分析视角重新评估财务比率与股票收益。《欧洲金融学杂志》，第 1–25 页。

Fama, E. F. 和 MacBeth, J. D.（1973年）。风险、收益与均衡：实证检验。《政治经济学杂志》，81(3):607–636。

Han, X.、Suer, A.、Foreman, B. 和 Jia, X.（2025年）。颅内压形态变化的拓扑可视化及实时数据轨迹映射。《IEEE生物医学与健康信息学杂志》。

Hartigan, J. A. 和 Wong, M. A.（1979）。AS 136算法：k均值聚类算法。《英国皇家统计学会杂志》。C辑（应用统计），28(1)：100–108。

Koljanˇci´c, N.、Park, S. A.、Gurnari, D.、D-lotko, P.、Hahn, J. 和 ˇSpa´anik, I.（2025年）。通过拓扑数据分析对二维气相色谱联用高分辨质谱数据进行贵腐葡萄酒的非靶向化学分析。《分析化学》。

Madukpe, V. N., Zulkepli, N. F. S., Noorani, M. S. M., 和 Gobithaasan, R.（2025年）。球映射器与传统映射器在调查空气污染物行为方面的比较分析。《环境监测与评估》，197(2):136。

Matejka, J. 和 Fitzmaurice, G.（2017年）。相同的统计数据，不同的图形：通过模拟退火生成外观各异但统计特征相同的数据集。载于《2017年CHI人类因素与计算系统会议论文集》，第1290–1294页。

Otway, A. 和 Rudkin, S.（2024）。左倾滞后的形态：2019年英国大选与英国选举地理的变化。可见 SSRN 4877434。

Qiu, W., Rudkin, S., 和 D-lotko, P.（2020）。通过对奥特曼Z值模型进行拓扑数据分析制图，深化对企业失败的理解。《专家系统与应用》，第113475页。

Rudkin, S.（2025a）。Python 中拓扑数据分析球映射器简介。arXiv 预印本 arXiv:2505.03022。

Rudkin, S.（2025b）。《R 中拓扑数据分析球映射器简介》。arXiv 预印本，arXiv:2504.14081。Rudkin, S., Barros, L., D-lotko, P., 和 Qiu, W.（2024a）。《脱欧投票的经济拓扑》。《区域研究》，58(3):601–618。Rudkin, S. 和 Dlotko, P.（2024）。《包容性的拓扑：欧洲地区与数字鸿沟》。载于《2024年第四届数字数据处理国际会议》（DDP），第 123–128 页。IEEE。Rudkin, S., Rudkin, W., 和 D-lotko, P.（2024b）。《比特币收益轨迹与比特币收益的可预测性》。《金融评论》。Rudkin, S. 和 Webber, D. J.（2023）。《区域增长路径与区域韧性》。可见于 SSRN 4333276。Singh, G., M´emoli, F., 和 Carlsson, G. E.（2007）。《用于高维数据集分析和三维物体识别的拓扑方法》。SPBG，91:100。

Tubadji, A. 和 Rudkin, S.（2025年）。《文化引力与通过迁移实现的增长再分配：空间计量经济学和拓扑数据分析带来的凝聚力启示》。《区域科学论文集》，104(1):100064。

Valerio, J., Vasconcelos-Filho, J. E., Stosic, B., de Oliveira, W. R., Santana, F. M., Antonino, A. C., 和 Duarte-Neto, P. J.（2025年）。鱼类耳石三维放射密度分布的拓扑分析：点采样对降维的影响。《Micron》，188：103731。

A 附加表格

本附录收录了正文中经删节的完整球汇总表。由于完整数据集中有 79 个数据点，此处仅提供全部 79 个球的列表。表 A1 和表 A2 汇总了 5 Y 变量，表 A3 和表 A4 则详细汇总了数据集中的 Y_1值。

表 A1 和表 A2 展示了球中呈现的预期结构。 Y_1 的均值范围为 -3 到 3，对应球位于子云任一边缘的情况。 Y_2 的值通常为整数，这意味着球只覆盖其中一个子云。不过，当球覆盖多个子云时，也会出现小数。由于子云的编号方式，研究这些小数值并没有太大的直接意义。 Y_3 是二次的，这意味着所有均值均为正。对于 X 形两端的子云，其平方和远大于靠近中心的子云。因此，我们会看到像 14 这样的球，其 Y_3 均值为 104.8。 Y_4 是噪声，因此不应观察到任何模式。我们发现确实没有模式。对于 Y_5，任何具有 $0 < X_1 < 3$ 和 $0 < X_2 < 3$的点，其值都为 1。许多具有 $Y_5 = 1$ 的点位于以 (0,0) 为中心的子云中，且 Y_3的值较低。其他点则位于以 (3,3) 为中心的子云中，且 Y_3的值较高。因此，我们看到了预期的模式。

表 A3 和表 A4 也展示了预期的模式。当球靠近 X 形点云左上至右下对角线上的中心时， Y_1 的均值接近 0。左下方的球具有较大的负 Y_1 值，例如球 19 至 27。与此同时，可以看到球 10 至 17 的 Y_1 值较高。球 10 至 17 位于 X 形右上方、以 (6,6) 为中心的子云中。一个 Y_1 球内的取值范围相应

B 完整直观阐释

为了详细说明拓扑数据分析球映射器（TDABM）图的构建过程，本附录将以一个二维数据集为例，逐步进行讲解。所用数据与正文直观阐释部分所用数据相同。形式上，该数据集包含 1000 个观测值（$N = 1000$），涉及 $K = 2$ 个变量； $N \times K$ 数据集为 X。每个变量均独立地取自标准正态分布，因此 $X_1 \sim N(0,1)$ 且 $X_2 \sim N(0,1)$。所得点云是一个二维高斯云。图 A1 和图 A2 逐步展示了球的添加过程。

X中的第一个数据点周围创建了第一个球。在子图 (a) 中，该球以蓝色显示，并用一个大菱形标出地标 1。所有蓝色点都被球 1 覆盖。在子图 (b) 中，第二个

表A1：X 数据集的Y变量摘要（球1至40）

Ball	Y_1	Y_2	Y_3	Y_4	Y_5	Size
1	0.656	1	65.43	-0.016	0	44
2	0.136	1	82.75	-0.039	0	40
3	-2.102	1	79.77	-0.060	0	11
4	-2.329	1	101.5	0.384	0	3
5	1.110	1	90.63	-0.195	0	15
6	2.524	1	78.00	0.379	0	9
7	-0.614	1.138	64.06	0.389	0	29
8	-1.702	3.222	42.15	-0.594	0	9
9	-1.027	1	128.7	1.366	0	2
10	0.644	3.609	38.40	-0.292	0	23
11	11.28	2.167	65.44	-0.059	0	24
12	9.193	3.455	43.60	0.352	0	22
13	12.92	2	83.73	-0.052	0	38
14	14.51	2	104.8	0.271	0	5
15	13.66	2	94.98	-0.063	0	11
16	11.75	2	69.47	-0.089	0	38
17	11.81	2	70.03	0.016	0	50
18	10.05	3.143	51.11	0.075	0	28
19	-12.17	3	75.39	-0.385	0	39
20	-13.01	3	85.15	-0.128	0	30
21	-10.86	3.154	59.37	0.083	0	26
22	-14.03	3	96.89	0.001	0	12
23	-10.83	3	61.05	0.046	0	5
24	-10.48	3	56.37	-0.156	0	12
25	-15.65	3	124.5	0.064	0	2
26	-12.82	3	82.40	-0.320	0	46
27	-14.01	3	97.54	-0.850	0	2
28	-7.179	6.680	26.52	0.165	0	25
29	-0.535	4	71.43	0.122	0	43
30	1.349	4	77.05	0.091	0	29
31	-1.474	4	85.80	0.020	0	25
32	-2.753	4	115.5	-0.293	0	2
33	-1.862	4	63.13	0.566	0	12
34	0.008	7.286	31.42	-0.371	0	28
35	0.006	4	87.18	-0.087	0	27
36	1.082	4	93.88	-0.131	0	13
37	1.761	6.667	34.48	-0.568	0	9
38	0.130	4.258	55.13	0.261	0	31
39	1.726	4	61.58	-0.068	0	8
40	-0.151	4	101.60	-0.393	0	5

注：本表列出了用于对 X 数据集着色的各个 5 Y变量的平均值。底层数据集包含 2 个变量 X_1 和 X_2，共 900 个观测值。数据经过平移，形成 9 组，每组包含 100 个点，中心分别位于 (-6,6)、(-3,3)、(3,3)、(6,6)、(0,0)、(-3,-3)、(3,-3)、(-6,-6) 和 (6,-6)。着色表示为 $Y_1 = X_1 + X_2 + \theta$，其中 $\theta \sim N(0,\ 0.2)$，Y_2 表示组号；$Y_3 = X_1^2 + X_2^2 + \theta$，其中 $\theta \sim N(0,\ 0.2)$；$Y_4 = \phi$，其中 $\phi \sim N(0,\ 1)$；而 Y_5 在 $0 < X_1 < 3$ 和 $0 < X_2 < 3$ 均满足时取值为 1。

表 A2：X 数据集 Y变量摘要（球 41–79）

Ball	Y_1	Y_2	Y_3	Y_4	Y_5	Size
41	-0.498	5	18.49	-0.091	0	50
42	-1.092	6.412	6.601	-0.243	0	17
43	0.774	5	22.43	-0.007	0	31
44	2.930	7	6.70	-0.285	0.625	8
45	0.415	5	13.93	-0.169	0	33
46	-0.968	4.692	29.62	-0.278	0	26
47	-2.397	5	18.74	1.010	0	9
48	-2.626	5.400	10.99	1.101	0	5
49	2.132	5	26.19	-0.853	0	5
50	5.161	6.188	14.60	0.130	0.313	16
51	5.026	6	13.67	0.115	0.391	23
52	5.781	6	19.25	-0.370	0	12
53	6.473	5.900	21.65	-0.116	0.100	40
54	2.953	7.5	5.351	-0.103	1	16
55	7.166	5.733	26.39	0.229	0	30
56	8.798	5.5	40.69	1.018	0	8
57	6.209	6	20.56	0.212	0	12
58	-5.449	7	15.43	0.143	0	49
59	-6.777	7	27.64	-0.078	0	10
60	-4.242	7.095	10.40	-0.061	0	21
61	-6.663	7	23.13	0.172	0	33
62	-4.240	7.308	9.815	-0.089	0	26
63	-6.165	7	22.75	0.270	0	6
64	-5.886	7	18.24	0.029	0	27
65	-7.053	7	27.87	-0.045	0	8
66	1.281	7.886	21.80	-0.161	0	35
67	-0.199	8	18.93	-0.154	0	51
68	-2.085	8	18.74	0.035	0	12
69	2.581	8	16.87	0.600	0	5
70	-1.851	8	29.46	-0.195	0	11
71	-0.641	8.240	9.457	-0.056	0	25
72	3.267	7.200	29.81	-0.552	0	5
73	0.435	8.034	14.77	-0.156	0	33
74	0.349	9	0.489	0.245	0.327	49
75	-1.718	8.813	3.073	0.159	0	16
76	-1.210	8.806	1.469	0.138	0	31
77	1.737	9	4.252	-0.506	0.250	4
78	1.471	8.739	2.217	-0.432	0.609	23
79	0.339	8.692	1.546	-0.009	0.115	26

注：本表列出了用于为 X 数据集着色的各个 5 Y变量的平均值。底层数据集包含 900 个观测值，涉及 2 个变量： X_1 和 X_2。数据经过平移，形成 9 组、每组 100 个点，中心分别位于 (-6,6)、(-3,3)、(3,3)、(6,6)、(0,0)、(-3,-3)、(3,-3)、(-6,-6) 和 (6,-6)。着色定义为 $Y_1 = X_1 + X_2 + \theta$，其中 $\theta \sim N(0,\ 0.2)$， Y_2 为组号；$Y_3 = X_1^2 + X_2^2 + \theta$，其中同样有 $\theta \sim N(0,\ 0.2)$； $Y_4 = \phi$，其中 $\phi \sim N(0,\ 1)$；而 Y_5 在 $0 < X_1 < 3$ 和 $0 < X_2 < 3$ 均满足时取值为 1。

表 A3： Y_1 摘要 X 数据集中的示例表格行（球 1-40）

Ball	Mean	标准差	Min	q25	q50	q75	Max	Size
1	0.656	0.770	-0.696	0.030	0.542	1.228	2.699	44
2	0.136	0.775	-1.652	-0.578	0.158	0.785	1.732	40
3	-2.101	0.752	-3.379	-2.608	-2.279	-1.431	-0.924	11
4	-2.329	0.593	-2.756	-2.756	-2.579	-1.652	-1.652	3
5	1.110	0.520	0.344	0.918	1.053	1.376	2.456	15
6	2.524	0.742	1.400	2.144	2.533	2.704	3.912	9
7	-0.614	0.843	-2.643	-0.701	-0.526	0.001	0.721	29
8	-1.702	0.601	-2.466	-2.019	-1.591	-1.541	-0.411	9
9	-1.027	0.131	-1.120	-1.120	-1.027	-0.935	-0.935	2
10	0.644	0.843	-0.664	0.003	0.530	1.203	2.324	23
11	11.28	0.725	9.513	10.88	11.36	11.71	12.74	24
12	9.190	1.026	7.364	8.253	9.271	9.930	10.55	22
13	12.91	0.746	11.73	12.29	12.86	13.33	14.44	38
14	14.51	0.553	13.96	14.05	14.44	14.78	15.32	5
15	13.66	0.542	12.74	13.27	13.92	14.13	14.20	11
16	11.75	0.819	10.36	11.03	11.66	12.44	13.25	38
17	11.81	0.779	10.44	11.27	11.70	12.44	13.37	50
18	10.05	1.101	8.038	9.210	9.969	10.94	11.63	28
19	-12.17	0.685	-13.74	-12.83	-12.19	-11.77	-10.67	39
20	-13.01	0.711	-14.49	-13.63	-12.98	-12.45	-11.63	30
21	-10.86	0.894	-12.19	-11.66	-10.84	-10.35	-9.06	26
22	-14.03	0.568	-15.03	-14.36	-13.80	-13.64	-13.29	12
23	-10.83	0.442	-11.40	-11.15	-10.77	-10.44	-10.39	5
24	-10.48	0.637	-11.32	-10.91	-10.54	-10.13	-9.06	12
25	-15.65	1.194	-16.50	-16.50	-15.65	-14.81	-14.81	2
26	-12.82	0.818	-15.02	-13.42	-12.78	-12.17	-11.62	46
27	-14.02	0.389	-14.29	-14.29	-14.02	-13.74	-13.74	2
28	-7.179	0.716	-9.288	-7.667	-7.054	-6.565	-6.245	25
29	-0.535	0.765	-2.123	-1.054	-0.526	0.184	0.982	43
30	1.349	0.799	0.100	0.555	1.618	1.786	3.180	29
31	-1.474	0.889	-3.111	-2.180	-1.558	-0.648	-0.074	25
32	-2.753	0.796	-3.315	-3.315	-2.753	-2.190	-2.190	2
33	-1.862	0.653	-2.881	-2.495	-1.793	-1.257	-1.054	12
34	0.008	0.938	-1.517	-0.902	0.194	0.867	1.426	28
35	0.006	0.837	-1.757	-0.604	-0.074	0.543	1.717	27
36	1.082	0.724	-0.148	0.510	1.275	1.640	2.163	13
37	1.761	0.751	0.896	1.108	1.483	2.261	3.102	9
38	0.130	0.788	-1.168	-0.463	0.192	0.871	1.774	31
39	1.726	0.755	0.705	1.135	1.590	2.487	2.678	8
40	-0.151	0.414	-0.604	-0.358	-0.156	-0.148	0.510	5

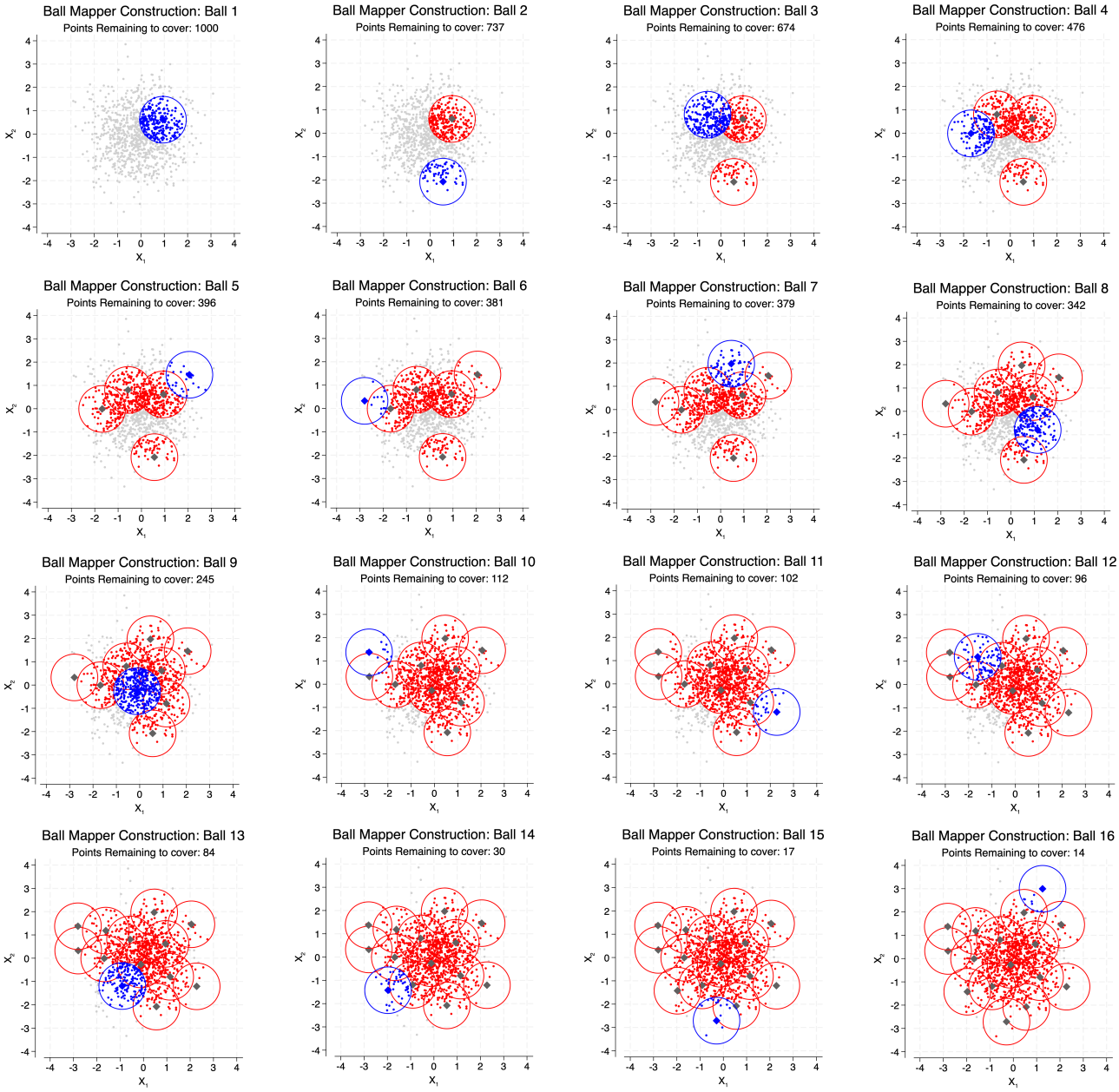
注：数值给出了 TDABM 图中每个球对应的 Y_1 的均值、标准差、最小值、最大值、四分位数和中位数。 $Y_1 = X_1 + X_2 + \theta$，其中 $\theta \sim N(0,\ 0。2)$ 是一个随机噪声分量。底层数据集包含 2 个变量 X_1 和 X_2 的 900 个观测值。数据经过转换，形成 9 组各含 100 个点的组，中心分别位于 (-6,6)、(-3,3)、(3,3)、(6,6)、(0,0)、(-3,-3)、(3,-3)、(-6,-6) 和 (6,-6)。

表 A4： Y_1 摘要 X 数据集中的示例表格行（球 41–79）

Ball	Mean	标准差	Min	q25	q50	q75	Max	Size
41	-0.498	0.796	-2.173	-0.987	-0.511	0.051	0.794	50
42	-1.092	0.701	-1.998	-1.664	-1.299	-0.712	0.231	17
43	0.774	0.726	-0.539	0.354	0.743	1.203	2.324	31
44	2.930	0.848	1.959	2.288	2.693	3.665	4.191	8
45	0.415	0.784	-0.967	-0.140	0.408	1.034	2.153	33
46	-0.968	0.772	-2.466	-1.539	-0.968	-0.411	0.779	26
47	-2.397	0.702	-3.371	-3.077	-2.173	-1.851	-1.503	9
48	-2.626	0.754	-3.479	-3.333	-2.492	-1.998	-1.829	5
49	2.132	0.558	1.556	1.799	1.988	2.324	2.996	5
50	5.161	0.904	3.053	4.759	5.026	5.877	6.669	16
51	5.026	0.666	3.580	4.500	5.096	5.525	6.212	23
52	5.781	0.728	4.831	5.357	5.586	6.236	7.178	12
53	6.473	0.816	5.096	5.754	6.661	6.990	8.295	40
54	2.953	1.097	1.378	2.218	2.601	3.746	4.964	16
55	7.166	0.716	5.689	6.853	7.028	7.522	8.861	30
56	8.798	0.702	7.826	8.367	8.579	9.358	9.950	8
57	6.209	0.763	4.865	5.850	6.139	6.796	7.507	12
58	-5.449	0.872	-7.001	-6.229	-5.347	-4.836	-3.770	49
59	-6.777	0.785	-7.774	-7.337	-6.783	-6.460	-5.292	10
60	-4.242	0.778	-5.317	-4.866	-4.148	-3.598	-2.635	21
61	-6.663	0.639	-7.899	-7.027	-6.638	-6.343	-5.203	33
62	-4.240	0.859	-5.375	-4.866	-4.418	-3.740	-2.325	26
63	-6.165	0.491	-6.687	-6.487	-6.268	-5.971	-5.310	6
64	-5.886	0.806	-7.615	-6.522	-5.844	-5.299	-4.523	27
65	-7.053	0.569	-7.899	-7.526	-6.945	-6.587	-6.409	8
66	1.281	0.807	-0.246	0.649	1.184	1.770	3.158	35
67	-0.199	0.802	-1.784	-0.676	-0.245	0.416	1.770	51
68	-2.085	0.518	-2.708	-2.504	-2.171	-1.729	-1.058	12
69	2.581	0.415	2.159	2.301	2.449	2.812	3.182	5
70	-1.851	0.561	-2.708	-2.324	-1.784	-1.246	-1.128	11
71	-0.641	0.759	-2.108	-0.963	-0.418	-0.144	1.002	25
72	3.267	0.684	2.261	3.102	3.158	3.904	3.910	5
73	0.435	0.855	-1.027	-0.245	0.379	1.091	2.449	33
74	0.349	0.680	-0.945	-0.060	0.346	0.815	1.861	49
75	-1.718	1.066	-3.738	-2.409	-1.761	-0.697	-0.251	16
76	-1.210	0.886	-3.132	-1.866	-1.135	-0.528	0.228	31
77	1.737	0.965	0.778	1.069	1.558	2.405	3.053	4
78	1.471	0.736	0.230	0.867	1.266	2.109	2.745	23
79	0.339	0.803	-1.349	-0.060	0.495	1.081	1.620	26

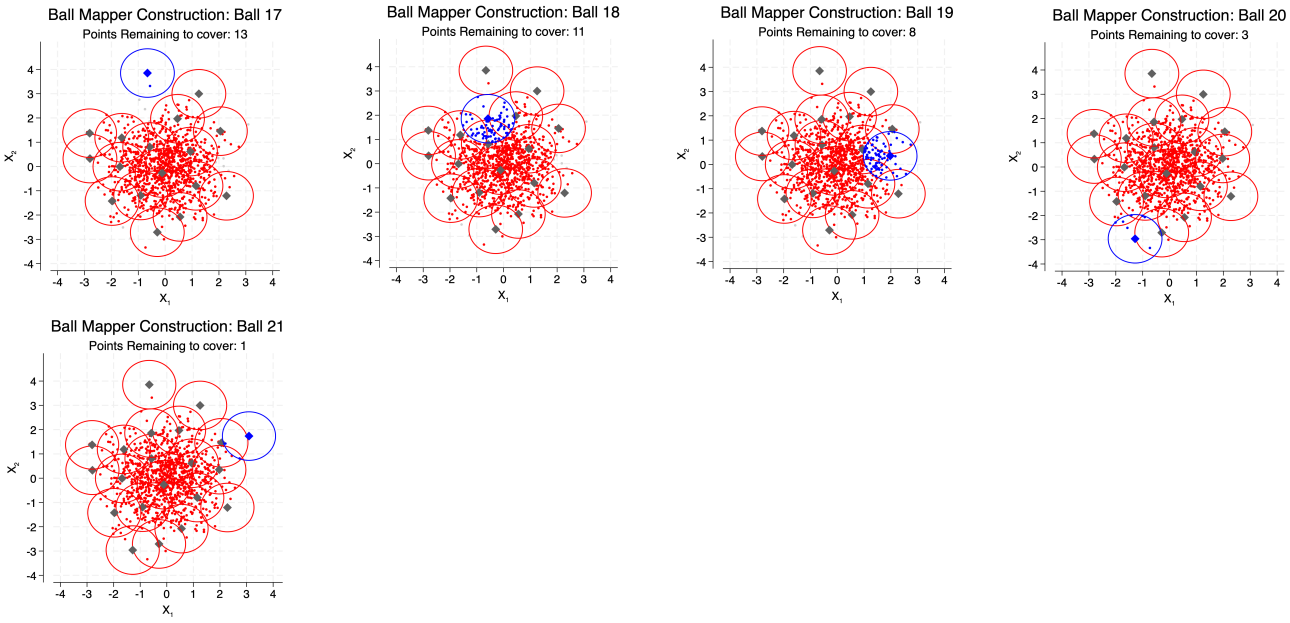
注：这些值给出了 TDABM 图中每个球的 Y_1 均值、标准差、最小值、最大值、四分位数和中位数。 $Y_1 = X_1 + X_2 + \theta$，其中 $\theta \sim N(0,\ 0.2)$ 是一个随机噪声分量。底层数据集包含 900 个观测值和 2 个变量， X_1 和 X_2。数据经过平移，形成 9 组，每组 100 个点，中心分别位于 (-6,6)、(-3,3)、(3,3)、(6,6)、(0,0)、(-3,-3)、(3,-3)、(-6,-6) 和 (6,-6)。

图 A1：Ball Mapper 图的逐步构建过程（球 1 至 16）



注：构建 二元高斯云的 TDABM 覆盖。已覆盖的点标为红色，
当前子图中正在创建的球 标为蓝色。此处 $X_1 \sim N(0,1)$ 和 $X_2 \sim N(0,1)$，球半径为 $\varepsilon = 1$. 00.

图 A2：Ball Mapper 图的逐步构建（球 17 至球 21）

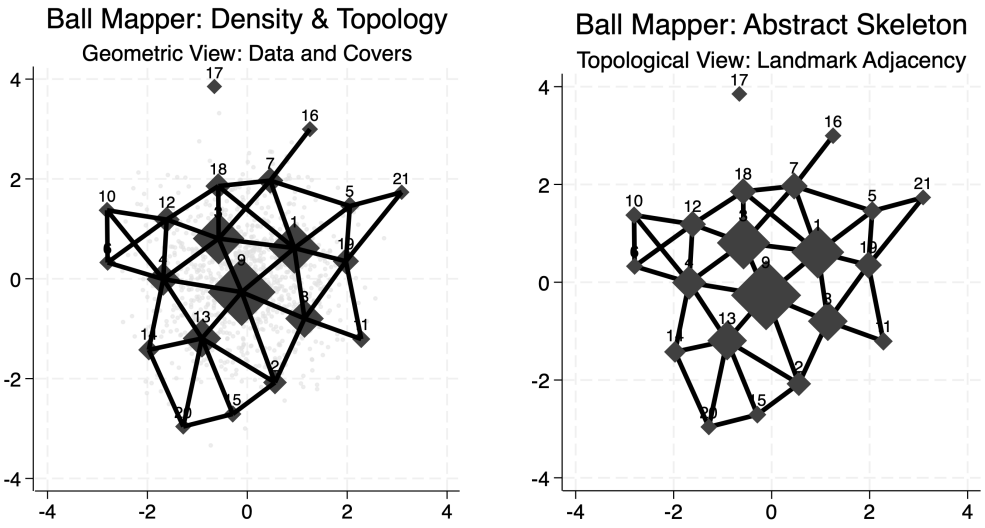


注：二元高斯云的 TDABM 覆盖构建过程。已覆盖的点标为红色，子图中正在创建的特定球标为蓝色。这里 $X_1 \sim N(0,\ 1)$ 和 $X_2 \sim N(0,\ 1)$，球半径为 $\varepsilon = 1.00$。

加入一个球。第二个球以数据集中的第 2 个点为中心绘制。新球覆盖蓝色点，而球 1 覆盖的点则显示为红色。这里前两个球之间没有重叠。球 3 绘制在球 1 附近，可以看到两者有所重叠。接着，球 4 位于球 3 左侧，而球5位于球 1 上方，靠近空间的右上方。球 6 覆盖了图左侧的几个点。球 7 位于图的右上方，与球 1 和球 3 重叠。加入第 8 个球，将球 1 和球 2 连接起来。球 9 位于点云密集的部分。随后，球 10 位于图左侧、球 6 的正上方，也只包含了另外几个点。球 11 位于右下方。球 12、13 和 14 是填补现有球之间空隙的其他示例。随后，球15位于图的底部中央。球 15 位于图的下方，并覆盖了其中的一小组点。球 16 和球 17 位于图的上边缘，而球 18 则用于填补空间较密集部分的另一处空隙。球 19 位于右侧中央，同样覆盖了空间中较密集的部分。球 20 和球 21 位于空间边缘，完成覆盖。

图构建的后续步骤可通过图 A3 轻松理解。首先，通过标志点的大小来体现数据的密度。通过在任何非空重叠处绘制边，体现地标彼此之间的接近程度。为便于理解该图，标志点均已编号。由于我们使用散点图来演示其功能，因此数字标注在对应地标附近。算法会记住球内的点，因为这些信息由算法存储。

图 A3：迈向 TDABM 图



(a) 大小、边和标签　　(b) 移除其他点

注：在二维高斯点云上构建 TDABM 图。这里的数据包含 1000 个观测值，其中 $X_1 \sim N(0,\ 1)$ 和 $X_2 \sim N(0,\ 1)$。在构建 TDABM 图时，所有点都被覆盖后，根据各地标所对应球中的点数确定地标的大小，并在球发生重叠的任意一对地标之间绘制边。由于数据中的信息随后都包含在图中，因此完整的数据点集就变得多余了。子图 (b) 展示了移除其他数据点后的同一张图。最后一步是移除坐标轴，并将地标抽象化。

单个数据点因此变得多余，可以移除。子图 (b) 展示了移除所有数据点后的图。有关结构和密度的信息仍然保留。此处的最后一步是移除坐标轴，将图绘制为抽象形式。网络的抽象不会改变边。最终的图看起来像一个网络，但它是对底层数据的拓扑忠实表示。

在本示例中，底层数据集具有两个维度。二维数据的散点图很容易理解和解读。因此，抽象过程以及从图中移除数据点，反而会使图的信息量降低。然而，当变量更多、无法在单张图中观察数据时，TDABM 过程便能够构建出可解释的可视化结果。