arXiv:2601.05400v1 [stat.AP] 8 Jan 2026

用 h 图表示非对称关系：发现跨期刊引文关系的原型模式

Aleix Alcacer 和 Irene Epifanio

摘要 本研究从一个新颖的角度探讨多维尺度分析问题。我们提出一种基于 h 图、可扩展且天然适用于非对称邻近数据的方法。该方法不对对象本身进行嵌入，而是嵌入用于定义各对象到其他对象或从其他对象到该对象的邻近性的变量。该方法易于实现，所得表示的质量也易于评估。本文通过在一张共同地图上可视化期刊的引用概况与被引概况之间的非对称关系，展示了该方法。根据欧几里得距离衡量，在 h 图中相距较远（或较近）的两个轮廓分别表示它们彼此不同（或相似）。这种表示方式可用于进行原型体分析（ADA）。我们利用 ADA 寻找原型期刊（或极端案例）。通过将数据集表示为这些原型期刊的凸组合，结果便于解读，即使是非专业人士也能轻松理解。本文还将该方法与其他方法论进行比较，结果显示我们提出的方法表现良好。为确保研究可复现，代码和数据均已公开。

1 引言

多维尺度分析是许多领域中的经典问题。然而，这仍然是一个备受关注的问题，因为在生物信息学等某些领域中，特征不可用（Hastie 等人，2009年，第18章），而且只能获得对象对之间的邻近性信息。因此，降维工具对于机器学习领域至关重要。

在图像匹配、文本或网络挖掘、认知心理学等某些领域，成对数据不具度量性，不相似度矩阵也不满足度量函数的数学

Aleix Alcacer 海梅一世大学，西班牙，电子邮件：aalcacer@uji.es Irene Epifanio 海梅一世大学和 ValgrAI，西班牙，电子邮件：epifanio@uji.es

要求（自反性、确定性、对称性、三角不等式）。在这些情况下，非度量性并非源于含噪测量，而是因为数据本质上不具度量性。因此，我们不应强行使其满足度量性（例如，给非对角线上的不相似性加上一个常数），否则可能会丢失真实信息（Pe̩kalska 和 Duin，2005年）。

绝大多数成熟的机器学习方法都只针对度量数据构建，且通常不考虑度量性违背，而是对不相似性进行转换（例如，通过对称化），或像经典尺度分析那样直接省略负特征值。

然而，近年来该领域发表的论文数量激增，表明人们重新开始关注非对称关系的降维问题（Okada 和 Imaizumi，2024年；Olszewski，2024年、2023年；Di Nuzzo 等，2025年；Bove 等，2021年）。Bove 和 Okada（2018年）概述了针对非对称成对关系的不同方法。

Epifanio（2013）提出了一种基于 h 图的方法论，可以考虑非度量性，包括不对称性和缺乏自反性（正的自我不相似性）。其思路是嵌入定义每个实例与其他实例之间邻近性的变量，而不是嵌入实例本身。在 h 图中，不相似性矩阵被视为一个数据矩阵，表示两个变量：元素“j”到所有其他元素的不相似性（$d_{j\cdot}$），以及所有其他元素到“j”的不相似性（$d_{\cdot j}$）。在这种映射中，两个变量之间的欧几里得距离表示变量差值的样本标准差。如果两个变量相似，那么它们的差值——以及该差值的标准差——就会很小。

投影到欧几里得空间后，我们便可以使用许多统计技术，例如原型体分析（ADA）（Vinu´e 等，2015年），正如 Vinu´e 和 Epifanio（2017年）所做的那样。ADA 是一种无监督统计学习技术，用于识别数据集中的代表性“极端”案例（称为原型体）。这些原型体是实际观测值，可用于将所有其他数据点近似表示为它们的凸组合。ADA 依赖于欧几里得距离和凸组合，因此，将数据投影到欧几里得空间（在其中这些运算定义明确且有意义）可以使该方法正常运行。ADA 以便于人类理解的方式概括数据，因为人们更容易理解极端值。

鉴于聚类技术和 ADA 两种方法都旨在揭示数据中的结构，因此聚类技术也可作为 ADA 的替代方法。不过，两者在概念上的重要差异意味着它们是否适用取决于具体情境。聚类通常用集中趋势来表示各组，而 ADA 则通过极端但具有代表性的观测值来刻画模式；当数据变化由边界案例而非中心案例解释得更好时，这有助于理解。数据呈现出明显分离的组时，聚类通常最为有效，因为此时聚类中心能够有效概括各个子组。相比之下，如果群组边界模糊，或者数据构成一个连续体而非彼此独立的簇，那么 ADA 可能会得出更有参考价值的结果，因为其原型体能够捕捉界定变化范围的极端情况。此外，ADA 还会提供比例或混合权重，说明每个观测值与各个原型体的关联程度，从而细致地描述个案如何组合这些极端原型。一项比较

两种技术的教学示例由 Cabero 和 Epifanio（2020年）给出，直观展示了它们各自的表现及解释优势。

本文通过一个关于统计学期刊间相互引文的数据库，逐步展示该方法论。第 2 节介绍 h 图和 ADA 方法论。第 3 节介绍数据集。第 4 节展开其应用。最后，第 5 节给出结论。

为确保结果可复现，代码和数据可在 https://epifanio.uji.es/RESEARCH/asyada.zi 获取。p.

2 方法论

2.1 H 图

Corsten 和 Gabriel（1976）提出了 h 图，用于可视化数据矩阵 $\mathbf{X}$ 中多个变量之间的关系。要在二维空间中计算 h 图，需对矩阵 $\mathbf{S}$（$\mathbf{X}$ 的方差–协方差矩阵）进行特征值分解。$\mathbf{S}$ 始终为半正定矩阵，因此其特征值始终为正。令 λ_1 和 λ_2 为最大的两个特征值，$\mathbf{q}_1$ 和$\mathbf{q}_2$ 为它们对应的单位特征向量。二维 h 图为：

$$\mathbf{H}_2 = \left(\sqrt{\lambda_1}\mathbf{q}_1,\ \sqrt{\lambda_2}\mathbf{q}_2\right).$$

行$\mathbf{h}_i$ 与$\mathbf{h}_j$ 之间的欧几里得距离，约等于变量“i”与“j”之差的样本标准差（Seber，1984）。对于完整矩阵$\mathbf{H}$，二者则相等。因此，当最大的两个特征值占总方差的大部分时，矩阵$\mathbf{H}_2$ 可作为有效的低维表示。Corsten 和 Gabriel（1976）提出了以下拟合优度度量，用于评估二维 H 图绘制；其值越接近 1，表示拟合效果越好：

$$\frac{\lambda_1^2 + \lambda_2^2}{\sum_j \lambda_j^2}$$

设 $\mathbf{\Delta}$为一个不相似度矩阵，其元素为 δ_{ij}，表示对象“i”与对象“j”之间的不相似度。当 $\mathbf{\Delta}$ 不对称时，考虑$\mathbf{D} = [\mathbf{\Delta} | \mathbf{\Delta}']$，它是由 $\mathbf{\Delta}$和 $\mathbf{\Delta}'$按列组合而成的矩阵，其中$'$表示转置。若将$\mathbf{D}$ 视为数据矩阵，则前 n 列包含 $d_{\cdot j}$类变量，即其他对象与“j”之间的不相似度；而后面的列块包含 $d_{j\cdot}$类变量，即“j”与其他对象之间的不相似度。应用 H 图绘制后，$\mathbf{H}_2$ 有 $2n$ 行。前 n 行近似表示 $d_{\cdot j}$轮廓或描述，而从第 $n+1$行到第 $2n$行则近似表示 $d_{j\cdot}$轮廓。

对于多维尺度分析，H 图绘制应用于$\mathbf{D}$ ，这意味着如果 $d_{\cdot j}$ 和$d_{\cdot i}$ 轮廓相似，它们在图中的位置就会彼此接近；同理也适用于

$d_{j\cdot}$和 $d_{i\cdot}$轮廓。此外，$d_{\cdot j}$和 $d_{i\cdot}$轮廓会映射到统一表示中，从而能够直接比较。当 $d_{\cdot j}$ 轮廓的位置接近 $d_{i\cdot}$ 轮廓时，表明两者相似。因此，通过计算表示中 $d_{\cdot j}$ 和 $d_{j\cdot}$轮廓之间的欧几里得距离，我们可以识别非对称程度较高或较低的对象——也就是说，分别找出其 $d_{\cdot j}$和 $d_{j\cdot}$轮廓彼此高度吻合或差异显著的对象。

总之，h 图的目标不是精确保留点间不相似性，也不是维持这些不相似性的排序。相反，它着重于维持不相似性变量之间的关系，即这些轮廓本身。正如 Epifanio（2013）的示例 1 所示，存在这样一种情况：i 与 j 之间的不相似性很小（MA 与 VL 之间的距离最小），但这一点并未反映在 h 图投影中，因为 MA 与 VL 之间的距离仅为第二小。也就是说，不相似性的排序并未得到保留。

还需注意，尽管显然可以分别表示变量 $d_{i\cdot}$和 $d_{\cdot j}$，但这样做会因未将二者结合考虑而导致信息丢失。尤其是，我们将无法确定某一期刊的被引概况是否与另一期期刊的引用概况相似。同样，也无法分析某一期刊的不对称性，因为同一期刊的被引概况和引用概况不会表示在一个公共空间中。还应牢记，在分别对被引概况和引用概况进行 h 图投影时，各坐标轴不一定具有相同的解释，因此不同空间无法进行有意义的合并。

2.1.1 h 图的优点

除了 h 图结果易于解读之外，由于所有轮廓都映射在同一表示中，h 图还具有若干优点：(a) h 图基于特征向量提供明确的解，从而避免了其他方法中常见的局部极小值问题。此外，即使对象和个体的角色互换，所得解仍保持一致。(b) h 图在不相似性尺度的线性变换下具有不变性，这意味着即使尺度发生变化，视觉构型仍保持不变（Epifanio，2013）。(c) h 图的计算效率高，能够处理非常大的矩阵（Schwarz 和 Karlsson，2019），即具有良好的可扩展性。(d) h 图的拟合优度易于评估；(e) 如果某些观测值是离群值，可以按照 Daigle 和 Rivest（1992）的建议计算稳健 h 图。

2.2 原型体分析（ADA）

原型体分析基于这样一种理念：数据可以表示为属于数据集的一组原型模式的混合（有关该领域的综述，请参见 Alcacer 等（2025b））。统计学中的原型概况与日常生活中的直观含义相同——它们代表极端或纯粹的类型，并且比

中心点更容易理解，因为人们往往更容易理解相互对比或对立的组成部分（Thurau 等人，2012）。

让我们回顾多变量情形下的 ADA。设 $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ 为一个包含 n 个观测值和 m 个变量的 $n \times m$ 数据矩阵。

在 ADA 中，通过代表性观测值的凸组合定义三个矩阵，以近似数据结构：

1. k原型体记为 $\mathbf{z}_j$，是矩阵 $\mathbf{Z}$ 的行，其中每个 $\mathbf{z}_j$对应于原始数据矩阵 $\mathbf{X}$ 中一个特定的观测数据点。2. 一个 $n \times k$矩阵 $\boldsymbol{\alpha} = (\alpha_{ij})$，其中包含用于将每个观测值 $\mathbf{x}_i$ 近似表示为原型体凸组合的混合系数：

$$\hat{\mathbf{x}}_i = \sum_{j=1}^{k} \alpha_{ij} \mathbf{z}_j.$$

3. 一个 $k \times n$ 矩阵 $\boldsymbol{\beta} = (\beta_{jl})$，由用于确定从数据中选择原型体的二元系数组成。每个原型体 $\mathbf{z}_j$ 定义为：

$$\mathbf{z}_j = \sum_{l=1}^{n} \beta_{jl} \mathbf{x}_l,$$

其中，$\beta_{jl} \in \{0, 1\}$确保每个原型体都是 $\mathbf{X}$ 中的一个特定观测值。

为估计这些矩阵，目标是使残差平方和（RSS）最小：

$$RSS = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \mathbf{z}_j\|^2 = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} \mathbf{x}_l\|^2, \tag{1}$$

其中，$|\cdot|$表示欧几里得范数。

该优化问题受以下约束条件限制：

$_k$1. 混合系数的凸性约束：$\sum \alpha_{ij} = 1$，其中 $\alpha_{ij} \geq 0$，对于$_{j=1}$

$i = 1, \ldots, n$

$_n$2. 原型体的二元选择约束：$\sum \beta_{jl} = 1$，其中 $\beta_{jl} \in \{0, 1\}$，且$_{l=1}$

$j = 1, \ldots, k.$

这些约束确保每个观测值都由实际数据点的凸组合近似，从而提高可解释性和经验有效性。

ADA 背后的混合整数优化问题采用 Vinu´e 等人（2015年）提出的两阶段算法求解。第一阶段称为 BUILD，通过选取一组候选原型体来初始化解。第二阶段称为 SWAP，会迭代优化这一选择：只要将初始集合中的元素替换为未选中的观测值能够降低

残差平方和（RSS），就进行替换。我们的实现采用 Epifanio 等人（2018年）开发的 R 实现（R 开发核心团队，2025年）。Vinu´e 和 Epifanio（2021年）通过抽样提高了原型体算法的可扩展性。此外，Vinu´e 和 Epifanio（2021年）还对该算法进行了并行化，从而节省更多时间。

为确定适当的原型体数量，我们采用肘部准则。这是一种先前已应用于 Cutler 和 Breiman（1994年）、Eugster 和 Leisch（2009年）以及 Vinu´e 等（2015年）论文中的方法。该方法包括绘制 RSS 与原型体数量之间的关系，并确定 RSS 下降速率发生急剧变化的点——通常称为肘部。

2.3 H 图 +ADA

与 Vinu´e 和 Epifanio（2017年）一样，为了通过 H 图绘制从非对称关系中获得原型体，我们注意到每个对象“j”都有两个轮廓：一个表示 $d_{j\cdot}$，另一个近似表示 $d_{\cdot j}$。二者均通过二维向量在同一构型中表示，即 $\mathbf{h}_j$ 和 $\mathbf{h}_{j+n}$，其中 $j = 1$、...、 n。因此，我们对一个由 H 图轮廓两个区块的表示组合而成的 $n \times 4$ 矩阵 $\mathbf{X}$ 应用 ADA；也就是说，$\mathbf{X}$的第 j 行由 $\mathbf{h}_j$ 和 $\mathbf{h}_{j+n}$ 构成。

ADA 已在多项先前研究中结合 H 图投影使用，并取得了良好效果。Vinu´e 和 Epifanio（2017）在足球队表现分析中将其用于非对称不相似度；Epifanio 等人（2023年）将其应用于足部形状研究，Alcacer 等人（2021年）将其用于鞋码推荐系统，Epifanio 等人（2020年）则将其应用于多个含有缺失数据的数据集。此外，Vinu´e 等人（2015年）在服装设计研究中，将其用于西班牙女性躯干三维二值图像的非度量但对称的不相似度。此外，Vinu´e 等人（2015年）还针对非度量但对称的不相似度开展了一项模拟研究，以确定哪种多维尺度分析（MDS）方法更适合在多维尺度投影后还原原型体。研究将 H 图与其他六种 MDS 方法进行了比较：经典（度量）多维尺度分析、Kruskal 非度量多维尺度分析、Sammon 非线性映射、等距映射（Tenenbaum 等人，2000年）、局部线性嵌入（Roweis 和 Saul，2000年）以及扩散映射（Nadler 等人，2006年）。结果表明，与其他 MDS 方法相比，H 图在还原原始原型体方面表现更佳。

在可扩展性方面，由于 h 图和 ADA 都具有可扩展性，因此二者的顺序组合也具有可扩展性。

下面详细说明它们的计算复杂度。对于 h 图，首先我们计算协方差矩阵，这主要涉及矩阵乘法，即 O(n^3)，然后计算特征向量。计算特征向量的复杂度取决于所采用的方法和矩阵。尽管存在一些专门的算法，即使使用 $O(1)$ 也可以运行（Sun 等，2020），但在最坏情况下，使用标准算法可以在多项式时间 $O(n^3)$ 内完成计算。至于 ADA，其计算复杂度在很大程度上取决于用于估计 $\boldsymbol{\alpha}$ 的方法。如果使用非负最小二乘（NNLS）（Lawson 和 Hanson，1974；Cutler 和 Breiman，1994），其复杂度为 $O(nk^3)$（Alcacer 等，2025b）。因此，总体计算复杂度是以下两项的乘积：

枚举所有可能的原型组合的成本与矩阵$\boldsymbol{\alpha}$更新过程的成本，即 $O(\binom{n}{k} nk^3)$。

3 数据库

Epifanio（2014）使用 h 图分析了 2011年 JCR版本（Clarivate 发布的《期刊引证报告》）中“统计学与概率论”学科类别的 25 种期刊之间的非对称引用关系。本文考察的是 2025年 JCR版本中同一学科类别按影响因子排名前 25 的期刊，该版本提供 2024年数据。表 1 按字母顺序列出了这些期刊的名称及其编号。如前所述，h 图 + ADA 具有可扩展性，因此可用于大型数据库，不受限制。不过，出于教学和演示目的，我们选择使用仅包含 25 个观测值的小型数据库，以便更简洁地讨论结果。其他近期研究也采用了这一做法，例如 Okada 和 Imaizumi（2024）的研究。

表 1 所选期刊列表

#	期刊名称
1	ANNALS OF PROBABILITY
2	《统计年刊》
3	《统计学及其应用年度评论》
4	《贝叶斯分析》
5	《生物统计学》
6	《化学计量学与智能实验室系统》
7	《计量经济学》
8	ECONOMETRICS AND STATISTICS
9	计量经济学期刊
10	FUZZY SETS AND SYSTEMS
11	IEEE-ACM《计算生物学与生物信息学汇刊》
	数学
12	JOURNAL OF STATISTICAL SOFTWARE
13	JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
14	英国皇家统计学会 B 系列统计学期刊
	统计方法学
15	法律, 概率与风险
16	MULTIVARIATE BEHAVIORAL RESEARCH
17	概率工程力学
18	质量技术与定量管理
19	空间统计学
20	《统计分析与数据挖掘》
21	STATISTICAL SCIENCE
22	统计调查
23	STOCHASTIC ENVIRONMENTAL RESEARCH AND RISK ASSESSMENT
24	TECHNOMETRICS
25	WILEY INTERDISCIPLINARY REVIEWS-COMPUTATIONAL STATISTICS

跨期刊引用显然具有非对称性，因为期刊“i”引用期刊“j”的程度未必与期刊“j”引用期刊“i”的程度相同。同样，三角不等式未必成立，而且期刊会引用自身（缺乏自反性）。

我们不采用原始的期刊间引用频次，而是采用经相关性因子标准化后的频次（Pudovkin 和 Garfield，2002年），以调整期刊规模差异（Klavans 和 Boyack，2006年）。Pudovkin 和 Garfield（2002年）将期刊 i 与期刊 j 的期刊相关性记为 $R_{i\to j}$，定义如下：

$$R_{i\to j} = \frac{H_{i\to j}\cdot 10^6}{\mathrm{Pap}_j \cdot \mathrm{Ref}_i}$$

其中，$H_{i\to j}$ 表示当年期刊 i 引用期刊 j 的次数（即引用所有年份发表在 j 上的论文的次数），Pap_j 表示当年期刊 j 发表的论文数，Ref_i 表示当年期刊 i 引用的参考文献数。

由于数据可用性有限，我们将分析范围限定为2022年和2023年发表的可引用文章。因此，我们只考虑2024年对这些文章的引用。

鉴于我们的目标是进行模式检测，调整构型的尺度（扩大或缩小）或许能提供更有意义的见解（Seber，1984年）。因此，我们不使用相关性测度值，而是使用其秩（相关性值最高者的秩最小），这与 Epifanio（2013年，2014年）的方法一致。这种相对化方式在植物生物学中也很常见，非对称测度尤其适用于分析群落数据（Podani 和 Mikl´os，2002年）。对于并列值，我们赋予其并列项的平均秩。对于因没有引用而未定义的条目，我们赋予其已取得的最大秩加一（本研究中为171）。如需相似性，可将不相似性用171减去其数值转换为相似性。

4 结果

图 1 展示了这些期刊的 h-图映射。二维 h 图的拟合优度为 74.3%，表明其呈现效果良好。被引期刊的位置以黑色标示，施引位置则以红色圆圈标示。

我们先来考察施引档案。左上方是一组专注于一般理论和方法论的期刊（2、5、13、14、20）。左侧中央是一组综述期刊（3、21、25）。附近还有两种期刊，分别是 12 和 22，它们都是开放获取期刊，一种关注调查，另一种关注软件。中央部分是处于理论与应用之间的期刊（4、8、19、24）。右上角是一组社会科学和计量经济学期刊（7、9、16）。最后，右侧中央是特定领域的期刊，例如专注于特定领域应用的期刊（6、11、17、18、23）、概率论（1、15）和模糊集（10）。

然而，被引概况并未如此分组。尽管存在一些组，例如左上方的（2、5、13、14、21），它们对应于一般理论和方法论以及

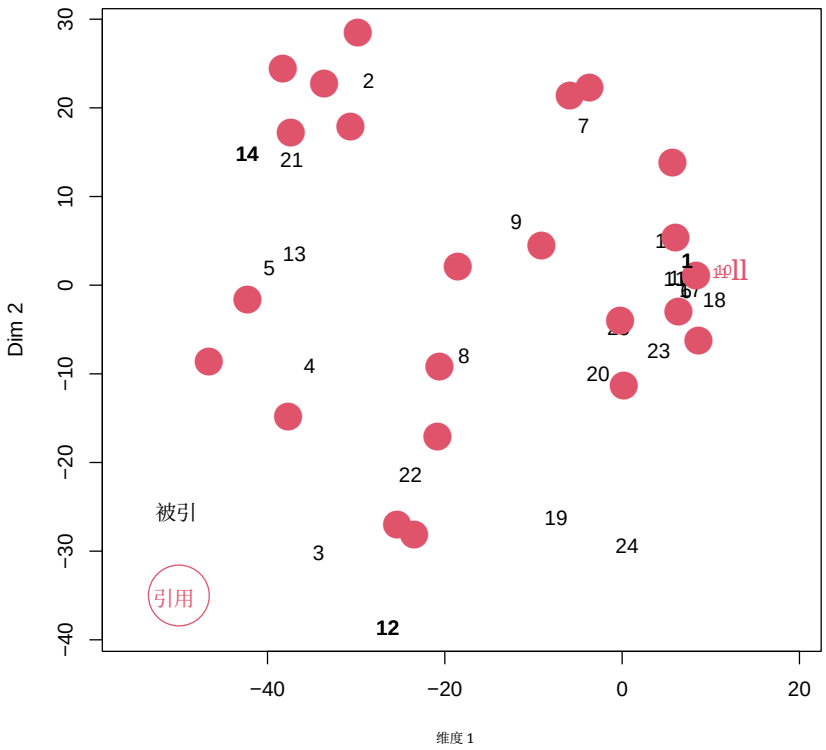

图 1 25 种统计学期刊的 h-图映射。表 1 提供了数字对应的代码。粗体表示 $k = 3$ 的原型体。

综述期刊；还有另一组期刊，主要对应特定领域的期刊（1、6、10、11、15、16、17、18、23），以及《统计分析与数据挖掘》（20）——这是一份介于理论与应用之间、具有计算统计学特征的期刊；此外还有 WILEY INTERDISCIPLINARY REVIEWS-COMPUTATIONAL STATISTICS（25），这是一份专门评述计算统计学领域的综述期刊。

请注意，这两种期刊（20 和 25），以及《统计学及其应用年度评论》（3）——该期刊不仅综述方法论进展，也介绍使这些方法得以实施的计算工具——是最不对称的期刊；也就是说，它们主要引用某些期刊，却被其他类型的期刊引用。例如，《统计分析与数据挖掘》（20）主要引用期刊 13、14 和 5，但其引用者主要来自统计学类别之外。

另一方面，最对称的期刊，即引用和被引用方式相似的期刊，包括 LAW, PROBABILITY & RISK（15）、FUZZY SETS AND SYSTEMS（10）和 IEEE-ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS（11）。它们属于非常特定的领域。

4.1 ADA

ADA 已应用于 h-图表示，以发现模式并概括数据。我们选择使用 ADA，而非聚类，因为图1显示，即使仅凭目视检查，数据也没有呈现出明显分离的组；因此，在此情境下，ADA 是一种信息量更大的方法。此外，我们计算了平均轮廓系数——这是一项衡量整体聚类质量的指标——并观察到其最高值 0.46 出现在 $k = 3$。由于低于 0.5 的值表明聚类结构较弱（Kaufman 和 Rousseeuw, 1990），这一发现印证了图1所传达的直观印象。

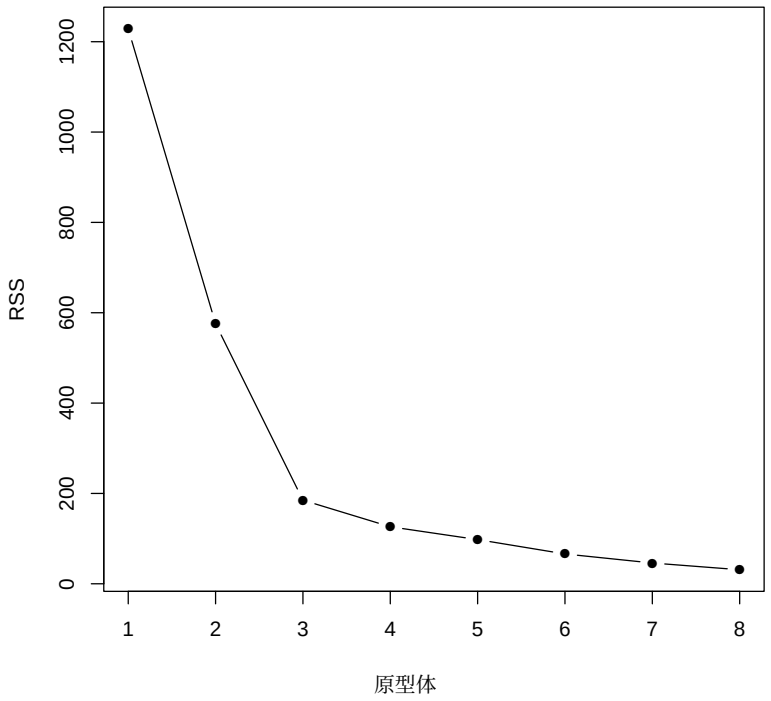

图 2 25 种统计学期刊 h 图的 ADA 碎石图。

图2显示了碎石图。在 $k = 3$ 处出现肘部。图3展示了 $k = 3$ 的 α 值，其中原型体为：ANNALS OF PROBABILITY（1）、JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES B-STATISTICAL METHOD- OLOGY（14）和 JOURNAL OF STATISTICAL SOFTWARE（12），分别代表一份特定期刊、一份理论与方法期刊，以及一份综合性应用计算统计期刊。根据这一表示，例如，ECONOMET- RICS AND STATISTICS（8）可描述为由 47% 的原型体 1、26% 的原型体 2 和 27% 的原型体 3 混合而成。这反映了该期刊的范围：它刊载计量经济学、统计方法学和计算方面的论文。类似地，ANNUAL REVIEW OF STATISTICS AND ITS APPLICATION（3）可近似表示为由 37% 的原型体 2 和 63% 的原型体 3 组成。这也符合该期刊的范围：它综述方法学进展和计算工具。

尽管我们重点关注 $k = 3$ 结果，但根据所需的细节程度，我们可以考虑更多或更少的原型体。对于 $k = 1$，原型体是中心对象，对应于 ECONOMETRICS AND STATISTICS (8)，即一份综合性期刊。对于 $k = 2$，原型体分别是 STOCHASTIC ENVIRONMENTAL RESEARCH AND RISK ASSESSMENT (23) 和 JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION (13)，前者是应用型期刊，后者是方法论期刊。尽管原型体不一定是嵌套的，但在这个数据集中确实如此。对于 k= 4，原型体 JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES B-STATISTICAL METHODOLOGY (14) 和 JOURNAL OF STATISTICAL SOFT- WARE(12) 与使用 $k = 3$ 得出的结果一致；此外，特定期刊 LAW, PROBABILITY & RISK (15) 也成为原型体，而一份综述期刊 WILEY INTERDISCIPLINARY REVIEWS-COMPUTATIONAL STATISTICS (25) 也成为原型体。对于 $k = 5$，原型体 JOURNAL OF THE ROYAL STATISTI- CAL SOCIETY SERIES B-STATISTICAL METHODOLOGY (14)、JOURNAL OF STATISTICAL SOFTWARE (12) 和 LAW, PROBABILITY & RISK 再次保留；综述期刊则由 STATISTICAL SCIENCE (21) 延续；此外还出现了新的原型体类型 TECHNOMETRICS (24)，这是一份应用统计期刊，但属于工程领域。

4.2 与其他方法的比较

我们考察了两种成熟的方法，它们均有相应的 R 软件包，且此前已用于分析期刊之间的交叉引文活动：(a) 多维展开分析，例如 Schneider（2009年）和 Epifanio（2014年）所采用的方法；(b) 网络表示，例如 Calero Medina 和 van Leeuwen（2012年）以及 Epifanio（2014年）所采用的方法。

4.2.1 多维展开分析

在多维展开法中，数据通常表示为两组元素之间的不相似性，这两组元素由 n_1 个体和 n_2 对象组成。其目标

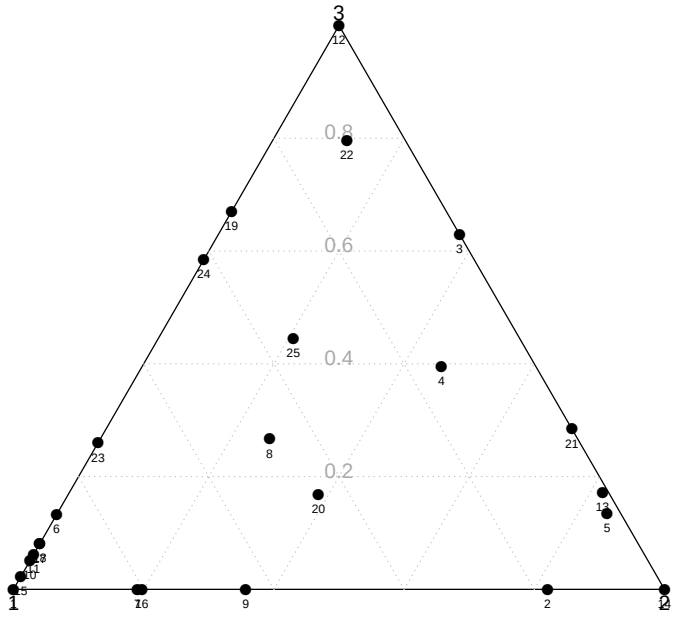

图3 25 种统计学期刊 h 图的 ADA 三元图。

是确定一个共同的定量尺度，以便直观考察这两组之间的关系。多维展开法可被视为多维尺度分析的一个特例。

设 $\mathbf{\Delta}$ 表示维度为 $n_1 \times n_2$ 的观测差异矩阵，其元素 δ_{ij} 表示个体 i 与对象 j 之间观测到的差异。目标是通过最小化以下应力函数，估计维度为 $(n_1 + n_2) \times p$ 的联合构型矩阵 $\mathbf{X}$：

$$\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left(\delta_{ij} - d_{ij}(\mathbf{X}_1, \mathbf{X}_2)\right)^2,$$

其中，距离函数定义为

$$d_{ij}(\mathbf{X}_1, \mathbf{X}_2) = \left(\sum_{s=1}^{p} (x_{1is} - x_{2js})^2\right)^{1/2}.$$

矩阵 $\mathbf{X}$ 被划分为两个子矩阵：$\mathbf{X}_1$，维度为 $n_1 \times p$，表示个体构型；以及 $\mathbf{X}_2$，维度为 $n_2 \times p$，表示对象构型。在当前应用中，这两个集合相同，因此 $n_1 = n_2 = n = 25$，且 $p = 2$。

多维展开分析的结果如图 4 所示。该分析使用 R 包 SMACOF 计算得出（De Leeuw 和 Mair，2009年；Mair 等，2022年）。左侧展示的是最佳解，其中对象和个体分别由被引轮廓和施引档案表示。右侧则交换了这两种角色。两种情况下的构型相似，但并不完全相同。如果达到全局最小值，无论角色如何分配，解都将保持不变；但在低维解中，局部极小值往往更频繁地出现（Borg 和 Mair，2017年），这就是同时展示两种角色的原因。

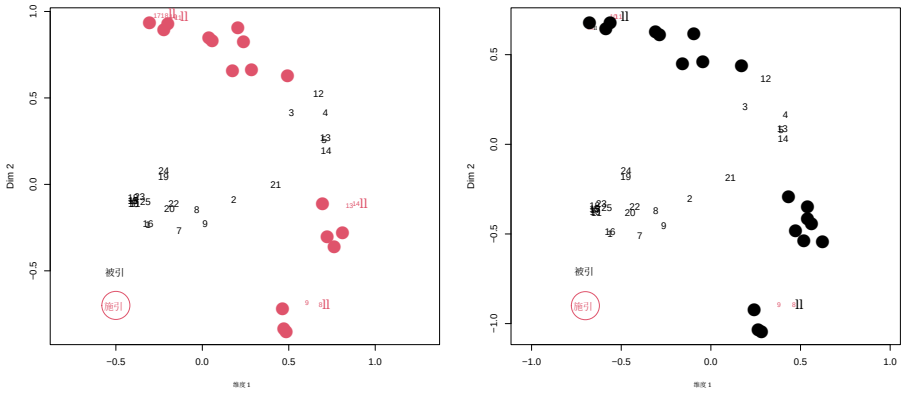

图 4 25 种统计学期刊的展开法解。表 1 列出了各编号对应的代码。

被引位置位于图的中央区域，而施引位置则分成两组。一方面，右下方的施引组又由两组构成：一组是社会科学（16）和计量经济学期刊（7、8 和 9）以及概率论期刊（1）；另一组是关于一般理论和方法论的期刊（2、5、13、14），以及一些应用领域的期刊（20），还有一份综述期刊（3）。图 4 左侧的期刊（4）也属于这一组。另一方面，上方的组分为两组：一组分布较广，由综述期刊（21、22、25）和软件期刊（12），以及介于理论与应用之间的期刊（19、24）构成；另一组则较为紧密，由特定领域的期刊（6、10、11、15、17、18、23）构成。这些施引概况与其对应的被引概况相距甚远，尽管二者相似。例如，LAW, PROBABILITY & RISK（15）是被引次数最多且（自）引次数最多的期刊，FUZZY SETS AND SYSTEMS（10）也有同样的情况，其自引率达到 20% 或以上。展开映射中还发现了更多不一致之处：ANNALS OF PROBABILITY（1）和 MULTIVARIATE BEHAVIORAL RESEARCH（16）这两份期刊彼此非常接近，而这两份期刊

其实差异很大。尽管《统计学及其应用年度评论》（3）的引用概况与其他综述期刊（如 STATISTICAL SCIENCE（21））相似，但它们彼此相距较远，而在 h 图解中却彼此接近。

4.2.2 网络表示

不相似度矩阵可以转换为相似度矩阵，并视为有向加权图，再通过社会分析可视化技术进行表示，如 Calero Medina 和 van Leeuwen（2012）所示。图 5展示了一个网络表示，其中引文链接的阈值设为 100，布局采用弹簧嵌入算法，并使用 R 软件包 sna 绘制（Butts，2008）。之所以设置阈值，是因为邻接矩阵较为稠密，导致网络表示难以阅读（Ghoniem 等，2005）。即使设置了阈值，且我们的数据集规模较小，可视化结果仍不易阅读，只有孤立的期刊（1、6、10、11、15、17、18、23）例外；这些期刊对应于特定领域的期刊。此外，如果只表示链接，我们就会丢失有关各期刊剖面彼此接近程度的信息。

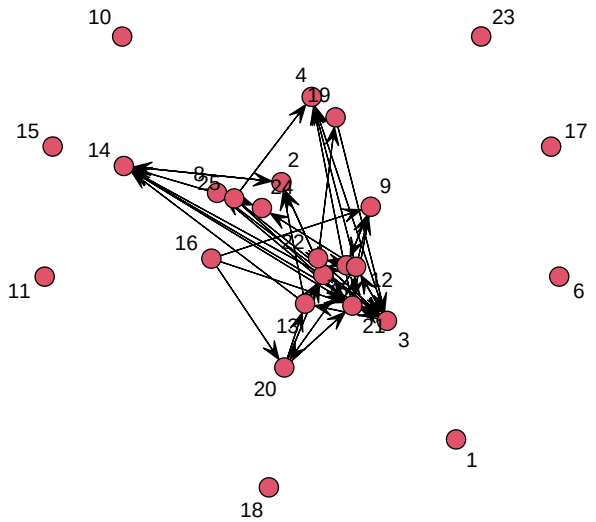

图 5 25 种统计学期刊的网络表示。表 1 列出了这些编号对应的代码。

5 结论

本研究展示了 h 图作为一种实用且有效的方法，在表示非对称不相似性方面的潜力。我们通过映射和分析期刊之间的非对称关系来说明这一点，既考察期刊的引用情况，也考察其被引情况。h 图能够同时表示这两种角色。与展开法和网络表示相比，h 图得出的结果信息更丰富。本文还重点介绍了使用 h 图的若干优势，具体包括：通过统一的欧几里得表示实现直观的可解释性；采用明确的基于特征向量的解析解，避免陷入局部极小值；在线性变换下保持尺度不变性；对大型矩阵具有较高的计算效率；以及能够直接评估拟合优度。

此外，通过原型体分析发现了原型概况，从而使我们能够更容易地发现数据结构。我们识别出三种原型期刊：一份学科专门期刊、一份理论与方法期刊，以及一份综合应用计算统计学期刊。当 k 取较大值时，会呈现出更多结构细节。特别是，当 $k = 4$ 时，一份综述期刊被识别为额外的原型体。

尽管在本案例中，由于数据中不存在明显分离的群组结构，我们采用了 ADA，但 h 图支持应用聚类技术，以帮助揭示潜在的数据结构。

当怀疑存在离群观测值时，可以按照 Daigle 和 Rivest（1992）提出的方法计算稳健 h 图。类似地，也可以按照 Moliner 和 Epifanio（2019）所述的方法计算稳健 ADA。

未来研究可以从多个方向展开，包括通过 h 图表示符号型不相似性。首先，在某些情况下，确切的不相似性未知，但可以获得其范围（Groenen 等人，2006年）或直方图（Groenen 和 Winsberg，2006年）等信息。例如，可以根据 D’Enza 等人（2021年）的思想，对区间不相似性进行 H 图绘制。其次，另一个方向是修改双原型分析（Alcacer 等人，2024年），通过对行和列考虑相同的参数，直接从不相似度矩阵中寻找原型概况。第三，可以扩展 H 图绘制，以处理二模三元不相似性（对象× 对象 × 来源）（Okada 和 Imaizumi，1997年），方法是按照 Alcacer 等人（2025a年）提出的方法，将对象 ×对象各侧的矩阵并排堆叠。第四，也是最后一个方向，可以考虑监督多维尺度分析，即将类别标签或结果变量纳入考量，并通过监督主成分分析（Barshan 等人，2011年）定义监督 h 图。

补充信息

代码和数据可作为补充信息从 https://epifanio.uji.es/RESEARCH/asyada.zip 获取。

致谢 本研究部分获西班牙科学与创新部 PID2022-141699NB-I00 和 PID2020-118763GA-I00 项目，以及瓦伦西亚自治区政府 CIPROM/2023/66 项目的支持。

参考文献

Alcacer, A., Benitez, R., Bolos, V. J., and Epifanio, I. (2025a年). 对二模三向非对称差异进行多维尺度分析：寻找原型轮廓并进行聚类。*arXiv* 预印本 *arXiv:2511.15813*. Alcacer, A., Epifanio, I., and Gual-Arnau, X. (2024年). 双原型分析：基于极值同时学习观测值和特征。*IEEE*模式分析与机器智能汇刊, 46(12):8228–8239. Alcacer, A., Epifanio, I., Mair, S., and Mørup, M. (2025b). 一篇关于原型分析的综述。*arXiv* 预印本 *arXiv:2504.12392*. Alcacer, A., Epifanio, I., Valero, J., and Ballester, A. (2021年). 结合分类与基于用户的协同过滤来匹配鞋类尺码。数学, 9(7). Barshan, E., Ghodsi, A., Azimifar, Z., and Jahromi, M. Z. (2011年). 监督主成分分析：在子空间和子流形上进行可视化、分类与回归。模式识别, 44(7):1357–1371. Borg, I. and Mair, P. (2017年). 多维尺度分析中初始构型的选择：局部极小值、拟合与可解释性。奥地利统计学杂志, 46(2):19–32. Bove, G. and Okada, A. (2018年). 非对称成对关系分析方法。《数据分析与分类进展》, 12(1):5–31. Bove, G., Okada, A., and Vicari, D. (2021年). 非对称邻近度数据分析方法. 施普林格. Butts, C. T. (2008). 使用 sna 进行社会网络分析。统计软件杂志, 24:1–51. Cabero, I. and Epifanio, I. (2020年). 寻找二元问卷的原型模式。*SORT*, 44(1):39–66. Calero Medina, C. M. and van Leeuwen, T. N. (2012年). 种子期刊引文网络图：一种基于网络理论的方法。《美国信息科学与技术学会会刊》, 63(6):1226–1234. Corsten, L. C. A. and Gabriel, K. R. (1976). 比较方差矩阵的图形探索。生物测量学, 32(4):851–863. Cutler, A. and Breiman, L. (1994年). 原型分析。*TECHNOMETRICS*, 36(4):338–347. Daigle, G. and Rivest, L.-P. (1992年). 稳健双标图。加拿大统计学杂志, 20(3):241–255. De Leeuw, J. and Mair, P. (2009年). 使用主化进行多维尺度分析：R 中的 SMA-COF。统计软件杂志, 31:1–30. Di Nuzzo, C., Vicari, D., and Vitozzi, A. (2025年). 一种用于可视化非对称邻近矩阵中流的 Candecomp/Parafac 方法。载于意大利统计学会科学会议, 第159–164页。施普林格. D'Enza, A. I., Schisa, V., and Palumbo, F. (2021年). 区间数据的主成分分析：常见方法及其变体。*Statistica Applicata-ItalianJournal of Applied Statistics*, 3:249–270. Epifanio, I. (2013年). 用于展示非度量差异矩阵的 h 图。《统计分析与数据挖掘》, 6(2):136–143。

Epifanio, I.（2014）。利用h图映射期刊间的不对称引文关系。信息科学与技术协会学报，65(6):1293–1298。Epifanio, I., Gimeno, V., Gual-Arnau, X., and Ib´a˜nez-Gual, M. V.（2023）。形状与大小空间中的原型曲线：通过代表性极端值发现曲线大数据的显著特征。*LaMatematica*，2(3):635–658。Epifanio, I., Ib´a˜nez, M. V., and Sim´o, A.（2020）。含缺失数据的原型分析：通过观察少数基于极端轮廓的样本，查看所有样本。美国统计学家，74(2):169–183。Epifanio, I., Ib´a˜nez, M. V., and Sim´o, A.（2018）。基于标志点的原型形状及其缺失数据处理扩展。《数据分析与分类进展》，12(3):705–735。Eugster, M. J. and Leisch, F.（2009）。从蜘蛛侠到英雄——R中的原型分析。*JOURNAL OF STATISTICAL SOFTWARE*，30(8):1–23。Ghoniem, M., Fekete, J.-D., and Castagliola, P.（2005）。关于使用节点—链接和基于矩阵的表示提高图的可读性：一项受控实验与统计分析。信息可视化，4(2):114–135。Groenen, P., Winsberg, S., Rodriguez, O., and Diday, E.（2006）。I-scal：区间不相似性的多维尺度分析。《计算统计学与数据分析》，51:360–378。Groenen, P. J. and Winsberg, S.（2006）。直方图不相似性的多维尺度分析。载于数据科学与分类，第161–170页。施普林格。Hastie, T., Tibshirani, R., and Friedman, J.（2009）。《统计学习要素》。数据挖掘、推断与预测。第2版，施普林格出版社，纽约。Kaufman, L. and Rousseeuw, P.（1990）。《发现数据中的组：聚类分析导论》。约翰·威立，纽约。Klavans, R. and Boyack, K. W.（2006）。为科学映射寻找更好的关联度量。《美国信息科学与技术学会会刊》，57(2):251–263。Lawson, C. L. and Hanson, R. J.（1974）。《求解最小二乘问题》。普伦蒂斯·霍尔，恩格尔伍德克利夫斯。Mair, P., Groenen, P., and de Leeuw, J.（2022）。基于主化的多维尺度分析：R中的SMACOF。*JOURNAL OF STATISTICAL SOFTWARE*，102。Moliner, J. and Epifanio, I.（2019）。稳健多变量与函数型原型分析及其在金融时间序列分析中的应用。*Physica A*：统计力学及其应用，519:195 – 208。Nadler, B., Lafon, S., Coifman, R. R., and Kevrekidis, I. G.（2006）。动力系统的扩散映射、谱聚类与反应坐标。应用与计算调和分析，21:113–127。Okada, A. and Imaizumi, T.（1997）。二模三维邻近度的非对称多维尺度分析。分类学杂志，14(2):195–224。Okada, A. and Imaizumi, T.（2024）。非对称关系的应用多维尺度分析。施普林格。Olszewski, D.（2023）。一种保拓扑的非对称邻域检索可视化工具。专家系统与应用，225:120175。

Olszewski, D.（2024）。用于降维与数据可视化的非对称等距映射。载于国际人工神经网络会议，第102– 115页。施普林格。Pe̛kalska, E. and Duin, R. P. W.（2005）。《模式识别中的不相似性表示：基础与应用》。世界科学出版社。Podani, J. and Mikl´os, I.（2002）。主坐标分析中的相似系数与马蹄效应。《生态学》，83(12):3331–3343。Pudovkin, A. I. and Garfield, E.（2002）。寻找语义相关期刊的算法程序。《美国信息科学与技术学会会刊》，53(13):1113–1119。R 开发核心团队（2025）。*R*：一种用于统计计算的语言与环境。R 统计计算基金会，维也纳，奥地利。ISBN 3- 900051-07-0。Roweis, S. T. and Saul, L. K.（2000）。通过局部线性嵌入实现非线性降维。《科学》，290(5500):2323–2326。Schneider, J. W.（2009）。利用多维展开法映射期刊间的交叉引文活动：对绘图研究的启示。载于第十二届国际科学计量学与信息计量学会议论文集（*ISSI 2009*），第443–454页。BIREME/PAHO/WHO与里约热内卢联邦大学，巴西里约热内卢。Schwarz, A. and Karlsson, L.（2019）。用于非对称特征值问题的可扩展特征向量计算。《并行计算》，85:131–140。Seber, G.（1984）。多变量观测。约翰·威立。Sun, Z., Pedretti, G., Ambrosi, E., Bricalli, A., and Ielmini, D.（2020）。时间复杂度为 o (1) 的存内特征向量计算。《先进智能系统》，2(8):2000042。Tenenbaum, J. B., de Silva, V., and Langford, J. C.（2000）。非线性降维的全局几何框架。《科学》，290(5500):2319–2323。Thurau, C., Kersting, K., Wahabzada, M., and Bauckhage, C.（2012）。面向可持续性的描述性矩阵分解：采用对立原则。《数据挖掘与知识发现》，24(2):325–354。Vinu´e, G. and Epifanio, I.（2017）。面向体育分析的原型体分析。《数据挖掘与知识发现》，31(6):1643–1677。Vinu´e, G. and Epifanio, I.（2021）。用于大规模函数型数据异常检测的稳健原型体。《数据分析与分类进展》，15(2):437–462。Vinu´e, G., Epifanio, I., and Alemany, S.（2015）。原型体：一种定义代表性原型数据的新方法。《计算统计学与数据分析》，87:102 – 115。