arXiv:2601.05392v1 [stat.ME] 8 Jan 2026

包含名义多选题的问卷原型案例

Aleix Alcacer 和 Irene Epifanio

摘要原型分析是一种探索性工具，可将一组观测值解释为纯粹（极端）模式的凸组合。当这些模式对应于样本中的实际观测值时，便称为原型体。我们首次提出将原型点分析应用于名义观测值，具体而言，是从仅允许单选的包含名义型多选题的问卷中识别原型案例。这种方法有助于我们更好地理解名义型数据集，其作用类似于它在多变量情境中的应用。我们将这种方法与原型分析和概率原型分析进行比较，并通过一个真实案例——德国信用数据集——展示该方法的优势。

关键词：原型分析、原型点分析、名义观测值、虚拟变量

1 引言

名义变量将特征划分为两个或多个不同类别，在研究中十分常见。名义数据是通过观察而非测量获得的，其类别无序且间距不等，也不存在有意义的零点。例如，包含多选题的调查问卷中就会出现这类数据。在我们的案例中，参与者只能从预先设定的一组选项中选择一个答案。

我们首次为这类数据引入探索性工具原型点分析（ADA），帮助人们理解、描述、可视化并提取易于解读的信息，即使是非专业人士也能使用。ADA 是一种无监督统计

Aleix Alcacer 西班牙卡斯特利翁-德拉普拉纳海梅一世大学 电子邮件：aalcacer@uji.es Irene Epifanio 西班牙海梅一世大学及 valgrAI 电子邮件：epifanio@uji.es

学习技术，旨在通过已知为原型体的k 个纯模式的凸组合（或混合）来近似样本数据。这些原型体代表样本中的极端观测值。将它们纳入样本有助于提升可解释性，而它们作为极端案例的特性则有助于理解数据。研究表明，当数据通过其极端组成部分呈现，或通过对比不同观测值的特征呈现时，人们能更有效地理解数据 [8]。 [10] 提出了 ADA，作为原型分析（AA）的一种衍生方法；AA 由 [2]开发。与 ADA 类似，AA 旨在通过原型的混合来近似实值连续多变量数据；不过，原型并非实际案例，而是数据点的组合。此外，[6] 提出了 AA 的概率框架（PAA），通过在参数空间中运作来处理名义观测值 [7] 。尽管 AA 和 ADA 最初是为实值观测值设计的，本研究旨在将原型方法扩展到名义数据。对于二分变量或二元变量这一特殊情况，即仅有两个类别的名义变量， [1] 提出使用 ADA，而非 AA 或 PAA。对于有序数据，ADA 也曾用于一种基于有序刻板模型的两步法 [3]。

第2节介绍方法，第3节介绍数据和结果。最后，第4节给出一些结论。代码和数据可在 http://www3.uji.es/~epifanio/RESEARCH/nom.zip 获取。

2 方法

名义变量不能以类别形式用于 AA 或 ADA，因为它们与数值变量一起使用。虚拟变量编码可以将类别转换为 AA 和 ADA 能够处理的格式，因为字符列会被转换为一系列二元列。因此，第一步是将名义矩阵的列转换为虚拟变量（二元）列。第二步，我们将 ADA 应用于这个$n \times m$二元矩阵$\mathbf{X}$，其中包含n个观测值和m 列。

在 ADA 中，定义了三个矩阵：a) k 个原型体 $\mathbf{z}_j$，它们作为 $k \times m$ 矩阵 $\mathbf{Z}$ 的行表示，其中包含 $\mathbf{X}$ 的具体行；b) 一个 $n \times k$ 矩阵$\alpha = (\alpha_{ij})$，其中包含用于将每个观测值 $\mathbf{x}_i$近似表示为原型体（$\hat{\mathbf{x}}_i = \sum \alpha_{ij}\mathbf{z}_j$）凸组合的混合系数；以及 c) 一个 $k \times n_{j=1}$矩阵 $\beta = (\beta_{jl})$，其中包含用于定义每个原型体（$\mathbf{z}_j = \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l$）的二元系数。为确定这些矩阵，我们的目标是在满足特定约束的条件下，将以下残差平方和（RSS）最小化（其中 $\|\cdot\|$ 表示向量的欧几里得范数）：

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

满足以下约束：

1) $\sum_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ for $i = 1, \ldots, n$ and

2) $\sum_{l=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \in \{0, 1\}$ and $j = 1, \ldots, k$.

此问题中矩阵的估计可以使用 [10] 提出的算法来完成。

原型分析的核心理念是，我们可以识别出一组原型模式，并认为数据可以表示为这些模式的混合。在名义数据或二元数据的情境下，原型模式也必须是名义型或二元型，以反映数据来源群体的特性。例如，如果妊娠是二元变量之一，那么将原型观测视为一名怀孕程度为0.8次的女性显然不合逻辑。换言之，原型模式必须是名义型或二元型，才能保持清晰且易于解释；这对原型技术至关重要。它们应当代表可观测的现象，而非抽象概念。

此外，若要将数据描述为混合，我们必须假定观测值存在于向量空间中，也就是说，它们可以按比例缩放（在此处为区间 $[0, 1]$内）并进行组合。兼顾这些理念的一种合适方案是对$\mathbf{X}$应用ADA，因为可行的原型模式将属于观测样本。事实上，ADA最初就是为寻找真实存在而非虚构的模式而开发的 [10]。相反，应用AA或PAA得到的原型不一定是二元的，也可能不属于可行解集。事实上， [7]在其实验中将通过AA或PAA得到的原型二元化了。

事实上，对于可行解并非连续的问题采用连续优化方法，可能会导致严重失效 [4, 章 13]。即使考察连续解附近的所有可行二元解，也无法保证该方法能够得出令人满意的解。一项涉及二元观测值的模拟实验由 [1], 开展；研究发现，与 AA 和 PAA 依次得到的原型相比，ADA 得到的原型体与真实原型更为接近。具体而言，ADA 的平均误分类误差最低。

因此，我们建议使用 ADA 处理名义型或二元观测值。

3 结果

我们考察了 [7],所使用的德国信用数据集，该数据集描述了个人及其信用申请。此数据集可在 [5]获取，其中包含 1,000 人的信息和五个名义变量（括号内为类别数量）：信贷用途（10）、就业期限（5）、个人状况和性别

（4）、就业状况（4）和信用风险（2）。本分析旨在识别信用申请的原型剖面。首先，我们创建虚拟变量，然后应用 ADA、AA 和 PAA。对于 AA 和 PAA 的结果，我们将所有大于 0.5 的值替换为 1，其余值替换为 0，以此进行二值化。为了便于比较，我们使用 $k = 10$，这是 [7]使用的值。

表1展示了由 ADA 得出的原型体。其中三个轮廓对应不良信用风险状况。使用 AA 和 PAA 时可能出现的问题之一是，没有任何类别超过二值化阈值。事实上，在 PAA 的第 6 个原型轮廓及变量“信贷用途”中就出现了这种情况。图1展示了按照 [7] 中的说明，通过单纯形可视化投影的 alpha 值。“不良”观测值位于原型体 6、9 和 10 附近。

为评估原型剖面的互补性，进而评估其极端程度，我们计算了原型剖面之间的汉明距离（二进制向量之间的曼哈顿距离）。表 2 汇总了各方法所得原型剖面之间的距离，并列出这些距离的总和

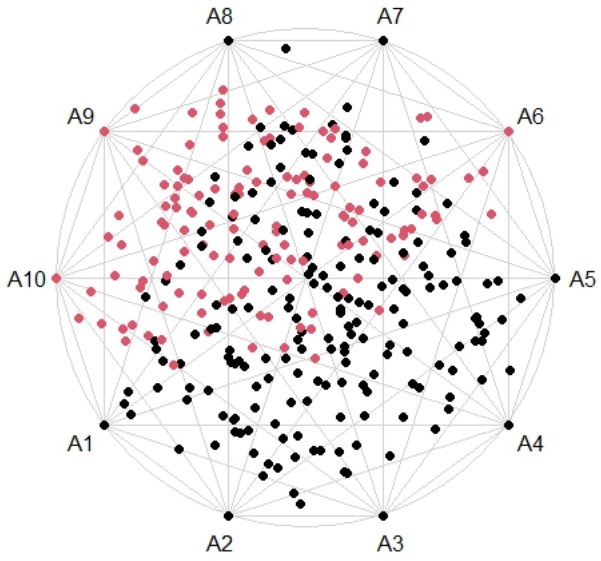

图1 德国信用数据集的单纯形可视化。信用风险因子以黑色（“良好”）和红色（“不良”）显示。

（见“Total”列）。因此，原型体能够识别出更多互补画像。

表1 ADA 得出的原型体描述，以及被选为原型体的行数（第一列）。

案例	信贷用途	就业期限	性别与个人状况	就业状况	信用风险
23	汽车（新车）	... <1 年	男性：单身	非熟练工—居民	Good
591	收音机/电视机	4 <=... <7 年	女性	非熟练工—居民	Good
32	家具/设备	1 <=... <4年	男性：单身	熟练员工/官员	Good
15	汽车（新车）	1 <= ... <4 年	女性	技术员工/公务员	Good
129	汽车（二手）	失业者	男性：单身	个体经营者/高资质员工	Good
125	家具/设备	... <1 年	女性	技术员工/公务员	Bad
1	收音机/电视机	.. >= 7 年	男性：单身	熟练员工/官员	Good
146	商业	4 <= ... <7 年	男性：单身	熟练员工/官员	Good
107	汽车（新车）	.. >= 7 年	男性：单身	个体经营者/高资质员工	Bad
436	收音机/电视机	1 <=... <4 年	男性：已婚/丧偶	熟练员工/官员	Bad

表 2 各方法的原型剖面之间具有每种汉明距离的条目数。

距离	0	4	5	6	7	8	10	总计	ADA	10	8
0	460	28	8	612	AA	10	12	0	420	28	
8	604	PAA	10	12	12	266	28	6	590		

4 结论

我们提出使用 ADA 识别名义数据中的原型模式，以增进对数据集的理解。实际应用的结果展示了 ADA 对包含名义型多选题的问卷的优势，可作为其他成熟方法的替代或补充。尽管统计学在很大程度上依赖于这样一项原则：对数据集中的大量元素进行平均是有益的，但本文采取了不同的方法。我们聚焦于少量代表性观测（原型观测），并通过这些极端案例的混合来解释数据的构成。我们的研究结果表明，这种方法能够提供大量信息，并且是使数据集更易于理解和使用的宝贵工具，即使对于非专业人士也是如此。

作为未来的研究工作，一个并非简单直接的扩展方向是处理混合数据场景，其中包括实值数据和类别数据，以及缺失值。此外，从计算角度来看，在处理超大数据集时，[10] 提出的ADA算法可能会较慢。在这种情况下，可以使用 [9]提出的方法来利用大数据集计算ADA。

致谢 本研究得到瓦伦西亚自治区政府 CIPROM/2023/66、西班牙科学与创新部 PID2022-141699NBI00 和 PID2020-118763GA- I00 的支持。

参考文献

[1] Cabero I，Epifanio I（2020）为二元问卷寻找原型模式。SORT 44(1)：39–66[2] Cutler A，Breiman L（1994）原型分析。Technometrics 36(4)：338–347[3] Fern´andez D，Epifanio I，McMillan LF（2021）有序数据的原型分析。Information Sciences 579：281–292[4] Fletcher R（2000）《实用优化方法》，第2版。约翰·威立父子出版公司[5] Hofmann H（1994）Statlog（德国信用数据）。UCI机器学习资料库，DOI：https://doi.org/10.24432/C5NC77[6] SethS，Eugster MJ（2016）概率原型分析。MachineLearning 102(1)：85–113[7] Seth S，Eugster MJA（2016）名义观测值的原型分析。IEEE Trans Pattern Anal Mach Intell 38(5)：849–861[8] Thurau C，Kersting K，Wahabzada M，Bauckhage C（2012）用于可持续性的描述性矩阵分解：采用对立原则。Data Mining and Knowledge Discovery 24(2)：325–354[9] Vinu´e G，Epifanio I（2021）用于大规模函数型数据异常检测的稳健原型点。Advances in Data Analysis and Classification 15(2)：437–462[10] Vinu´e G，Epifanio I，Alemany S（2015）原型体：定义代表性原型数据的新方法。Computational Statistics & Data Analysis 87：102–115