arXiv:2301.01364v1 [stat.AP] 3 Jan 2023

幂变换数据集的对应分析笔记

Choulakian V.，蒙克顿大学，加拿大
vartan.choulakian@umoncton.ca

2023年1月

R ′结 ′语

我们旨在清晰、简明地阐述自2009年起由Grenacre发起的幂变换数据或 Box-Cox 变换数据的对应分析。我们区分两类数据集：严格正值数据集和含零值的数据集；本文重点讨论后者。

关键词：列联表；组成数据；双重中心化；幂变换；交互作用；零值；出租车对应分析。

AMS 2010 分类号：62H25, 62H30

1 引言

对应分析（CA）和对数比分析（LRA）是两种常用方法，用于分析和可视化一个大小为 $I \times J$ 的双向列联表或组成数据集 N $=(n_{ij})$，其中行的取值为 $i = 1, ..., I$，列的取值为 $j = 1, ..., J.$。LRA 被认为是分析和可视化 N 的理想方法，因为它具有行列尺度不变性；但只有在 $n_{ij} > 0.$ 时才能应用。对应分析则可在 $n_{ij} \geq 0$ 时应用于数据，但它不具有行列尺度不变性，因为它取决于行边际和列边际。

本文的核心主题是 Greenacre（2010年，结果1），我们将其重新命名为定理，并将其应用于 N。Greenacre定理涉及当 $n_{ij} > 0$ 时，$\mathbf{N}^{(\alpha)} = (n_{ij}^{\alpha})$ 的对应分析在 $\alpha \in (0, 1]$ as $\alpha \to 0$ 时收敛至 N 的均匀加权对数比分析。不过，Greenacre（2011年，2022年）试图

证明其在 $n_{ij} \geq 0$ 时的适用性——实用性；也就是说，当数据集 N 含有零值单元格时。我们旨在进一步阐明后一种情况。

本文组织如下：第2节介绍预备知识；第3节介绍关于严格正数据集的对应分析的Greenacre定理；第4节介绍含有零值单元格的非负数据集的对应分析；第5节讨论三个含零单元格的真实数据集示例；最后，我们在第6节总结。

1.1 参考文献

对应分析和对数比分析基于三种不同原理：对应分析基于Benz´ecri的分布等价性原理，RC关联模型基于Yule尺度不变原理，而组成数据分析基于Aitchison子组成相干原理。Choulakian等人（2023年）对这三种原理作了近期讨论。

Benz´ecri（1973）是关于对应分析的参考著作。Beh和Lombardo（2014）对对应分析及其变体作了全面综述。

对数比分析包括两种各自发展成熟的方法：Goodman（1979、1981a、1981b、1991、1996）提出的用于分析列联表的RC关联模型，以及Aitchison（1986）提出的一组组成向量（组成数据分析）；另请参见Greenacre（2018）等人的相关研究。

关于对应分析与对数比分析之间的关系，已有许多论文对此进行讨论；例如，Goodman（1996）、Cuadras等（2006）、Cuadras和Cuadras（2015年）、Greenacre（2009、2010）、Beh和Lombardo（2022）以及Choulakian（2022）。

2 列联表分析基础

我们考虑一个双向列联表 $\mathbf{N} = (n_{ij})$，其行类别为 $i = 1, ..., I$，列类别为 $j = 1, ..., J$；$\mathbf{P} = \mathbf{N}/n = (p_{ij})$ 是列联表 N 对应的 $I \times J$阶对应矩阵（概率表）。按惯例定义 $p_{i+} = \sum_{j=1}^{J} p_{ij}$，$p_{+j} = \sum_{i=1}^{I} p_{ij}$，向量 $\mathbf{r} = (p_{i+}) \in \mathbb{R}^I$，向量 $\mathbf{c} = (p_{+j}) \in \mathbb{R}^J$，以及 $\mathbf{M}_I = Diag(\mathbf{r})$，即对角元素为 p_{i+} 的对角矩阵；$\mathbf{M}_J = Diag(\mathbf{c})$ 的定义类似。我们假设 $\mathbf{M}_I$ 和 $\mathbf{M}_J$ 分别是 $I \times I$ 阶和 $J \times J$ 阶的正定度量矩阵；这意味着 $\mathbf{M}_I$ 和 $\mathbf{M}_J$ 的对角元素严格为正。

2.1 行类别与列类别的独立性

a) I 行类别与 J 列类别相互独立，则

$$\sigma_{ij} = p_{ij} - p_{i+}p_{+j} = 0 \tag{1}$$

对于 $i = 1$、...、I 和 $j = 1$、...、J，其中 σ_{ij} 是相对于独立模型 $p_{i+}p_{+j}$. 的 p_{ij} 的残差矩阵

注 1：列联表 N $= (n_{ij})$ 也可以表示（编码）为一个大小为 $n \times (I+J)$ 的指示矩阵 Z $= [\mathbf{Z}_I \mathbf{Z}_J] = [(z_{\alpha i})\ (z_{\alpha j})]$，其中，当个体 α 不具有行变量的 i 水平时，$z_{\alpha i} = 0$；当个体 α 具有行变量的 i 水平时，$z_{\alpha i} = 1$；当个体 α 不具有列变量的 j 水平时，$z_{\alpha j} = 0$；当个体 α 具有列变量的 j 水平时，$z_{\alpha j} = 1$。注意，N $= \mathbf{Z}_I'\mathbf{Z}_J$，且 $\sigma_{ij} = p_{ij} - p_{i+}p_{+j}$ 是 $\mathbf{Z}_I$ 的第 i 列与 $\mathbf{Z}_J$ 的第 j 列之间的协方差。

b) 独立性假设 $\sigma_{ij} = 0$ 也可以通过另一种方式解释为

$$\begin{aligned} \Delta_{ij} &= (\frac{p_{ij}}{p_{i+}p_{+j}} - 1) = 0 \\ &= \frac{1}{p_{i+}}(\frac{p_{ij}}{p_{+j}} - p_{i+}) \\ &= \frac{1}{p_{+j}}(\frac{p_{ij}}{p_{i+}} - p_{+j}); \end{aligned} \tag{2}$$

这就是行列同质性模型。Benz´ecri（1973，第31页）将条件概率向量（$\frac{p_{ij}}{p_{+j}}$对于 $i = 1$、...、I 且 j 固定）称为第 j 列的剖面。他还将 $\frac{p_{ij}}{p_{i+}p_{+j}}$ 的元素称为概率测度（p_{ij}）相对于乘积测度$p_{i+}p_{+j}$的密度函数。在 Goodman（1996）以及 Beh 和 Lombardo（2014，第123页）中，元素 $\frac{p_{ij}}{p_{i+}p_{+j}}$ 被称为皮尔逊比率。

c) 表示独立性假设 $\sigma_{ij} = 0$以及行列同质性模型 $\Delta_{ij} = 0$ 的第三种方式，是通过 (w_i^R, w_j^C) 加权对数线性表述（方程 (3)），假设 $p_{ij} > 0$ 并定义 G_{ij} =log(p_{ij}),

$$\lambda(p_{ij}, w_i^R, w_j^C) = G_{ij} - G_{i+} - G_{+j} + G_{++} = 0, \tag{3}$$

其中 $G_{i+} = \sum_{j=1}^{J} G_{ij} w_j^C$、 $G_{+j} = \sum_{i=1}^{I} G_{ij} w_i^R$和 $G_{++} = \sum_{j=1}^{J} \sum_{i=1}^{I} G_{ij} w_j^C w_i^R$; $w_j^C > 0$ 和 $w_i^R > 0$满足 $\sum_{j=1}^{J} w_j^C = \sum_{i=1}^{I} w_i^R = 1$，它们是先验固定或依赖于数据的概率权重。两种常用权重是边际权重 ($w_i^R = p_{i+}$, $w_j^C = p_{+j}$) 和均匀权重 ($w_i^R = 1/I$, $w_j^C = 1/J$)。这在 Goodman（1996）方程 7 或 Goodman（1991）方程 2.2.6 中是隐含的；而在 Egozcue 等人（2015）中则是明确给出的。

公式（3）等价于对数比

$$\log(\frac{p_{ij} p_{i_1 j_1}}{p_{ij_1} p_{i_1 j}}) = 0 \quad \text{for } i \neq i_1 \text{ and } j \neq j_1,$$

Goodman（1979年，公式 2.2）将其称为“零关联”模型。

公式（3）也等价于

$$p_{ij} = \frac{\exp(G_{i+}) \exp(G_{+j})}{\exp(G_{++})},$$

由此我们推断：在独立性假设下，边际行概率向量（p_{i+}）与加权几何均值向量（exp(G_{i+})）成比例；列也具有类似性质；例如参见 Egozcue 等人（2015年）。

2.2 交互分解

假设独立性—齐性—零关联模型并不成立，那么三种等价的模型表述（1）、（2）、（3）中的每一种都可以通过添加 k 双线性项进行推广，以解释 I 行与 J 列之间的非独立—非齐性—关联，即交互作用，其中 $k = rank$(N)-1。

我们用 τ_{ij}. 表示交互指标 (1)、(2)、(3) 中的任意一个

Benz´ecri（1973，第 2 卷，第 31–32 页）强调了行权重和列权重或度量在多维数据分析中的重要性；因此，在法国数据分析界，任何研究都从三元组 (X, M$_I$, M$_J$) 开始，其中 X 表示数据集，M$_I$ =($Diag(m_i^r)$) 是定义在行上的度量矩阵，M$_J$ = ($Diag(m_j^c)$) 是定义在列上的度量矩阵。我们遵循相同的步骤，但 X 是预处理后的数据，其中：

a) 在协方差分析中，X =(τ_{ij}) =($IJ\sigma_{ij}$)以及

$(\mathbf{M}_I, \mathbf{M}_J) = (Diag(1/I)\ Diag(1/J))$。b) 在对应分析中，$\mathbf{X} = (\tau_{ij}) = (\Delta_{ij})$以及 $(\mathbf{M}_I, \mathbf{M}_J) = (Diag(p_{i+})\ Diag(p_{+j}))$；c) 在对数比分析中，$\mathbf{X} = (\tau_{ij}) = (\lambda(p_{ij}, w_i^R, w_j^C))$以及 $(\mathbf{M}_I, \mathbf{M}_J) = (Diag(w_i^R)\ Diag(w_j^C))$，其中 $\sum_{j=1}^{J} w_j^C = \sum_{i=1}^{I} w_i^R = 1.$

注意，在 a) 中，我们将 σ_{ij} 乘以 IJ，从理论上统一协方差分析与对应分析，详见关于双重中心化的小节。

我们通过奇异值分解（SVD）或出租车奇异值分解（TSVD）将式 (1)、(2) 和 (3) 中的交互作用分解为

$$\tau_{ij} = \sum_{\alpha=1}^{k} f_\alpha(i) g_\alpha(j) / \delta_\alpha, \tag{4}$$

其中，f_α（i）是沿第 α 个主维度的第 i 行主坐标，g_α（j）是沿第 α 个主维度的第 j 列主坐标。此外，δ_α 是第 α 个主轴的离散度量。

备注 2：

a) 在 SVD 情形下，参数 (f_α（i） g_α（j） δ_α) 满足以下条件：对于 α、$\beta = 1$、...、k

$$\delta_\alpha^2 = \sum_{i=1}^{I} {f_\alpha}^2(i) m_i^r = \sum_{j=1}^{J} g_\alpha^2(j) m_j^c$$

$$0 = \sum_{i=1}^{I} f_\alpha(i) m_i^r = \sum_{j=1}^{J} g_\alpha(j) m_j^c$$

$$0 = \sum_{i=1}^{I} f_\alpha(i) f_\beta(i) m_i^r = \sum_{j=1}^{J} g_\alpha(j) g_\beta(j) m_j^c \ \text{ for } \alpha \neq \beta.$$

b) 在 TSVD 情况下，参数（f_α（i） g_α（j） δ_α）满足以下条件：对于 α、$\beta = 1$、...、k

$$\delta_\alpha = \sum_{i=1}^{I} |f_\alpha|(i) m_i^r = \sum_{j=1}^{J} |g_\alpha(j)| m_j^c$$

$$0 = \sum_{i=1}^{I} f_\alpha(i) m_i^r = \sum_{j=1}^{J} g_\alpha(j) m_j^c$$

$$0 = \sum_{i=1}^{I} f_\alpha(i)\ sign(f_\beta(i)) m_i^r = \sum_{j=1}^{J} g_\alpha(j)\ sign(g_\beta(j)) m_j^c \ \text{ for } \alpha > \beta.$$

关于 TSVD 的说明可参见 Choulakian（2006、2016年）。

备注 3

a) 在 $(\tau_{ij}) = (IJ\sigma_{ij})$ 的情形下，双线性分解 (4) 也称为跨电池分析，最初由 Tucker（1958年）提出；后来，Tenenhaus 和 Augendre（1996年）在对应分析领域重新引入了这一方法，并指出，对某些列联表进行的 Tucker 奇异值分解比对应分析能揭示更有趣（且更易解释）的结构。

b) 在 $(\tau_{ij}) = (\Delta_{ij})$ 的情形下，对应分析分解有多种解释。就数据分析而言，Benz´ecri（1973）将其解释为对行剖面和列剖面进行加权主成分分析。对应分析的另一种实用解释可与 Tucker 的跨电池分析相媲美，即 Hotelling（1936年）的典型相关分析，参见 Lancaster（1958年）和 Goodman（1991年，1996年）。

c) 在 $(\tau_{ij}) = (\lambda(p_{ij}, w_i^R, w_j^C))$ 和 $(\mathbf{M}_I, \mathbf{M}_J) = (Diag(w_i^R)\ Diag(w_j^C))$ 的情形下，其中 (w_i^R, w_j^C) 为预先指定；我们将此情形记为 TLRA 或 LRA；一般预先指定权重的示例见 Egozcue 和 Pawlowsky-Glahn（2016年）。对于重要的特殊情形 $\lambda(p_{ij}, w_i^R, w_j^C) = \lambda(p_{ij}, 1/I, 1/J)$，我们得到均匀加权（或出租车度量）对数比分析 uwLRA（或 uwTLRA）。

d) 我们关注三个交互指标 (1)、(2) 和 (3) 的尺度依赖性或尺度独立性。三个指标取决于 p_{ij}，其中 $p_{ij} = n_{ij} / \sum_{i,\ j} n_{ij\cdot}$。为强调这种依赖关系，我们将三个交互指标中的任意一个表示为 $\tau_{ij}(n_{ij}) = \tau(p_{ij}, m_i^R, m_j^C)$。遵循 Yule（1912年），我们给出如下

定义1：若对于任意尺度 $a_i > 0$ 和 $b_j > 0$，$\tau_{ij}(n_{ij}) = \tau_{ij}(a_i n_{ij} b_j)$，则交互指标 $\tau_{ij}(n_{ij})$ 是尺度不变的。

显然，三个采用数据依赖边际权重（p_{i+}、p_{+j}）的指标都不是尺度不变的。

关于关联指标 (3)，我们有以下引理，参见 Choulakian（2022）或 Choulakian等（2023）。

引理 1：a) 具有预先指定权重（w_i^R、w_j^C）的关联指标（3）

具有尺度不变性。也就是说,

$$\lambda(n_{ij}, w_i^R, w_j^C) = \lambda(a_i n_{ij} b_j / n^*, w_i^R, w_j^C),$$

对于任意尺度 $a_i > 0$、 $b_j > 0$ 和 $n^* = \sum_{i,\ j} a_i n_{ij} b_j$.

b) 一阶近似为: $\lambda(n_{ij}, w_i^R, w_j^C) \approx (\frac{p_{ij}}{w_j^C w_i^R} - \frac{p_{i+}}{w_i^R} - \frac{p_{+j}}{w_j^C} + 1)$。

2.3 双重中心化

交互矩阵 (τ_{ij}) 经过双重中心化，即:

$$\begin{aligned} 0 &= \sum_{i=1}^{I} m_i^r m_j^c \tau_{ij} \\ &= \sum_{j=1}^{J} m_i^r m_j^c \tau_{ij}. \end{aligned} \tag{5}$$

根据 Tukey（1977年，第10章），双重中心化有两种：行加列和行乘列；我们将它们分别称为加性和乘性，下面将对此进行描述。

我们考虑三元组 $(\mathbf{Y}, \mathbf{M}_I, \mathbf{M}_J)$，其中 $Y_{ij} = h(p_{ij})$，且 h是一个函数，而 $(\mathbf{M}_I, \mathbf{M}_J) = (Diag(m_i^r)\ Diag(m_j^c))$。我们计算三个均值、两个边际和总计: $Y_{i+} = \sum_{j=1}^{J} Y_{ij} m_j^c$、 $Y_{+j} = \sum_{i=1}^{I} Y_{ij} m_i^r$和 $Y_{++} = \sum_{j=1}^{J} \sum_{i=1}^{I} Y_{ij} m_i^r m_j^c$.

乘性双重中心化为

$$\tau_{ij} = Y_{ij} - \frac{Y_{i+} Y_{+j}}{Y_{++}}, \tag{6}$$

其中 $\text{rank}(\tau_{ij}) = \text{rank}(Y_{ij}) - 1$.

加性双重中心化为

$$\tau_{ij} = Y_{ij} - Y_{i+} - Y_{+j} + Y_{++}, \tag{7}$$

其中 $\text{rank}(\tau_{ij}) = \text{rank}(Y_{ij}) - 1$ 或 $\text{rank}(Y_{ij}) - 2$.

这两种双重中心化的一个明显区别是，(7) 对行或列加性常数具有不变性；定义1和引理1a由此成为一个推论。

我们考虑 $Y_{ij} = h(p_{ij})$ 的两种函数形式。

a) $Y_{ij} = h(p_{ij}) = \frac{p_{ij}}{m_i^r m_j^c}$是联合概率测度 p_{ij} 相对于乘积测度 $m_i^r m_j^c$. 的密度函数。首先，当 $m_i^r = 1/I$and$m_j^c = 1/J$.时，由式 (6) 得到协方差交互作用 (τ_{ij}) =($IJ\sigma_{ij}$)。因此，秩($IJ\sigma_{ij}$) =秩(IJp_{ij}) -1.其次，当 $m_i^r = p_{i+}$and$m_j^c = p_{+j}$.时，由式 (6) 或式 (7) 得到对应分析交互作用 (τ_{ij}) =(Δ_{ij})。因此，秩(Δ_{ij}) =秩($\frac{p_{ij}}{p_{i+}p_{+j}}$) -1.第三，当 $m_i^r = w_i^R$and $m_j^c = w_j^C$.时，由式 (7) 得到引理 1b 的右侧。

b) $Y_{ij} = h(p_{ij})$ =对数 p_{ij}针对 $p_{ij} > 0$.首先，由（7）得到对数交互（τ_{ij}） =($\lambda(p_{ij}, w_i^R, w_j^C)$)，其中 $m_i^r = w_i^R$且 $m_j^c = w_j^C$.。因此，秩（$\lambda(p_{ij}, w_i^R, w_j^C)$) =秩(log p_{ij}) -1, 如 Goodman（1991，表 11）所示；或者秩(log p_{ij}) -2，如 Goodman（1991，表 10）所示。其次，据我们所知，由（6）得到的对数交互尚未得到应用，这很可能是因为它不是行列尺度不变的；另见第 3.1 小节。

Choulakian等人（2023年）使用残差符号质量指数 QSR，在上述不同情形中选择一种最优情形。

3 正值幂变换数据的对应分析

此处陈述 Greenacre定理，并在附录中给出数学证明。

设 ($p_{ij}^{(\alpha)} = \frac{n_{ij}^{\alpha}}{\sum_{i,j} n_{ij}^{\alpha}}$)（ $i = 1, ..., I$）和（ $j = 1, ..., J$）分别为幂变换严格正数据的对应表和对应分析交互矩阵 ($\Delta_{ij}^{(\alpha)} = \frac{p_{ij}^{(\alpha)} - p_{i+}^{(\alpha)} p_{+j}^{(\alpha)}}{p_{i+}^{(\alpha)} p_{+j}^{(\alpha)}}$)。

定理（Greenacre（2010年，结果 1）
在假设 $n_{ij} > 0$ 下，对于 $\alpha \to 0$ 和 $\alpha > 0$

$$\begin{aligned}\Delta_{ij}^{(\alpha)} &= \frac{p_{ij}^{(\alpha)} - p_{i+}^{(\alpha)} p_{+j}^{(\alpha)}}{p_{i+}^{(\alpha)} p_{+j}^{(\alpha)}} \\ &\approx \alpha\lambda(p_{ij}, w_i^R = 1/I, w_j^C = 1/J) \\ &= \alpha(\log p_{ij} + \frac{1}{IJ}\sum_{i,j}\log p_{ij} - \frac{1}{I}\sum_{i}\log p_{ij} - \frac{1}{J}\sum_{j}\log p_{ij})\end{aligned}$$

此外， $p_{i+}^{(\alpha)} \approx 1/I$对于 $i = 1, ..., I$以及 $p_{+j}^{(\alpha)} \approx 1/J$对于 $j = 1, ..., J$.

理解该定理的一种方法是考察以下使用引理1b得到的近似序列，以及这一事实： $p_{i+}^{(\alpha)} \approx 1/I$对于 $i = 1, ..., I$且$p_{+j}^{(\alpha)} \approx 1/J$对于 $j = 1, ..., J$当 $\alpha \to 0$.

$$\begin{aligned}\lambda(p_{ij}^{(\alpha)}, w_i^R = 1/I, w_j^C = 1/J) &= \alpha\lambda(p_{ij}, w_i^R = 1/I, w_j^C = 1/J) \\ &\approx (IJp_{ij}^{(\alpha)} + 1 - Ip_{i+}^{(\alpha)} - Jp_{+j}^{(\alpha)}) \\ &\approx (IJp_{ij}^{(\alpha)} - 1) \\ &\approx \Delta_{ij}^{(\alpha)}.\end{aligned}$$

3.1 备注

a) 该定理很有意思，但在实证情境中并无用处，因为可以直接计算对数交互 $\lambda(p_{ij}, w_i^R = 1/I, w_j^C = 1/J)$。

b) 在上述定理中，假设 $n_{ij} > 0$ 至关重要：如果$\alpha = 0$，那么对于所有 (i, j), $n_{ij}^{\alpha} = n_{ij}^{0} = 1$ ；也就是说，矩阵 $(n_{ij}^{\alpha} = n_{ij}^{0} = 1)$ 的秩为 1；因此，定理中等式两边均为零。

c) 在附录中上述定理的证明里，可以看到

$$\lim_{\alpha \to 0} \frac{\Delta_{ij}^{(\alpha)}}{\alpha} = \lambda(p_{ij}, w_i^R = 1/I, w_j^C = 1/J).$$

d) 假设 $n_{ij} \geq 1$；那么另一种幂变换是对数变换，也称为 Box-Cox变换，$L = (log n_{ij} = \lim_{\alpha \to 0} \frac{1}{\alpha}(n_{ij}^{\alpha} - 1))$。我们尚未见过对应分析应用于 L 的案例，参见关于双重中心化的第 2.3 小节。请注意，Greenacre定理中的假设 $n_{ij} > 0$ 比使用对数变换所需的假设 $n_{ij} \geq 1$ 更弱。Goodman 基于对数变换建立的 RC 模型，是用于分析列联表（计数严格为正的表），而非组成数据；组成数据中，以百分比表示的数据通常既有严格为正且小于 1 的值，也有大于 1 的正值；例如，参见 Greenacre 和 Lewi（2008年）研究的考古组成 CUPS数据。

e) Greenacre定理与Goodman的无边际对应分析（多重因子对应分析）之间存在某种关联，参见 Goodman（1996年，公式（46））。对具有行边际和列边际（p_{i+}）的概率表 $P = (p_{ij})$ 进行多重因子对应分析时，

分别为（p_{+j}）和（p_{+j}）时，对与 P 相关的矩阵 Q $=$（$q_{ij} = a_i p_{ij} b_j$）分两步进行对应分析：

步骤 1：根据对数交互指数的尺度不变性，见引理1a，在假设 $n_{ij} > 0$ 下，存在唯一的概率矩阵 Q $=(q_{ij})$，它通过严格正值尺度（a_i, b_j）与 P $= (p_{ij})$ 相关联，并保持第 i 行与第 j 列之间的尤尔关联不变；也就是说，

$$\lambda(p_{ij}, w_i^R = 1/I, w_j^C = 1/J) = \lambda(q_{ij} = a_i p_{ij} b_j, w_i^R = q_{i+} = 1/I, w_j^C = q_{+j} = 1/J$$

No注意，Q 具有与（$p_{ij}^{(\alpha)}$）类似的均匀行和列边际权重。

步骤 2：P 的多重因子对应分析是 Q 的对应分析表示

$$IJq_{ij} - 1 = \sum_{\alpha=1}^{k} f_\alpha(i) g_\alpha(j) / \delta_\alpha.$$

更多详情请参见 Choulakian（2022）。

4 含零值单元格的幂变换数据集的对应分析

大型组成数据集或列联表中出现零值单元格的情况相当常见。对于这种情况，令 m 为零值单元格的数量；因此，$IJ-m$ 是表中严格正值单元格的数量。我们看到，对于这类表，备注 b 中讨论的对数变换并不适用，而简单的幂变换 n_{ij}^{α} 对于 $\alpha \in [0, 1]$ 则适用。也就是说，矩阵 $(lim\ n_{ij}^{\alpha})$ 随着 $\alpha \to 0$收敛到一个指示矩阵 Z $=(z_{ij})$，其中 $z_{ij} = 1$，如果 $n_{ij} > 0$，而 $z_{ij} = 0$，如果 $n_{ij} = 0.$。将这种情况与备注 a) 中在假设 $n_{ij} > 0$ 下的观察结果进行比较颇有启发性，此时指示矩阵变为秩为 1 的 Z $=$ ($z_{ij} = 1$)。在此，值得考虑两种不同的情况：稀疏表中至少有两行或两列非成比例，且其中至少有两处零值；以及恰有一列（或一行）包含至少一个零值单元格的表。在后一种情况下，零项会产生主导性影响，这与 Benz´ecri 在对应分析中讨论的重型单元格和重型列（或行）的情况，以及 Choulakian（2008）在出租车对应分析中讨论的情况非常相似。

4.1 含有 m 个零值单元格的列联表

这里我们讨论指示矩阵 Z 的一列中有 m 个 $m=1$、……、$I-1$ 个零值单元格的情况，该矩阵的大小为 $I \times J.$。根据 Benz´ecri 的分布等价性原则，在对应分析（或出租车对应分析）中，成比例的行或列可以合并；因此，对 Z 进行对应分析（或出租车对应分析），与对 2×2 列联表进行对应分析（或出租车对应分析）完全相同

$$\mathbf{R} = \begin{bmatrix} 0 & m(J-1) \\ (I-m) & (I-m)(J-1) \end{bmatrix}, \tag{8}$$

它只有一个对应分析（或出租车对应分析）主维度。

引理 2
a) 对于 $\mathbf{R}$ 的对应分析，离散度称为惯量，$\rho_1^2 = \frac{m}{IJ}$.b) 对于 $\mathbf{R}$ 的出租车对应分析，出租车离散度 $\delta_1 = {}^{4m}_{(J-1)(I-m)} \quad \frac{}{(IJ-m)^2}$.

证明 a)：根据 Goodman（1996年）中的方程 2.1.4，对于一个 2×2 列联表，基于相关系数的“非独立性”为

$$\rho_1^2 = \frac{(p_{11}p_{22} - p_{12}p_{21})^2}{p_{1+}p_{2+}p_{+1}p_{+2}}. \tag{9}$$

将（9）中 R 的各元素的概率值代入后，得到

$$\begin{aligned} \rho_1^2 &= \frac{m^2(J-1)^2(I-m)^2}{(I-m)m(J-1)(J-1)I(I-m)J} \\ &= \frac{m}{IJ}. \end{aligned}$$

b）的证明：我们需要计算 R 的 $\sigma_{ij} = p_{ij} - p_{i+}p_{+j}$ $i=1$、2 和 $j=1$、2 的交叉协方差值

$$\begin{aligned} \mathbf{\Sigma} &= \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \\ &= \begin{bmatrix} \sigma_{11} & -\sigma_{11} \\ -\sigma_{11} & \sigma_{11} \end{bmatrix}, \end{aligned} \tag{10}$$

因为（$\sigma_{ij} = p_{ij} - p_{i+}p_{+j}$）已进行行和列中心化。

存在一个主轴 $\mathbf{u}_1' = (1 \ -1)$。因此，

$$\begin{aligned}\delta_1 &= ||\mathbf{\Sigma u}_1||_1 \\ &= 4|\sigma_{11}| \\ &= 4\frac{m(J-1)(I-m)}{(IJ-m)^2}.\end{aligned}$$

4.2 稀疏列联表

设 $\mathbf{Z} = (z_{ij})$ 是一个关联矩阵，即一个存在-缺失数据集，其中 $z_{ij} = 0$ 表示第 i 个个体中不存在水平 j， $z_{ij} = 1$ 表示第 i 个个体中存在水平 j。对应分析（或出租车对应分析）是分析此类列联表的一种常用方法，例如参见 Choulakian 和 Abou-Samra（2020年）。令 $p_{ij} = z_{ij}/\sum_{i,\ j} z_{ij}$，并假设边际 $p_{+j} > 0$ 和 $p_{i+} > 0$，则对应分析（或出租车对应分析）的数据重构公式为

$$p_{ij} = p_{+j}p_{i+}(1 + \sum_{\alpha=1}^{k} f_\alpha(i)g_\alpha(j)/\delta_\alpha). \tag{11}$$

现在假设我们应用均匀加权对数比分析（或 TLRA）；观察到 $\log_2(z_{ij}+1) = z_{ij}$，则数据重构公式变为

$$p_{ij} = p_{+j}/I + p_{i+}/J - 1/(IJ) + \sum_{\alpha=1}^{k} f_\alpha(i)g_\alpha(j)/\delta_\alpha; \tag{12}$$

这是一个熟悉的公式，称为方差分析因子分析（因子分析与方差分析），参见 Mandel（1971年）。

要在（11）和（12）之间选择“最佳”方案，我们在 Taxicab框架中使用残差符号质量指数（QSR），参见 Choulakian（2021年）和 Choulakian等人（2023年）。

5 示例

这里我们分析三个公开可用的列联表，并提供汇总结果。在计算中，我们使用 R 软件包 ca 和 TaxicabCA。

5.1 示例 1：作者数据集

Greenacre 和 Lewi（2009年）讨论了包含一个计数为零的作者列联表。该数据集的大小为 12×26，收录于 Greenacre 等人（2022年）在 R 软件包中提供的对应分析 ca 中。

我们考虑 $\mathbf{N}^{(\alpha)} = (n_{ij}^{\alpha})$ 的幂变换数据集，其中 $\alpha = 10\hat{}(-4)$；Greenacre（2022年）使用了 α 的这个取值。使用 R 软件包中的 ca，对 $\mathbf{N}^{(\alpha)}$ 得到的前两个离散度-惯量值为：

$$0.003204, \quad 0;$$

而根据引理 2a，通过大小为 2×2、包含 $m = 1$、$I = 12$ 和 $J = 26$ 的合并指示矩阵 R，得到第一个主惯量的值：$1/(12 * 26) = 0.003205$。

使用 R 软件包中的 TaxicabCA，对 $\mathbf{N}^{(\alpha)}$ 得到的前两个离散度值为：

$$0.011369, \quad 9e-06;$$

而根据引理 2b，通过大小为 2×2、包含 $m = 1$、$I = 12$ 和 $J = 26$ 的 R 矩阵，得到第一个主出租车离差的值：0.011373。

此示例展示了对于 $\alpha = 10\hat{}(-4)$，$\mathbf{N}^{(\alpha)} = (n_{ij}^{\alpha})$ 中一个零值单元格的显著影响。下一个示例将展示，在同一列中存在多个零值单元格时，也会得到类似的结果。

5.2 示例 2：RBGlass1 数据集

RBGlass1 是一个大小为 105×11 的组成数据集，其中变量 Sb 列有 $m = 26$ 个零值。该数据集收录于 Carlson 和 Roth （ 2022 年）的 archdata R 软件包中。

我们考虑 $\alpha = 10\hat{}(-4)$ 的幂变换数据集 $\mathbf{N}^{(\alpha)} = (n_{ij}^{\alpha})$。通过 R 软件包中的 ca 函数，$\mathbf{N}^{(\alpha)}$ 的前两个离散度-惯量值为：

0.0225，0;

然而，根据引理 2a，通过大小为 2×2 的合并指示矩阵 R，其中包含 $m = 26$、$I = 105$ 和 $J = 11$，我们得到第一个主惯量的值：$26/(11 * 105) = 0.022511$。

通过 R 软件包中的 TaxicabCA，$\mathbf{N}^{(\alpha)}$的前两个离散度值为：

$$0.0645, \quad 7e-06;$$

而根据引理 2b，通过大小为 2×2 的 R 矩阵，取 $m = 26$、 $I = 105$ 和 $J = 11$，可得第一个主出租车离差的值：0.0645。

5.3 示例 3：啮齿动物丰度数据

我们考虑 TaxicabCA R 软件包中的啮齿动物数据集，该数据集大小为 28 × 9。它是一个丰度数据集，记录了加利福尼亚州 28 个城市中 9 种大鼠的数量。Choulakian（2017）通过比较对应分析图和出租车对应分析图对其进行了分析；Choulakian（2021）指出，该数据集具有准二块对角结构；此外，Choulakian（2022）采用 Goodman 边际自由对应分析和边际自由出租车对应分析方法对其进行了分析。

令 N 为原始数据集，其大小为 28×9；N 中零计数的表观比例为 66.27%。R 语言 TaxicabCA 软件包中的函数 ”CombineCollinearRowsCols” 会合并 N 中成比例的行和列；我们看到 Nmerged 的大小为 21×9。因此，在对应分析框架下，N 中零计数的实际比例等于 Nmerged 中零计数的比例，即 58.73%。同理，指示矩阵 Z 的实际大小等于 Zmerged 的大小，即 14×9，其零值比例为 60.32%。

我们考察幂变换数据集 $\mathrm{N}^{(\alpha)} = (n_{ij}^{\alpha})$，其中 $\alpha = 10\hat{}(-4)$。对 $\mathrm{N}^{(\alpha)}$ 和 Zmerged 应用 R 软件包中的 ca，得到的图完全相同。同样，对 $\mathrm{N}^{(\alpha)}$ 和 Zmerged 应用 R 软件包中的 TaxicabCA，得到的图也完全相同。

可以通过运行附录中的 R代码查看这四幅图。

6 结论

我们尝试清晰简明地展示自 Grenacre 于 2009年 开始研究幂变换数据或 Box-Cox 变换数据的对应分析（CA）。我们区分了两类数据集：严格正值数据集和含有零值的数据集。具体而言，我们展示了当幂趋近于零时，零项对幂变换数据对应分析的主导影响；在这种情况下，幂变换数据集几乎变成一个 0-1 指示矩阵。

另一种方法是向任何零值单元格加上一个正数，然后按照公式（11）的推导使用对数变换；相关研究可参见 Lubbe 等人（2021年）和 Choulakian 等（2023年）。

Acknowledgements.

Choulakian 的研究得到了加拿大自然科学与工程研究委员会的支持。

Appendix 1: The R code

运行 R代码将生成论文中讨论的数值结果和图。R代码使用以下三个软件包。

a) Greenacre 等人（2022）开发的 ca软件包用于进行对应分析并生成对应分析图。

b) Allard 和 Choulakian（2019）开发的 TaxicabCA软件包用于进行出租车对应分析并生成出租车对应分析图。

c) Carlson 和 Roth（2022）开发的 archdata软件包用于 RBGlasse1数据集。

安装程序包 install.packages(c(”ca”, ”TaxicabCA”, ”archdata”)) 加载程序包 library(TaxicabCA) library(ca) library(archdata) #选择数据集 dataMatrix =as.matrix(rodent) dataMatrix <- t(author) data(RBGlass1) dataMatrix <- RBGlass1[, -1]dim(dataMatrix) #计算合并数据和指示矩阵 dataMerged <- CombineCollinearRowsCols(dataMatrix, rows = T, cols= T) dim(dataMerged) IndicatorM = 1-(dataMatrix == 0) 合并指示矩阵 <- CombineCollinearRowsCols(IndicatorM, rows = T, cols= T) dim(合并指示矩阵) #计算幂变换数据及其边际 alpha <- 10ˆ(-4) dataPowered <- dataMatrixˆalpha sum(dataPowered) apply(dataPowered, 2, function(x) sum(x))

apply(dataPowered, 1, function(x) sum(x)) #啮齿动物数据集的对应分析图 plot(ca(dataMatrix)) plot(ca(dataPowered)) plot(ca(IndicatorM)) # 出租车对应分析图 tca.Data <- tca(dataMatrix, nAxes=2,algorithm = "exhaustive") plot(tca.Data, axes = c(1,2),labels.rc = c(2, 2)) tca.Data <- tca(dataPowered, nAxes=2,algorithm ="exhaustive") plot(tca.Data, axes = c(1, 2),labels.rc = c(2, 2)) tca.Data <- tca(IndicatorM, nAxes=2,algorithm ="exhaustive") plot(tca.Data, axes = c(1,2),labels.rc = c(2, 2))

Appendix 2:

定理（Greenacre（2010年，结果 1）
在假设 $n_{ij} > 0$ 下,

$$\begin{aligned}\Delta_{ij}^{(\alpha)} &= \frac{p_{ij}^{(\alpha)} - p_{i+}^{(\alpha)} p_{+j}^{(\alpha)}}{p_{i+}^{(\alpha)} p_{+j}^{(\alpha)}} \\ &\approx \alpha\lambda(p_{ij}, w_i^R = 1/I, w_j^C = 1/J) \\ &= \alpha(\log p_{ij} + \frac{1}{IJ}\sum_{i,j}\log p_{ij} - \frac{1}{I}\sum_{i}\log p_{ij} - \frac{1}{J}\sum_{j}\log p_{ij})\end{aligned}$$

对于 $\alpha \to 0$ 和 $\alpha > 0$.

证明：根据麦克劳林-泰勒级数展开，在$\alpha = 0$的邻域内， $n_{ij}^{\alpha} =$ exp(α log n_{ij}) $= 1 + \alpha$ log $x + O(\alpha^2)$，其中 $r(\alpha) = O(h(\alpha))$

表示 $\lim_{\alpha\to 0}|\frac{r_{(\alpha)}}{h(\alpha)}| = constant > 0.$ 因此，对于 $i = 1, ..., I$ 以及 $j = 1, ..., J$

$$
\begin{aligned}
p_{ij}^{(\alpha)} &= \frac{n_{ij}^{\alpha}}{\sum_{i,j} n_{ij}^{\alpha}} \\
&= \frac{1 + \alpha \log n_{ij} + O(\alpha^2)}{IJ + \sum_{i,j} \alpha \log n_{ij} + O(\alpha^2)} \qquad \text{(A1)} \\
&= \frac{[1 + \alpha \log n_{ij} + O(\alpha^2)] \left[IJ + \alpha \sum_{i,j} \log n_{ij} + O(\alpha^2)\right]}{\left[IJ + \alpha \sum_{i,j} \log n_{ij} + O(\alpha^2)\right]^2} \\
&= \frac{IJ + \alpha IJ \log n_{ij} + \alpha \sum_{i,j} \log n_{ij} + O(\alpha^2)}{[IJ + O(\alpha)]^2}. \qquad \text{(A2)}
\end{aligned}
$$

由式（5a）可得

$$
\begin{aligned}
p_{i+}^{(\alpha)} &= \sum_j p_{ij}^{(\alpha)} \\
&= \frac{J + \alpha \sum_j \log n_{ij} + O(\alpha^2)}{IJ + O(\alpha)}; \qquad \text{(A3)} \\
p_{+j}^{(\alpha)} &= \frac{I + \alpha \sum_i \log n_{ij} + O(\alpha^2)}{IJ + O(\alpha)}; \qquad \text{(A4)} \\
p_{i+}^{(\alpha)} p_{+j}^{(\alpha)} &= \frac{IJ + \alpha J \sum_i \log n_{ij} + \alpha I \sum_j \log n_{ij} + O(\alpha^2)}{[IJ + O(\alpha)]^2}.
\end{aligned}
$$

由（A1）和（A4），可得

$$
\begin{aligned}
\Delta_{ij}^{(\alpha)} &= \frac{p_{ij}^{(\alpha)} - p_{i+}^{(\alpha)} p_{+j}^{(\alpha)}}{p_{i+}^{(\alpha)} p_{+j}^{(\alpha)}} \\
&= \frac{\left[IJ + \alpha IJ \log n_{ij} + \alpha \sum_{i,j} \log n_{ij} + O(\alpha^2)\right] - \left[IJ + \alpha J \sum_i \log n_{ij} + \alpha I \sum_j \log n_{ij} + O(\alpha^2)\right]}{IJ + \alpha J \sum_i \log n_{ij} + \alpha I \sum_j \log n_{ij} + O(\alpha^2)} \\
&= (\frac{\alpha IJ}{IJ}) \frac{\log n_{ij} + \frac{1}{IJ} \sum_{i,j} \log n_{ij} - \frac{1}{I} \sum_i \log n_{ij} - \frac{1}{J} \sum_j \log n_{ij} + O(\alpha)}{1 + \alpha \frac{1}{I} \sum_i \log n_{ij} + \alpha \frac{1}{J} \sum_j \log n_{ij} + O(\alpha^2)} \\
&= \alpha \frac{\lambda(p_{ij}, w_i^R = 1/I, w_j^C = 1/J) + O(\alpha)}{1 + O(\alpha)};
\end{aligned}
$$

根据引理 1，这就是所需的结果。

推论

根据 (A3) 和 (A4)，我们可知，对于 $i = 1, ..., I$， $p_{i+}^{(\alpha)} = 1/I + O(\alpha)$, 而对于 $j = 1, ..., J.$, $p_{+j}^{(\alpha)} = 1/J + O(\alpha)$

参考文献

Aitchison J (1986)《组成数据的统计分析》。伦敦：Chapman and Hall Allard J, Choulakian V (2019) R 中的 TaxicabCA 软件包 Beh E, Lombardo R (2014)《对应分析：理论、实践与新策略》。纽约：Wiley Beh E, Lombardo R (2022)《对应分析与 Cressie-Read 散度统计量》。国家应用统计研究所，伍伦贡大学，澳大利亚，工作论文 06-22 Benz´ecri JP (1973)《数据分析：第 2 卷：对应分析》。巴黎：Dunod Carlson D.L, Roth G (2022) R 中的 archdata 软件包 Choulakian V (2006) 出租车对应分析。《心理测量学》，71，333-345 Choulakian V (2008) 具有一个重型列的列联表的出租车对应分析。《心理测量学》，73(2)，309-319 Choulakian V (2016) 基于诱导范数的矩阵分解。《统计学、优化与信息计算》，4，1-14 Choulakian V (2017) 稀疏列联表的出租车对应分析。《意大利应用统计学杂志》，29 (2-3)，153-179 Choulakian V (2021) 对应分析和出租车对应分析中主维度内在质量的量化。可在 arXiv 上查阅：2108.10685 Choulakian V (2022) 关于 Goodman 的边际自由对应分析的一些注记。https ://arxiv.org/pdf/2202.01620.pdf Choulakian V, Allard J, Mahdi S (2023) 出租车对应分析与出租车对数比分析：对列联表和组成数据的比较。即将在《奥地利统计学杂志》发表。Cuadras CM, Cuadras D, Greenacre M (2006) 表示分类数据的不同方法比较。《统计学通讯：模拟与计算》，35(2)，447-459 Cuadras CM, Cuadras D (2015) 列联表多元分析的统一方法。《开放统计学杂志》，5，223-232

Egozcue JJ, Pawlowsky-Glahn V (2016) 改变单纯形中的参考测度及其加权效应。《奥地利统计学杂志》，45(4)，25-44 Egozcue JJ, Pawlowsky-Glahn V, Templ M, Hron K (2015) 使用单纯形几何研究列联表中的独立性。《统计学通讯：理论与方法》，44 :18，3978-3996 Goodman LA (1979) 对具有有序类别的交叉分类中关联进行分析的简单模型。《美国统计学会会刊》，74,537-55 Goodman LA (1981a) 具有有序类别的列联表的关联模型与二元正态分布。《生物统计学》，68，347-355 Goodman LA (1981b) 对具有有序类别的交叉分类进行分析的关联模型与典型相关。《美国统计学会会刊》，76，320-334 Goodman, LA (1991) 交叉分类数据分析中的度量、模型与图形展示。《美国统计学会会刊》，86 (4)，1085-1111 Goodman LA (1996) 一种用于分析交叉分类数据的通用方法：协调并综合 Pearson、Yule 和 Fisher 的若干方法，以及对应分析和关联分析的若干方法。《美国统计学会会刊》，91，408-428 Greenacre M (2009) 对应分析中的幂变换。《计算统计学与数据分析》，53(8)，3107-3116 Greenacre M (2010) 对数比分析是对应分析的极限情形。《数学地球科学》，42，129-134 Greenacre M (2011) 衡量子组成不一致性。《数学地球科学》，43，681–93 Greenacre M (2022) 卡方标准化结合 Box-Cox 变换，是组成数据分析中对数比变换的一种有效替代方法。可见：https ://arxiv.org/abs/2211.06755 Greenacre M, Lewi P (2009) 组成数据、列联表和比率尺度测量分析中的分布等价性与子组成一致性。《分类学杂志》，26，29-54 Greenacre M, Nenadic O, Friendly M (2022) R 中的 ca 软件包 Hotelling H (1936) 两组变量之间的关系。《生物统计学》28，321-377

Lancaster HO（1958年）。双变量分布的结构。《数学统计年刊》29，719–736。Mandel J（1971年）。非加性数据的一种新方差分析模型。《技术计量学》13(1)，1-18。Lubbe S、Filzmoser P、Templ M（2021年）。含大量零值的组成数据零值替换策略比较。《化学智能实验室系统》210，104248。Tenenhaus M、Augendre H（1996年）。Tucker组间电池因子分析与偏最小二乘典型相关分析。载于《第28届统计学日交流论文摘要集》，693-697。Tucker LR（1958年）。一种组间电池因子分析法。《Psychometrika》23，111-136。Tukey JW（1977年）。《探索性数据分析》。Addison-Wesley：马萨诸塞州雷丁。Yule GU（1912年）。关于测量两个属性之间关联的方法。《皇家统计学会杂志》75，579-642。