

# IBM SPSS Categories 21

Jacqueline J. Meulman

Willem J. Heiser



注意：使用本信息及其支持的产品之前，请阅读注意事项第 279 页码下的一般信息。

此版本适用于 IBM® SPSS® Statistics 21 及所有后续发布和修订，除非在新版本中另有说明。

Adobe 产品屏幕截图重印已获得 Adobe Systems Incorporated 的许可。

Microsoft 产品屏幕截图重印已获得 Microsoft Corporation 的许可。

受许可保护材料 – IBM 所有

**Copyright IBM Corporation 1989, 2012.**

美国政府用户受限权利 – 使用、复制或披露受与 IBM Corp. 签订的 GSA ADP Schedule Contract 的限制。

---

# 前言

IBM® SPSS® Statistics 是一种用于分析数据的综合系统。类别 可选附加模块提供本手册中描述的其他分析方法。此 类别 附加模块必须与 SPSS Statistics Core 系统一起使用，并已完全集成到了该系统中。

## 关于 IBM Business Analytics

IBM Business Analytics 软件提供决策者赖以提高业务绩效的完整、一致和准确的信息。包括**业务智能**、**预测分析**、**财务状况和战略管理**以及**分析应用程序**在内的一整套产品组合让您即刻、清楚地了解当前绩效并依此采取行动，以及能够预测未来的成果。结合丰富的行业解决方案、被证明的实践经验 and 专业的服务，无论公司规模大小，都能促使其获得最高的产能、自信自觉地做出决定并得到更好的成绩。

作为产品组合的一部分，IBM SPSS Predictive Analytics 软件帮助公司预测未来实践并采取积极行动，促使其获得更好的业务成果。全世界的商业政府和学术客户依赖 IBM SPSS 技术，因其具有竞争力的优势，能够吸引、留住和发展客户，同时减少欺诈和减轻风险。通过将 IBM SPSS 软件融入日常运营中，公司成为具有预测性的企业，能够引导和自觉做出决策，以满足业务目标，实现可观的竞争优势。欲知更多信息或联系代表，请访问 <http://www.ibm.com/spss>。

## 技术支持

我们提供有“技术支持”以维护客户。客户可就 IBM Corp. 产品使用或某一受支持硬件环境的安装帮助寻求技术支持。如要联系技术支持，请访问 IBM Corp. 网站，网址为 <http://www.ibm.com/support>。在请求协助时，请准备好您和您组织的 ID 以及支持协议。

## 针对学生的技术支持

如果您是使用任何学生版 IBM SPSS 软件产品的学生，请访问我们专为学生提供的在线**教育解决方案** (<http://www.ibm.com/spss/rd/students/>) 页面。如果您是使用大学提供的 IBM SPSS 软件副本的学生，请联系所在大学的 IBM SPSS 产品协调员。

## 客户服务

如果对发货或帐户存在任何问题，请联系您当地的办事处。请先准备好您的序列号以供识别。

## 培训讲座

IBM Corp. 提供公开的以及现场的培训讲座。所有讲座都是以实践小组为特色的。讲座将定期在各大城市开展。有关这些讲座的更多信息，请前往 <http://www.ibm.com/software/analytics/spss/training>。

## 致谢

最优尺度过程及其 IBM® SPSS® Statistics 实现由数据理论尺度系统小组 (DTSS) 开发, 该小组由莱顿大学社会与行为科学学院教育与心理学系的人员组成。

Willem Heiser、Jacqueline Meulman、Gerda van den Berg 和 Patrick Groenen 参与了最初的 1990 个过程的开发。Jacqueline Meulman 和 Peter Neufeglise 参与了分类回归、对应分析、分类主成份分析和多维尺度过程的开发。另外, Anita van der Kooij 对 CATREG、CORRESPONDENCE 和 CATPCA 做出了突出贡献。Willem Heiser、Jacques Commandeur、Frank Busing、Gerda van den Berg 和 Patrick Groenen 参与了 PROXSCAL 过程的开发。Frank Busing、Willem Heiser、Patrick Groenen 和 Peter Neufeglise 参与了 PREFSCAL 过程的开发。

## 部分 I: 用户指南

<b>1</b>	<b>分类数据的最优尺度过程介绍</b>	<b>1</b>
	什么是最优尺度? . . . . .	1
	为什么要使用最优尺度? . . . . .	1
	最佳度量水平和测量级别 . . . . .	2
	选择最佳度量水平 . . . . .	2
	转换图 . . . . .	3
	类别代码 . . . . .	3
	哪种过程最适合于您的应用? . . . . .	5
	分类回归 . . . . .	5
	分类主成份分析 . . . . .	6
	非线性典型相关性分析 . . . . .	6
	对应分析 . . . . .	7
	多重对应分析 . . . . .	8
	多维尺度. . . . .	8
	多维展开. . . . .	9
	最优尺度图表中的宽高比 . . . . .	9
	推荐参考 . . . . .	9
<b>2</b>	<b>分类回归 (CATREG)</b>	<b>12</b>
	在”分类回归”中定义尺度 . . . . .	13
	分类回归: 离散 . . . . .	14
	分类回归: 缺失值. . . . .	15
	分类回归: 选项 . . . . .	16
	分类回归规则化 . . . . .	18
	分类回归: 输出 . . . . .	19
	分类回归: 保存 . . . . .	20
	分类回归: 转换图. . . . .	21
	CATREG 命令附加功能 . . . . .	22
<b>3</b>	<b>分类主成份分析 (CATPCA)</b>	<b>23</b>
	在 CATPCA 中定义度量和权重 . . . . .	25

分类主成份分析：离散化 . . . . .	26
分类主成份分析：缺失值 . . . . .	26
分类主成份分析：选项 . . . . .	27
分类主成份分析：输出 . . . . .	29
分类主成份分析：保存 . . . . .	30
分类主成份分析：对象图 . . . . .	31
分类主成份分析：类别图 . . . . .	32
分类主成份分析：载荷图 . . . . .	33
CATPCA 命令附加功能 . . . . .	34
<b>4 非线性典型相关性分析 (OVERALS)</b>	<b>35</b>
定义范围和尺度 . . . . .	37
定义范围 . . . . .	38
非线性典型相关性分析：选项 . . . . .	38
OVERALS 命令附加功能 . . . . .	39
<b>5 对应分析</b>	<b>40</b>
在“对应分析”中定义行范围 . . . . .	41
在“对应分析”中定义列范围 . . . . .	42
对应分析：模型 . . . . .	42
对应分析：统计量 . . . . .	44
对应分析：图 . . . . .	45
CORRESPONDENCE 命令附加功能 . . . . .	46
<b>6 多重对应分析</b>	<b>47</b>
在“多重对应分析”中定义变量权重 . . . . .	49
多重对应分析：离散化 . . . . .	49
多重对应分析：缺失值 . . . . .	50
多重对应分析：选项 . . . . .	51
多重对应分析：输出 . . . . .	52
多重对应分析：保存 . . . . .	53
多重对应分析：对象图 . . . . .	54

多重对应分析：变量图 . . . . .	55
MULTIPLE CORRESPONDENCE 命令附加功能 . . . . .	56
<b>7 多维尺度 (PROXSCAL)</b>	<b>57</b>
矩阵中的跨列近似值 . . . . .	59
多列中的近似值 . . . . .	59
单列中的近似值 . . . . .	60
从数据中创建近似值 . . . . .	61
从数据中创建度量 . . . . .	62
定义多维尺度模型 . . . . .	63
多维尺度：约束 . . . . .	64
多维尺度：选项 . . . . .	65
多维尺度：图，版本 1 . . . . .	66
多维尺度：图，版本 2 . . . . .	67
多维尺度：输出 . . . . .	67
PROXSCAL 命令附加功能 . . . . .	69
<b>8 多维展开 (PREFSCAL)</b>	<b>70</b>
定义多维展开模型 . . . . .	71
多维展开：约束 . . . . .	72
多维展开：选项 . . . . .	73
多维展开：图 . . . . .	75
多维展开：输出 . . . . .	77
PREFSCAL 命令附加功能 . . . . .	78
<b>部分 II：示例</b>	
<b>9 分类回归</b>	<b>80</b>
示例：地毯清洁用品数据 . . . . .	80
标准线性回归分析 . . . . .	80
分类回归分析 . . . . .	86

示例:臭氧数据 . . . . .	97
离散化变量 . . . . .	98
转换类型选择 . . . . .	98
定量最优化 . . . . .	109
转换效果 . . . . .	111
推荐参考 . . . . .	120
<b>10 分类主成份分析</b>	<b>122</b>
示例:检验社会体系的相关性 . . . . .	122
运行分析 . . . . .	123
维数 . . . . .	127
定量 . . . . .	128
对象得分 . . . . .	129
成份载荷 . . . . .	131
附加维度 . . . . .	132
示例:进食障碍症 . . . . .	134
运行分析 . . . . .	135
转换图 . . . . .	147
模型摘要 . . . . .	149
成份载荷 . . . . .	149
对象得分 . . . . .	151
检查病程结构 . . . . .	153
推荐参考 . . . . .	168
<b>11 非线性典型相关性分析</b>	<b>171</b>
示例:调查结果的分析 . . . . .	171
检查数据 . . . . .	172
解释集之间的相似性 . . . . .	178
成份载荷 . . . . .	181
转换图 . . . . .	182
单一类别与多类别坐标 . . . . .	185
质心和投影质心 . . . . .	185
备用分析 . . . . .	188
一般建议 . . . . .	193
推荐参考 . . . . .	193

## 12 对应分析 195

标准化 . . . . .	195
示例:咖啡品牌的认知度 . . . . .	196
运行分析. . . . .	196
维数 . . . . .	200
贡献率 . . . . .	201
图 . . . . .	202
对称标准化. . . . .	203
推荐参考 . . . . .	205

## 13 多重对应分析 206

示例:五金件的特征 . . . . .	206
运行分析. . . . .	206
模型摘要. . . . .	209
对象得分. . . . .	210
区分测量. . . . .	211
类别定量. . . . .	212
更详细地考察对象得分 . . . . .	213
忽略离群值. . . . .	215
推荐参考 . . . . .	219

## 14 多维尺度 221

示例:亲属关系项的检验 . . . . .	221
选择维数. . . . .	221
三维解 . . . . .	228
使用非缺省转换的三维解. . . . .	234
讨论 . . . . .	238
推荐参考 . . . . .	238

## 15 多维展开 240

示例:早餐食品喜好程度 . . . . .	240
生成退化解. . . . .	240
度量 . . . . .	242
公共空间. . . . .	243

运行非退化分析 . . . . .	244
度量 . . . . .	245
公共空间. . . . .	246
示例：早餐食品喜好程度的三维展开 . . . . .	246
运行分析. . . . .	247
度量 . . . . .	251
公共空间. . . . .	252
私有空间. . . . .	253
使用不同的初始配置 . . . . .	255
度量 . . . . .	257
公共空间. . . . .	258
私有空间. . . . .	258
示例：检查行为情况的得体性 . . . . .	260
运行分析. . . . .	260
度量 . . . . .	265
公共空间. . . . .	266
近似值转换. . . . .	267
更改近似值转换（有序）. . . . .	267
度量 . . . . .	269
公共空间. . . . .	270
近似值转换. . . . .	270
推荐参考 . . . . .	271

## 附录

<b>A 样本文件</b>	<b>272</b>
<b>B 注意事项</b>	<b>279</b>
<b>参考书目</b>	<b>281</b>
<b>索引</b>	<b>287</b>

# 部分 I: 用户指南



# 分类数据的最优尺度过程介绍

类别过程使用最优尺度来分析标准统计过程难以分析或无法分析的数据。本章描述每种过程所执行的工作、每种过程最适合的情形、过程之间的关系，以及这些过程与其对应的标准统计过程的关系。

注意：这些过程及其 IBM® SPSS® Statistics 实现是由 Data Theory Scaling System Group (DTSS) 开发的，该小组由莱顿大学 (Leiden University) 社会和行为科学分院的教育和心理学系的成员组成。

## 什么是最优尺度？

最优尺度的基本理念是将数值定量分配给每个变量的类别，从而允许将标准过程用于获取已量化变量的解。

最优尺度值根据所用过程的优化标准分配给每个变量的类别。与分析中的名义变量或有序变量的原始标签不同，这些尺度值具有度量属性。

在大多数类别过程中，每个尺度化变量的优化定量都是通过称为**交替最小平方**的迭代方法获取的，在这种方法中，在将当前定量用于找到一个解之后，该定量会使用该解进行更新。更新后的定量接着用于查找新解，新解再用于更新定量，依此类推，直至达到通知过程停止的某个标准。

## 为什么要使用最优尺度？

分类数据通常可以在市场研究、民意调查研究以及社会和行为科学研究中找到。事实上，很多研究人员几乎专门处理分类数据。

虽然大多数标准模型的改编版本都专用于分析分类数据，但是通常，对于具有以下特点的数据集，它们的效果并不好：

- 观察值过少
- 变量过多
- 每个变量的值过多

通过量化类别，最优尺度方法避免了上述情况下的问题。此外，即使在专用方法适用的情况下，它们同样很有用。

对最优尺度输出的解释通常基于图形显示，而不基于解释参数估计值。最优尺度方法提供了极佳的探索性分析，它很好地补充了其他 IBM® SPSS® Statistics 模型。通过缩小调查重点的范围，利用最优尺度使数据可视化可构成以解释模型参数为中心的分析的基础。

## 最佳度量水平和测量级别

在首次使用类别过程时，这可能是非常令人困惑的概念。在指定级别时，您指定的不是变量被度量的级别，而是它们被调整的级别。其理念是，无论变量是如何度量的，要量化的变量都可能存在非线性关系。

对于类别而言，存在三种基本级别的测量：

- **标定级别**表示变量的值代表无序类别。可能为名义变量的示例有地区、邮政编码区域、宗教信仰以及多选类别。
- **有序级别**表示变量的值代表有序类别。示例包括代表满意度或信心以及首选评级得分的态度尺度。
- **数值级别**表示变量的值代表具有有意义度规的有序类别，这样类别之间适合比较距离。示例包括以年为单位的年龄和以千美元为单位的收入。

例如，假定像下表这样对地区、工作和年龄这几个变量进行编码。

表 1-1  
地区、工作和年龄的编码方案

地区编码	地区值	工作编码	工作值	年龄
1	北部	1	实习生	20
2	南部	2	销售代表	22
3	东部	3	经理人	25
4	西部			27

所示值代表了每个变量的类别。地区为名义变量。地区有四个类别，没有内在顺序。值 1 到 4 仅表示四种类别；编码方案完全是任意的。而工作可视为有序变量。原始类别构成了从实习生到经理人的晋级过程。较大的代码代表公司体制中职位较高的工作。但是，只有顺序信息是已知的，这意味着相邻类别之间谈不上有距离。相反，年龄可视为数值变量。对于年龄，值之间的距离是有内在意义的。20 与 22 之间的距离与 25 与 27 之间的距离相同，而 22 和 25 之间的距离大于前面的两个距离。

## 选择最佳度量水平

变量中没有任何内在属性自动预定义了您应为变量指定哪种最佳度量水平，理解这一点非常重要。您可以通过有意义的并且可使解释更简单的任何方法研究数据。例如，通过在有序级别分析数值级变量，使用非线性转换可能在更小的维数中得出解。

下面的两个示例演示了为什么“显而易见”的测量级别可能不是最佳的最佳度量水平。假设某个变量将对象分类为年龄组。虽然年龄可调整为数值变量，但是实际情况可能是对于 25 岁以下的人，安全性与年龄成正向关系，而对于 60 岁以上的人，安全性与年龄成负向关系。在此情况下，最好将年龄视为名义变量。

另一个示例是，根据政治偏好对人进行分类的变量似乎本质上是标定的。但是，如果按从政治左派到政治右派的顺序对政党进行排序，则可能要使用有序分析级别对政党进行量化以符合此顺序。

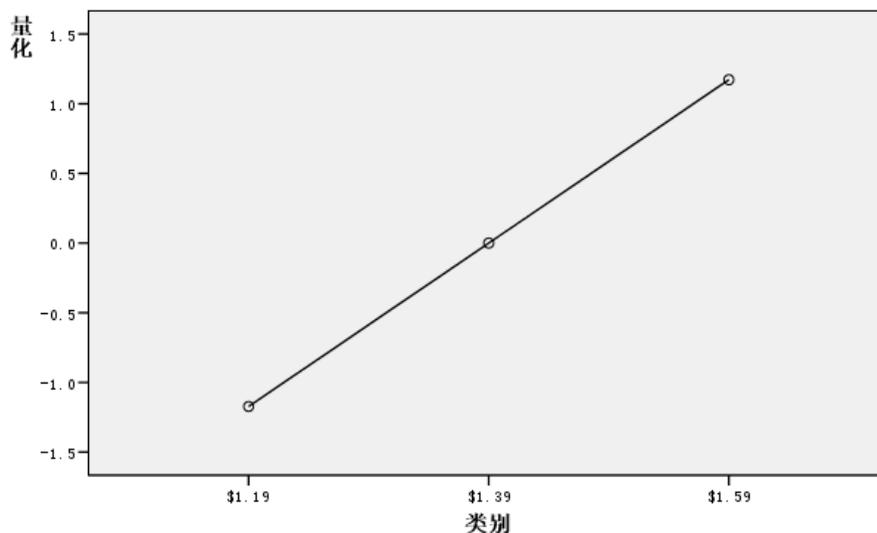
尽管变量没有预定义的属性可使其专属哪个级别，还是存在一些可以帮助新用户的通用准则。对于单名义量化，您通常不知道类别的顺序，但是您希望分析过程施加一个顺序。如果类别的顺序已知，则应该尝试有序量化。如果类别不可排序，则可以尝试多名义量化。

## 转换图

每个变量可调整到的不同级别对量化施加了不同的约束。变换图说明了从选定的最佳度量水平得出的量化和原始类别之间的关系。例如，当将变量视为数值时，将产生线性变换图。视为有序的变量将产生非递减变换图。视为标定的 U 形（或反转的 U 形）变量的变换图显示二次关系。通过完全改变类别的顺序，名义变量还可以生成没有明显趋势的变换图。下图显示了示例变换图。

变换图特别适合于确定选定的最佳度量水平的效果。如果多个类别得到了相似的量化，则确保可以将这些类别合并为一个类别。或者，如果视为标定的变量得到了显示递增趋势的量化，则有序转换可能生成相似的拟合。如果该趋势是线性的，则可能数值处理是合适的。但是，如果能确保合并类别或更改尺度级别，则分析不会有显著变化。

图片 1-1  
价格变换图（数值）



## 类别代码

当为分类变量进行编码时，应多加小心，因为一些编码方案可能生成不理想的输出或不完整的分析。下表中显示了工作的可行编码方案。

表 1-2  
工作的替代编码方案

类别	A	B	C	D
实习生	1	1	5	1
销售代表	2	2	6	5
经理人	3	7	7	3

某些类别过程要求定义每个所用变量的范围。此范围之外的任何值都视为缺失值。最小类别值总是 1。最大类别值由用户提供。此值不是变量类别的数目 — 而是最大类别值。例如，在表中，方案 A 的最大类别值为 3，方案 B 的最大类别值为 7，两种方案都对相同的三个类别进行编码。

变量范围确定了分析中将忽略哪些类别。分析中将忽略代码在定义范围之外的任何类别。这是忽略类别的简单方法，但是可能会产生不理想的分析结果。未正确定义的最大类别可能从分析中忽略有效类别。例如，对于方案 B，将最大类别值定义为 3 表示工作有编码为 1 到 3 的类别；经理人类别被视为缺失。因为实际上没有类别编码为 3，所以分析中的第三类别不包含个案。如果您本来就想忽略所有经理人类别，则此分析可能是合适的。但是，如果应该将经理人包括在内，则最大类别必须定义为 7，并且缺失值必须以大于 7 或小于 1 的值进行编码。

对于视为标定或名义的变量，类别范围不影响结果。对于名义变量，只有标签是重要的，而与该标签关联的值不重要。对于有序变量，通过量化保持类别的顺序；类别值本身不重要。生成相同类别排序的所有编码方案将得到相同的结果。例如，如果以有序级别分析工作，则表中前三个方案的作用是等效的。这三个方案的类别顺序是相同的。而方案 D 颠倒了第二类别和第三类别，将产生与其他方案不同的结果。

虽然一个变量的很多编码方案的作用相同，但是代码之间差值较小的方案为首选方案，因为代码影响过程产生的输出量。用 1 和用户定义的最大值之间的值编码的所有类别都是有效的。如果这些类别中有任何类别为空，则根据过程的不同，对应的定量将为系统缺失值或 0。虽然这两个赋值都不会影响分析，但是会为这些类别生成输出。因此，对于方案 B，工作有四个得到系统缺失值的类别。对于方案 C，也有四个得到系统缺失指示符的类别。相比之下，方案 A 没有系统缺失值定量。对视为标定或名义的变量使用连续整数作为代码将产生更少的输出，同时又不影响结果。

对视为数值的变量的编码方案比有序变量的编码方案更受限制。对于这些变量，连续的类别之间的差值非常重要。下表显示了年龄的三种编码方案。

表 1-3  
年龄的替代编码方案

类别	A	B	C
20	20	1	1
22	22	3	2
25	25	6	3
27	27	8	4

对数值变量的任何重新编码都必须保持类别之间的差值。使用原始值是确保保留差值的一种方法。但是，这可能产生很多有系统缺失指示符的类别。例如，方案 A 采用原始观察值。对于除“对应分析”之外的所有类别过程，最大类别值为 27，最小类别值设置为 1。前 19 个类别为空，并得到系统缺失指示符。如果最大类别远大于 1，则输出可能会很快变得相当繁琐，并且在 1 和最大值之间会有很多空类别。

为了减少输出的量，可以重新编码。但是，对于数值个案，不应使用“自动重新编码”工具。编码为连续整数将生成所有相邻类别之间的差值（值为 1），因此，所有定量都成为等间距的。在将变量视为数值时被视作重要的度规特征将因为重新编码为连续整数而受到破坏。例如，表中的方案 C 对应于自动重新编码年龄。类别 22 和 25 之间的差值从三更改为一，这种定量将影响后面的差值。

保持类别之间的差值的另一种重新编码方案是从每个类别减去最小类别值，然后在每个差上加 1。方案 B 出自于这种转换。从每个类别中减去最小类别值 20，然后对每个结果加 1。转换后的代码最小为 1，所有差值与原始数据的差值相同。最大类别值

现在为 8，第一个非零定量之前的零定量已全部排除。但是，从方案 B 得出的对应于每个类别的非零定量与从方案 A 得出的定量相同。

## 哪种过程最适合于您的应用？

这些过程中的四个过程（对应分析、多重对应分析、分类主成份分析和非线性典型相关性分析）所使用的方法属于称为**维度减少**的多变量数据分析的一般性领域。也就是说，变量间的关系尽可能以少量维数一如二或三一表示。这使您能够描述其原始丰富性和复杂性中难以理解的关系中的结构或模式。在市场研究应用中，这些方法可以是**知觉图**的形式。这些过程的主要优点是它们容纳具有不同最佳度量水平的数据。

类别回归描述分类响应变量和分类预测变量组合之间的关系。每个预测变量对响应变量的影响由相应的回归权重描述。与其他过程一样，可使用不同级别的最优尺度分析数据。

多维尺度和多维展开使用对象间的近似性描述低维空间中对象间的关系。

以下是每个过程的简要准则：

- 使用“类别回归”从分类自变量的组合中预测分类因变量的值。
- 使用“分类主成份分析”解释混合最佳度量水平的单个变量集中的变动模式。
- 使用“非线性典型相关性分析”评估两个或多个混合最佳度量水平的变量集的相关程度。
- 使用“对应分析”分析二阶列联表或可表示为二阶表的数据，例如品牌偏好或社会测量选择数据。
- 当您不想更强烈地假设以标定级别分析所有变量时，应使用“多重对应分析”分析分类多变量数据矩阵。
- 使用“多维尺度”分析近似性数据以查找低维空间中单个对象集的最小平方表示形式。
- 使用“多维展开”分析近似性数据以查找低维空间中两个对象集的最小平方表示形式。

## 分类回归

在分析目标为从一组自变量（预测变量）预测一个因变量（响应）的情况下，最适合使用类别回归。与所有最优尺度过程一样，尺度值分配给每个变量的每个类别，以使这些值对于回归是最优的。类别回归的解使已转换的响应与加权的已转换预测变量组合之间的平方相关性最大化。

**与其他类别过程的关系。**使用最优尺度的类别回归类似于具有两个集（其中一个集仅包含因变量）的最优尺度典型相关性分析。在后一种方法中，集的相似性是通过将每个集与位于所有集之间的某个未知变量进行比较而得出的。在类别回归中，已转换的响应与已转换的预测变量的线形组合的相似性是直接评定的。

**与标准方法的关系。**在标准线性回归中，分类变量或者可以重新编码为指示变量，或者以同样的方式视为区间级变量。在第一种方法中，该模型对于分类变量级别的每个组合都包含一个独立的截距和斜率。这产生了大量要解释的参数。在第二种方法中，对于每个变量只估计一个参数。但是，由于类别编码的任意性，不可能进行广义化。

如果某些变量不是连续的，则替代分析可用。如果响应是连续的并且预测变量是分类的，则通常采用方差分析。如果响应是分类的，而预测变量是连续的，则可能适用 Logistic 回归或判别分析。如果响应和预测变量都是分类的，则通常使用对数线性模型。

使用最优尺度的回归为每个变量提供三个尺度级别。这些级别的组合可解释任何单个“标准”方法都不适用的大量非线性关系。因此，最优尺度比具有最小附加复杂性的标准方法提供的灵活性更大。

此外，预测变量的非线性转换通常降低了预测变量之间的依赖性。如果将预测变量的相关矩阵的特征值与最优尺度化的预测变量的相关矩阵的特征值进行比较，则后者集合的可变性通常比前者更低。换言之，在类别回归中，最优尺度使预测变量相关矩阵的较大的特征值较小，而较小的特征值较大。

## 分类主成份分析

当您想解释混合最佳度量水平的单个变量集中的变化模式时，最适宜使用分类主成份分析。此项方法尝试在尽可能多地考虑到变化的同时减少变量集的维数。尺度值分配给每个变量的每个类别，以使这些值对于主成分解是最优的。分析中的对象根据量化的数据得到成分得分。成分得分图揭示了分析中对象间的模式，并且可以揭示数据中的异常对象。分类主成份分析的解使对象得分与指定的成分的数量（维数）的每个量化变量的相关性最大化。

分类主成份的一个重要应用是检查偏好数据，在此应用中，应答者根据其偏好对一定数量的项进行排序或评级。在常规 IBM® SPSS® Statistics 数据配置中，行为个人，列为项的测量值，行中的得分是偏好得分（例如，0 到 10 的范围），使数据依行而定。对于偏好倾向数据，可以将个人视为变量。使用转置过程可转置数据。级者成为变量，并且所有变量声明为有序。允许使用比 CATPCA 中的对象更多的变量。

**与其他类别过程的关系。**如果所有变量都声明为多名义，则分类主成份分析将生成等效于对相同变量运行的多重对应分析的分析。因此，分类主成份分析可视为多重对应分析的一种类型，在此类型中，某些变量声明为有序或数值。

**与标准方法的关系。**如果所有变量按数值级别调整，则分类主成份分析等效于标准主成分分析。

更广义地说，分类主成份分析是计算非数值尺度之间的相关性，并使用标准主成分或因子分析方法分析这些相关性的一种替代方法。仅将常规 Pearson 相关系数用作有序数据的关联性测量可能导致估计相关性时出现严重偏差。

## 非线性典型相关性分析

非线性典型相关性分析是具有多种不同应用的常规过程。非线性典型相关性分析的目标是分析两个或多个变量集之间的关系，而不似主成分分析那样，分析变量本身之间的关系。例如，您可能有两个变量集，其中一个变量集可能是一组应答者的人口统计学背景项，另一组变量集可能是对一组态度项的响应。分析中的尺度级别可以是名义、有序和数值的任何组合。最优尺度典型相关性分析通过同时将每个集中的典型变量与分配给对象的一组折算得分进行比较来确定集之间的相似性。

**与其他类别过程的关系。**如果有两个或多个变量集，并且每个集只有一个变量，则最优尺度典型相关性分析等效于最优尺度主成分分析。如果“每集单变量”分析中的所有变量为多名义，则最优尺度典型相关性分析等效于多重对应分析。如果有两个变量

集，其中一个变量集只包含一个变量，则最优尺度典型相关性分析等效于使用最优尺度的类别回归。

**与标准方法的关系。**标准典型相关性分析是一种统计方法，它查找相关性最大的一个变量集的线性组合和另一个变量集的线性组合。对于这组线性组合，典型相关性分析可查找后面的独立线性组合集（称为典型变量），直到最大数量等于较小集中的变量数。

如果分析中有两组变量集并且所有变量都定义为数值，则最优尺度典型相关性分析等效于标准典型相关性分析。虽然 IBM® SPSS® Statistics 没有典型相关性分析过程，但是可以从多变量方差分析中获得许多相关统计量。

最优尺度典型相关性分析有多种其他应用。如果有两个变量集，并且其中一个集包含声明为单名义的名义变量，则可以用类似于解释回归分析结果的方式解释最优尺度典型相关性分析结果。如果将变量视为多名义，则最优尺度分析是判别分析的替代方法。将变量分组成两个以上的集合将提供多种分析数据的方法。

## 对应分析

对应分析的目标是生成对应表的双标图。在对应表中，行和列变量假设为表示无序类别；因此，将始终使用标定最佳度量水平。检查这两种变量仅仅是为了获得其标定信息。也就是说，唯一需要考虑的是某些对象属于同一个类别而其他对象不属于。对于同一变量的类别之间的距离或顺序不作任何假设。

对应分析的一种特定用途是分析二阶列联表。如果表有  $r$  个活动行和  $c$  个活动列，则对应分析解中的维数最小为  $r$  减 1 或  $c$  减 1，以较小者为准。换言之，可在维度空间中完好地表示列联表的行类别或列类别。但是实际上，您希望在低维空间（如二维）中表示二阶表的行和列类别，因为二维图比多维空间表示形式更容易理解。

当使用的维数小于可能的最大维数时，分析中生成的统计量将描述行和列类别在低维表示形式中的表示效果。如果二维解表示形式的质量良好，则可以查看行点和列点图，以了解哪些行变量的类别相似，哪些列变量的类别相似，以及哪些行类别与哪些列类别相似。

**与其他类别过程的关系。**简单对应分析限于二阶表。如果有两个以上需要处理的变量，则可以组合变量以创建交互变量。例如，对于变量地区、工作和年龄，可组合地区和年龄以创建具有 12 个类别的新变量地区工作，如下表所示。此新变量与年龄一起构成了一个二阶表（12 行，4 列），可在对应分析中分析此表。

表 1-4  
地区和工作的组合

类别代码	类别定义	类别代码	类别定义
1	北部，实习生	7	东部，实习生
2	北部，销售代表	8	东部，销售代表
3	北部，经理人	9	东部，经理人
4	南部，实习生	10	西部，实习生
5	南部，销售代表	11	西部，销售代表
6	南部，经理人	12	西部，经理人

此方法的一个缺点是任意一对变量都可以组合。我们可组合工作和年龄，产生另一个 12 个类别的变量。或者，可以组合地区和年龄，生成一个新的 16 个类别的变量。这些交互变量的每一个都与其余的变量构成一个二阶表。对这三个表的对应分析不会生成相

同的结果，但每次分析都是有效的。此外，如果有四个或更多变量，则可以构建将一个交互变量与另一个交互变量进行比较的二阶表。要分析的可能存在的表的数量可以变得很大，即使只有几个表。可以选择其中一个表进行分析，或者可分析所有表。或者，可以使用多重对应分析过程同时检查所有变量，无需构建交互变量。

**与标准方法的关系。**在独立性作为分析中共同焦点的情况下，交叉表过程也可用于分析列联表。但是，即使在小型表中，检测背离独立性的原因可能也很困难。对应分析的作用在于显示任何规模的二阶表的此类模式。如果行和列变量之间存在关联（即如果卡方值很大），则对应分析可能有助于揭示关系的本质。

## 多重对应分析

多重对应分析尝试生成一个解，该解中同类别的对象绘制为紧靠在一起，而不同类别的对象绘制为远远分开。每个对象都尽可能地接近应用于该对象的类别的类别点。这样，类别可将对象分成数个均一子组。当变量将相同类别中的对象分类到同一子组中时，这些变量被视为是均一的。

对于单维解，多重对应分析将最优尺度值（类别定量）分配给每个变量的每个类别，以使总体上，平均而言，类别有最大分布。对于二维解决方案，多重对应分析查找与每个变量的类别的第一组定量无关的第二组定量，并再次尝试最大化分布，如此一直继续。因为变量的类别得到与维数一样多的得分，所以分析中的变量被假设为最佳度量水平的多名义。

多重对应分析还为分析中的对象分配得分，使类别量化为该类别中的对象的对象得分平均值，即质心。

**与其他类别过程的关系。**多重对应分析也称为齐性分析或双重尺度。当只有两个变量时，它提供了与对应分析类似但不相同的结果。对应分析生成了总结解的拟合与表示形式的质量的唯一输出，包括稳定性信息。因此，对应分析通常更适合于双变量情况下的多重对应分析。两种过程的另一区别是，多重对应分析的输入是数据矩阵，其中行为对象，列为变量；而对应分析的输入可以是相同的数据矩阵、常规相似性矩阵，或者联合列联表，后者是行和列都表示变量的类别的汇总矩阵。还可以将多重对应分析看作是在多名义级别调整的数据的主成分分析。

**与标准方法的关系。**可将多重对应分析视为多阶列联表的分析。多阶列联表也可使用交叉表过程进行分析，但是交叉表提供了每个控制变量的每个类别的单独的摘要统计量。使用多重对应分析时，通常可通过单个二维图总结所有变量之间的关系。多重对应分析的一种高级用法是用第一维的最优尺度值替换原始类别值，并执行二次多变量分析。由于多重对应分析用数值尺度值替换了类别标签，因此需要数值数据的很多不同过程可在多重对应分析之后应用。例如，因子分析过程生成等效于多重对应分析的第一维的主成分。第一维的成分得分等于对象得分，并且平方成份载荷等于区分测量。但是，第二个多重对应分析维度不等于因子分析的第二维。

## 多维尺度

当分析目标是查找单组对象或个案之间的一组距离测量中的结构时，最适合使用多维尺度。这是通过将观察值指定给概念低维空间中的特定位置实现的，这样使空间中的点之间的距离尽可能与给定的（不）相似性相匹配。结果是一个该低维度空间中的对象的最小平方表示形式，这在许多个案中将帮助您进一步理解您的数据。

**与其他类别过程的关系。**当您有用来创建距离的多变量数据，并且随后用多维尺度分析这些数据时，其结果类似于使用具有主要对象标准化的“分类主成份分析”分析数据。这类 PCA 也称为主坐标分析。

**与标准方法的关系。**类别多维尺度过程 (PROXSCAL) 提供了对 Statistics Base 选项中可用的尺度过程的多种改进 (ALSCAL)。PROXSCAL 提供了对某些模型的加速算法，可以对公共空间施加约束。此外，PROXSCAL 尝试使标准化初始应力而不是 S-stress (也称为**张力**) 达到最小。标准化初始应力通常是首选的，因为它是基于距离测量的，而 S-stress 是基于平方距离。

## 多维展开

当分析目标是查找两组对象（指行对象和列对象）之间的一组距离测量值中的结构时，最适合使用多维展开。这是通过将观察值指定给概念低维空间中的特定位置实现的，这样使空间中的点之间的距离尽可能与给定的（不）相似性相匹配。结果是一个该低维空间中的行和列对象的最小平方表示形式，这在许多情况下可帮助您进一步理解您的数据。

**与其他类别过程的关系。**如果数据由单组对象（正方形对称矩阵）之间的距离构成，则使用多维尺度。

**与标准方法的关系。**类别多维展开过程 (PREFSCAL) 提供了对 Statistics Base 选项中可用的展开功能的多种改进（通过 ALSCAL）。使用 PREFSCAL 可以对公共空间施加约束；此外，PREFSCAL 尝试使惩罚强度度量最小化，这有助于避免退化解（容易对其使用较老的算法）。

## 最优尺度图表中的宽高比

最优尺度图中的宽高比是等方性的。在二维图中，表示维度 1 中的一个单位的距离等于表示维度 2 中的一个单位的距离。如果更改二维图中某个维的范围，则系统将更改另一维的大小，以保持物理距离相等。对于优化尺度过程，不可覆盖等方宽高比。

## 推荐参考

有关最优尺度方法的一般性信息，请参见以下文章：

Barlow, R. E., D. J. Bartholomew, D. J. Bremner, 和 H. D. Brunk. 1972. *Statistical inference under order restrictions*. New York: John Wiley and Sons.

Benzécri, J. P. 1969. *Statistical analysis as a tool to make patterns emerge from data*. 输入: *Methodologies of Pattern Recognition*, S. Watanabe, ed. New York: Academic Press.

Bishop, Y. M., S. E. Feinberg, 和 P. W. Holland. 1975. *Discrete multivariate analysis: Theory and Practice*. Cambridge, Mass.: MIT Press.

De Leeuw, J. 1984. *The Gifi system of nonlinear multivariate analysis*. 输入: *Data Analysis and Informatics III*, E. Diday, et al., ed..

- De Leeuw, J. 1990. Multivariate analysis with optimal scaling. 输入: Progress in Multivariate Analysis, S. Das Gupta, 和 J. Sethuraman, eds. Calcutta: Indian Statistical Institute.
- De Leeuw, J., 和 J. Van Rijckevorsel. 1980. HOMALS and PRINCALS—Some generalizations of principal components analysis. 输入: Data Analysis and Informatics, E. Diday, et al., ed. Amsterdam: North-Holland.
- De Leeuw, J., F. W. Young, 和 Y. Takane. 1976. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, 41, .
- Gifi, A. 1990. Nonlinear multivariate analysis. Chichester: John Wiley and Sons.
- Heiser, W. J., 和 J. J. Meulman. 1995. Nonlinear methods for the analysis of homogeneity and heterogeneity. 输入: Recent Advances in Descriptive Multivariate Analysis, W. J. Krzanowski, ed. Oxford: Oxford University Press.
- Israëls, A. 1987. Eigenvalue techniques for qualitative data. Leiden: DSWO Press.
- Krzanowski, W. J., 和 F. H. C. Marriott. 1994. Multivariate analysis: Part I, distributions, ordination and inference. London: Edward Arnold.
- Lebart, L., A. Morineau, 和 K. M. Warwick. 1984. Multivariate descriptive statistical analysis. New York: John Wiley and Sons.
- 最大值, J. 1960. Quantizing for minimum distortion. *Proceedings IEEE (Information Theory)*, 6, .
- Meulman, J. J. 1986. A distance approach to nonlinear multivariate analysis. Leiden: DSWO Press.
- Meulman, J. J. 1992. The integration of multidimensional scaling and multivariate analysis with optimal transformations of the variables. *Psychometrika*, 57, .
- Nishisato, S. 1980. Analysis of categorical data: Dual scaling and its applications. Toronto: University of Toronto Press.
- Nishisato, S. 1994. Elements of dual scaling: An introduction to practical data analysis. Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc.
- Rao, C. R. 1973. Linear statistical inference and its applications, 2nd ed. New York: John Wiley and Sons.
- Rao, C. R. 1980. Matrix approximations and reduction of dimensionality in multivariate statistical analysis. 输入: Multivariate Analysis, Vol. 5, P. R. Krishnaiah, ed. Amsterdam: North-Holland.
- Roskam, E. E. 1968. Metric analysis of ordinal data in psychology. Voorschoten: VAM.

Shepard, R. N. 1966. Metric structures in ordinal data. *Journal of Mathematical Psychology*, 3, .

Wolter, K. M. 1985. *Introduction to variance estimation*. Berlin: Springer-Verlag.

Young, F. W. 1981. Quantitative analysis of qualitative data. *Psychometrika*, 46, .

# 分类回归 (CATREG)

分类回归通过为类别指定数值来量化分类数据，从而生成转换后变量的最优线性回归方程。分类回归也用缩写词 CATREG 来表示（代表 categorical regression）。

标准线性回归分析涉及使响应变量（因变量）和预测变量（自变量）的加权组合之间的平方差之和达到最小。变量通常是定量的，（名义）分类数据重新编码为二元变量或对比变量。因此，分类变量用于分离个案组，并且该技术估计每个组的独立的参数集。估计的系数反映了预测变量的变化对响应的影响程度。对于预测变量值的任何组合都可以预测响应。

另一种方法需要对分类预测变量值本身进行响应回归。这样，将为每个变量分别估计一个系数。但是，对于分类变量，类别值是任意的。以不同的方式编码类别将产生不同的系数，这样，在对同样的几个变量的分析进行比较时，难度就增大了。

CATREG 通过同时调整名义、序数和数值变量扩展了标准方法。该过程量化分类变量以使量化反映初始类别的特征。该过程以与处理数值变量相同的方式处理量化的分类变量。使用非线性转换允许在各种级别分析变量以查找最佳拟合模型。

**示例。**分类回归可用于描述工作满意度对工作类别、地理区域和旅行量的依赖程度。您可能会发现高满意度对应于经理和低旅行量。生成的回归方程可用于针对三个自变量的任何组合预测工作满意度。

**统计量和图。**频率、回归系数、ANOVA 表、迭代历史记录、类别量化、未转换的预测变量之间的相关性、转换后的预测变量之间的相关性、残差图和变换图。

**数据。**CATREG 在类别指示变量上运行。类别指示符应为正整数。可使用“离散化”对话框将小数值变量和字符串变量转换为正整数。

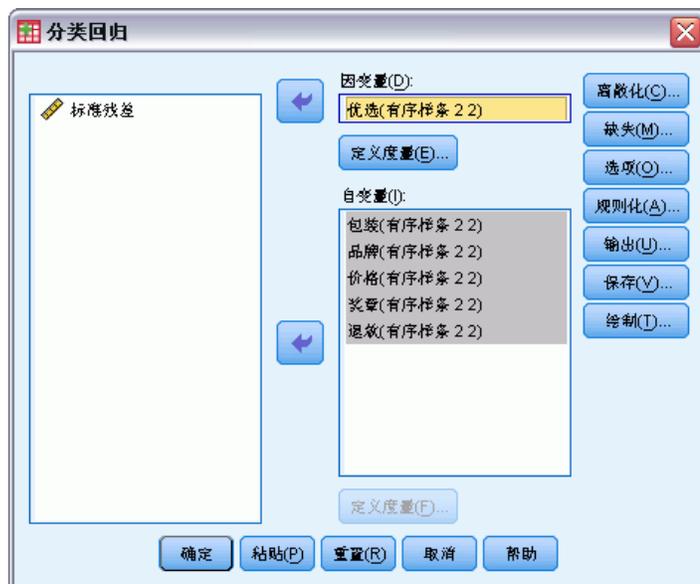
**假设。**只允许一个响应变量，但是预测变量的最大数目为 200。该数据必须至少包含三个有效个案，并且有效个案数必须大于预测变量数加一。

**相关过程。**CATREG 等效于使用最优尺度的分类典型相关性分析 (OVERALS)，该分析有两个变量集，其中一个只包含一个变量。将所有变量调整为数值级别对应于标准多重回归分析。

## 获取分类回归

- ▶ 从菜单中选择：  
分析 > 回归 > 最优尺度 (CATREG)...

图片 2-1  
“分类回归”对话框



► 选择因变量和自变量。

► 单击确定。

或者，更改每个变量的尺度级别。

## 在“分类回归”中定义尺度

可设置因变量和自变量的最佳度量水平。缺省情况下，它们调整为带两个内部结点的二度单调性样条（有序）。此外，还可以设置分析变量的权重。

图片 2-2  
“定义尺度”对话框



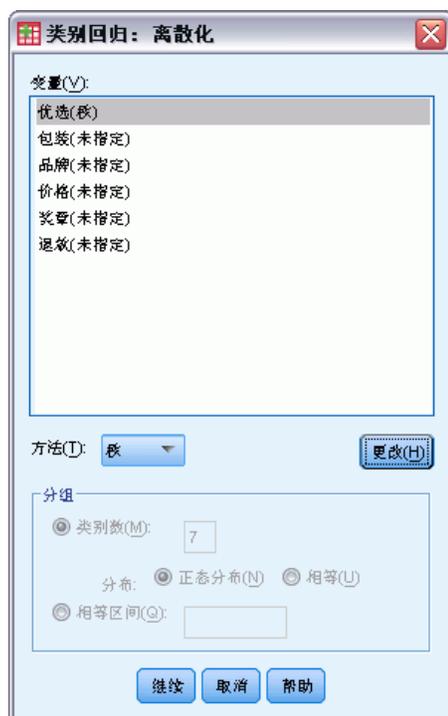
**最佳度量水平。**还可以选择用于量化每个变量的尺度级别。

- **有序样条。**已观察到的变量的类别顺序保存在最优尺度化变量中。类别点将位于一条通过原点的直线上（矢量）。生成的转换是一个选定次数的平滑单调的分段多项式。这些段是由内部结点的用户指定的数量和过程决定的位置来指定的。
- **名义样条。**已观察到的变量中保留在最优尺度化变量中的唯一信息是类别中对象的分组。已观察到的变量的类别的顺序不会保留。类别点将位于一条通过原点的直线上（矢量）。生成的转换是一个选定次数的平滑的、可能非单调的分段多项式。这些段是由内部结点的用户指定的数量和过程决定的位置来指定的。
- **有序。**已观察到的变量的类别顺序保存在最优尺度化变量中。类别点将位于一条通过原点的直线上（矢量）。生成的转换比有序样条转换拟合得好，但是平滑度较低。
- **名义。**已观察到的变量中保留在最优尺度化变量中的唯一信息是类别中对象的分组。已观察到的变量的类别的顺序不会保留。类别点将位于一条通过原点的直线上（矢量）。生成的转换比样条名义转换拟合得更好，但是平滑度较低。
- **数值。**类别被视为是有序的并且是等间距的（区间度）。已观察到的变量的类别顺序以及类别号之间的等间距保留在最优尺度化变量中。类别点将位于一条通过原点的直线上（矢量）。当所有变量都为数值级别时，分析与标准主成分分析类似。

## 分类回归： 离散

使用“离散化”对话框可以选择对变量重新编码的方法。除非另有指定，否则小数值变量将分组成具有近似正态分布的七个类别（如果变量的不同值的数目小于七，则将按此数目划分类别）。通过按照升序字母数值顺序分配类别指示符，字符串变量总是转换为正整数。字符串变量的离散化适用于这些整数。缺省情况下，其他变量保留原样。随后，分析中将使用离散化变量。

图片 2-3  
“离散化”对话框



**方法。**在分组、秩和乘之间选择。

- **分组。**重新编码为指定数量的类别或者按区间重新编码。
- **秩。**通过对个案排序来对变量进行离散化。
- **乘。**变量的当前值是乘以 10 且经过四舍五入的标准化值，并且加上了一个常数以使最低离散值为 1。

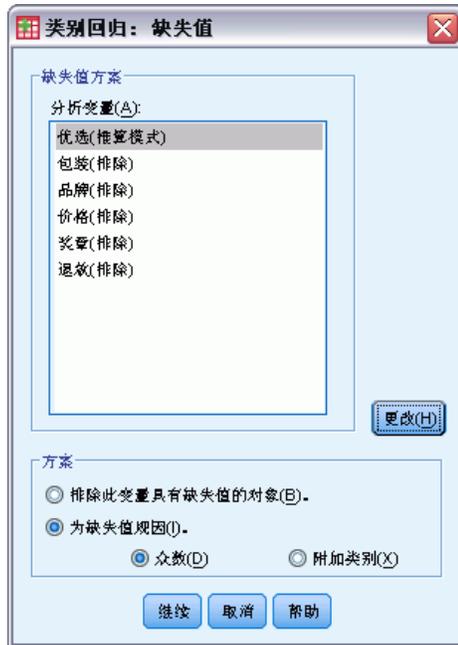
**分组。**在按分组离散化变量时，可以使用以下选项：

- **类别数。**指定一定数量的类别，并指定变量的值是否应该遵循在这些类别间的近似正态分布或均匀分布。
- **相等区间。**变量重新编码到由这些等大小的间隔所定义类别中。必须指定间隔的长度。

## 分类回归：缺失值

使用“缺失值”对话框可以选择处理分析变量和补充变量中的缺失值的方案。

图片 2-4  
“缺失值”对话框



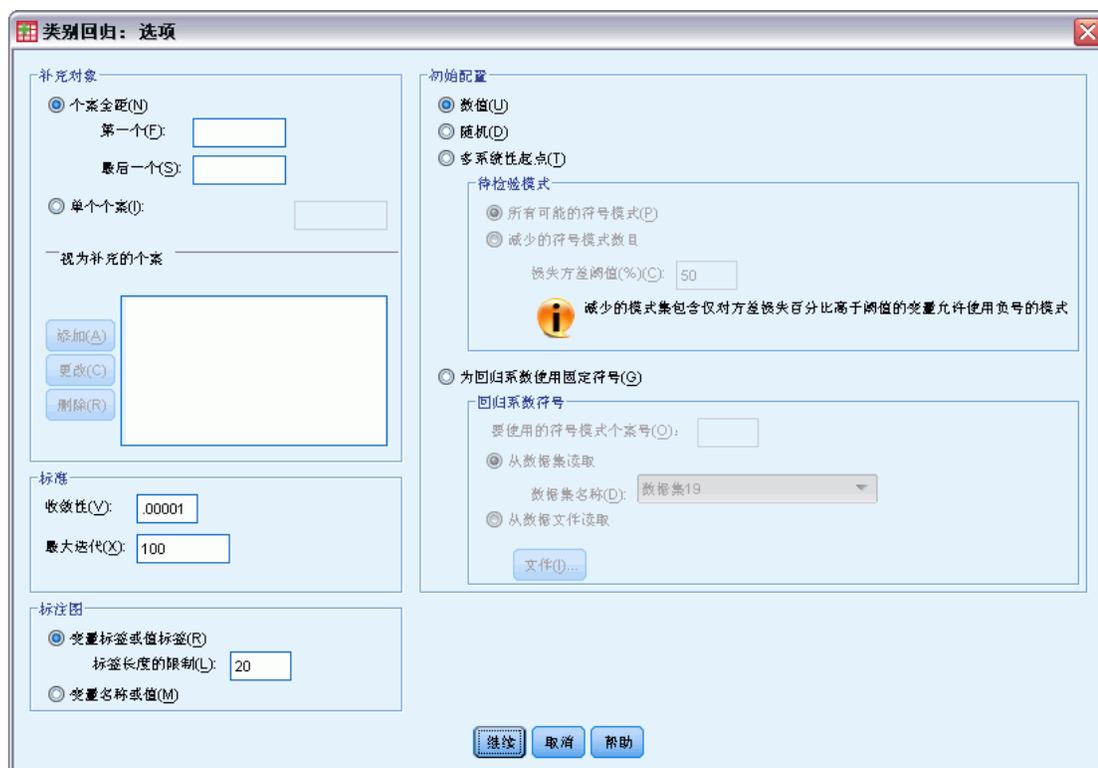
**方案。**选择排除带有缺失值的对象（列表删除）或插补缺失值（积极处理）。

- **排除此变量具有缺失值的对象。**选定的变量有缺失值的对象将从分析中排除。此方案不适用于补充变量。
- **插补缺失值。**选定的变量有缺失值的对象将对这些值归因。您可以选择插补方法。选择众数将缺失值替换为最频繁类别。当有多个众数时，将使用具有最小类别指示符的众数。选择附加类别将缺失值替换为相同的附加类别定量。这意味着此变量有缺失值的对象被视为属于同一（附加）类别。

## 分类回归：选项

“选项”对话框允许您选择初始配置样式，指定迭代和收敛条件，选择补充对象和设置绘图标记。

图片 2-5  
“选项”对话框



**补充对象。**使用此项可以指定您要视作补充对象的对象。只需键入补充对象的编号（或指定个案范围），然后单击添加。不可对补充变量加权（指定的权重将被忽略）。

**初始配置。**如果没有变量被视作名义变量，则选择数值配置。如果至少有一个变量被视为名义，则选择随机配置。

或者，如果至少一个变量具有有序或有序样条尺度级别，则通常的模拟拟合算法可能导致欠佳解。选择具有所有可能的待检验符号模式的多系统性起点，可以始终寻找最优解，但由于数据集中的有序和有序样条变量数量增加，因此所需的处理时间也快速增加。可以指定方差损失阈值百分比以减少检验模式数量，即阈值越高，排除的符号模式越多。采用此选项，尽管无法保证获得最优解，但也消除了得到欠佳解的可能性。此外，如果找不到最优解，欠佳解与最优解也不会差别太大。在要求多系统性起点时，每个起点的回归系数符号被写入外部 IBM® SPSS® Statistics 数据文件或当前会话中的数据文件。  
有关详细信息，请参阅第 20 页码分类回归：保存。

先前采用多系统性起点的运行结果允许您为回归系数使用固定符号。符号（采用 1 和 -1 指示）需要在指定数据集或文件的某行中。整数值的起始数字为包含待使用符号的文件对应该行的个案号。

**标准。**可以指定回归可在其计算中执行的最大迭代次数。还可以选择收敛标准值。如果上两次迭代之间的总拟合之差小于收敛值，或者达到了最大迭代次数，则回归停止迭代。

**标注图。**可用于指定在图中将使用变量和值标签还是变量名称和值。还可指定标签的最大长度。

## 分类回归规则化

图片 2-6  
“规则化”对话框

**方法。**规则化方法可以向 0 方向缩小回归系数估计，以降低其变异性，从而改善模型的预测误差。在要求规则化方法时，每个惩罚系数值的规则化模型和系数被写入外部 IBM® SPSS® Statistics 数据文件或当前会话中的数据集中。 [有关详细信息，请参阅第 20 页码分类回归：保存。](#)

- **Ridge 回归。** Ridge 回归引入惩罚项以缩小系数，惩罚项等于系数平方乘以惩罚系数的总和。该系数可从 0（无惩罚）到 1 变化；如果指定了范围与增量，过程将搜索“最佳”的惩罚值。
- **套索。**套索的惩罚项是基于绝对系数的总和，惩罚系数的指定与 Ridge 回归类似，但套索涉及更密集的计算。
- **弹性网络。**“弹性网络”简单地组合套索和 Ridge 回归惩罚，在指定的值网格中搜索以发现“最佳”的套索和 Ridge 回归惩罚系数。对于给定的套索与 Ridge 回归惩罚，“弹性网络”的计算量并不比套索多很多。

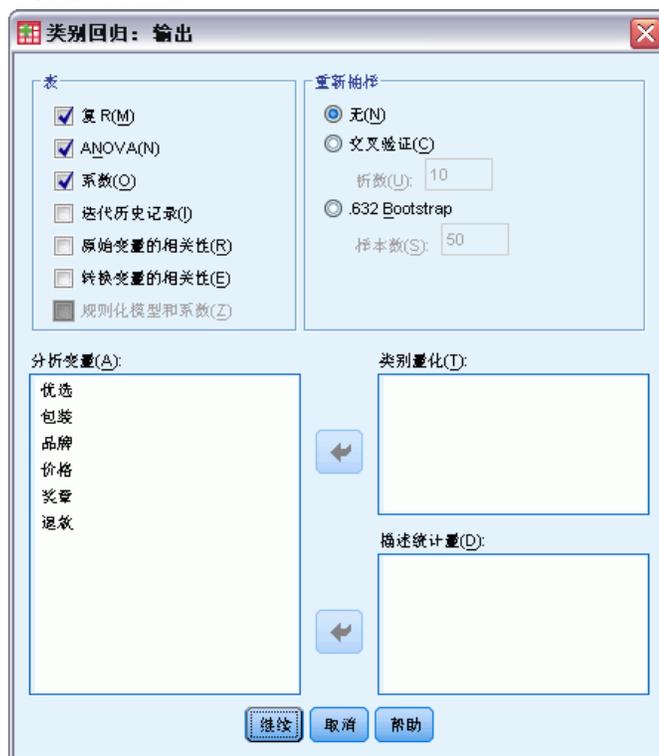
**显示规则化图。**这些是回归系数与规则化惩罚图。在搜索某个值范围以寻找“最佳”惩罚系数时，它提供了有关回归系数在该范围上如何变化的视图。

**弹性网络图。**对于“弹性网络”方法，由 Ridge 回归惩罚值产生单独的规则化图。所有可能图使用指定的最小和最大 Ridge 回归惩罚值所确定范围中的每个值。为部分 Ridge 惩罚允许您指定由最小和最大 Ridge 回归惩罚值所确定范围的值子集。只需键入惩罚值的编号（或指定值范围），然后单击添加。

## 分类回归： 输出

“输出”对话框允许您选择显示在输出中的统计。

图片 2-7  
“输出”对话框



**表。**为以下项目生成表：

- **复 R。**包括  $R^2$ 、调整后的  $R^2$  以及将最优尺度考虑在内的调整后的  $R^2$ 。
- **ANOVA。**此选项包括回归和残差平方和、均方和 F。将显示两个 ANOVA 表：一个表的回归自由度等于预测变量的数目，另一个表的回归自由度将最优尺度考虑在内。
- **系数。**此选项提供了三张表：包括 beta、beta 的标准误、t 值和显著性的“系数”表；含有将最优尺度自由度考虑在内的 beta 标准误的“系数最优尺度”表；以及含有零阶、部分和偏相关性、转换后预测变量的 Pratt 的相对重要性度量，以及转换前后容差的表。
- **迭代历史记录。**对于每次迭代，包括算法的起始值在内，将显示复 R 和回归误差。复 R 中的增量从第一次迭代开始列出。
- **初始变量的相关性。**显示未转换的变量之间的相关性的矩阵将被显示。

- **转换变量的相关性。**显示已转换的变量之间的相关性的矩阵将被显示。
- **规则化模型和系数**显示每个规则化模型的惩罚值、R 方和回归系数。如果指定了重新抽样方法，或指定了补充对象（检验个案），它还可显示预测误差或检验 MSE。

**重新抽样。**重新抽样方法为您提供有关模型预测误差的估计。

- **交叉验证。**交叉验证将样本分割为许多子样本，或群。然后，生成分类回归模型，并依次排除每个子样本中的数据。第一个模型基于第一个样本群的个案之外的所有个案，第二个模型基于第二个样本群的个案之外的所有个案，依此类推。对于每个模型，估计其预测误差的方法是将模型应用于生成它所排除的子样本。
- **.632 Bootstrap。**采用 Bootstrap 方法，可以通过放回方式从数据中随机抽取观察值，多次重复该过程以获得大量 Bootstrap 样本。为所有 Bootstrap 样本拟合模型，然后将该拟合模型所估计的模型预测误差应用到不 Bootstrap 样本中的个案。

**类别量化。**显示选定的变量转换后的值的表将显示。

**描述统计。**显示选定的变量的频率、缺失值和众数的表将显示。

## 分类回归：保存

使用“保存”对话框可以将预测值、残差和转换后的值保存到活动数据集和/或将离散化数据、转换后的值、规则化模型和系数以及回归系数符号保存到当前会话中的外部 IBM® SPSS® Statistics 数据文件或数据集中。

- 数据集在当前会话期间可用，但在后续会话期间不可用，除非显式将其保存为数据文件。数据集名称必须符合变量命名规则。
- 文件名或数据集名称对于保存的每种数据类型都必须各不相同。

图片 2-8  
保存对话框



只要在**规则化**对话框中选择规则化方式，即会保存规则化模型和系数。缺省情况下，该过程以唯一名称创建新数据集，当然您可以自行指定名称，或将其写入外部文件。

只要在**选项**对话框上使用多系统性起点作为初始配置，即会保存回归系数符号。缺省情况下，该过程以唯一名称创建新数据集，当然您可以自行指定名称，或将其写入外部文件。

## 分类回归：转换图

使用“图”对话框可以指定将生成转换图和残差图的变量。

图片 2-9  
“图”对话框



**变换图。**对于这些变量中的每个变量，将根据初始类别值绘制类别量化。空类别出现在水平轴上，但是不影响计算。这些类别由连接定量的行中的中断符标识。

**残差图。**对于这些变量中的每个变量，将根据类别指示符绘制残差（针对从除了有问题的预测变量之外的所有预测变量预测出的因变量进行计算），并且最优类别量化根据类别指示符乘以  $\beta$ 。

## CATREG 命令附加功能

如果将选定的内容粘贴到语法窗口中，并编辑生成的 CATREG 命令语法，则可以定制分类回归。使用命令语法语言还可以：

- 将已转换的变量保存到活动数据集时为其指定根名称（使用 SAVE 子命令）。

请参见命令语法参考以获取完整的语法信息。

# 分类主成份分析 (CATPCA)

此过程在减少数据维数的同时量化分类变量。分类主成份分析也表示为缩写词 CATPCA（代表 categorical principal components analysis）。

主成份分析的目标是将初始变量集缩减为表示初始变量中发现的大部分信息的较小不相关主成份集合。当大量变量妨碍有效解释对象（主体和单元）间关系时，该方法最为有用。通过减少维数，您只需解释少量主成份，而不是大量变量。

标准主成份分析假设数值变量间为线性关系。另一方面，通过最优尺度方法，可以将变量调整为不同级别。分类变量在指定维数内得到最优量化。因此，可以为变量间的非线性关系建模。

**示例。**分类主成份分析可用于以图形方式显示工作类别、工作部门、地区、旅行量（高、中、低）和工作满意度之间的关系。您可能会发现两个维占据了大量方差。第一维可能将工作类别与地区分开，而第二维可能将工作部门与旅行量分开。您可能还会发现较高的工作满意度与中等旅行量相关。

**统计量和图。**频率；缺失值；最佳度量水平；众数；按质心坐标、矢量坐标、每变量和每维总计解释的方差；矢量量化变量的成份载荷；类别量化和坐标；迭代历史记录；转换后变量和相关矩阵特征值的相关性；初始变量与相关矩阵特征值的相关性；对象得分；类别图；联合类别图；转换图；残差图；投影质心图；对象图；双标图；三标图和成份载荷图。

**数据。**字符串变量值总是按升序字母数值顺序转换为正整数。用户定义的缺失值、系统缺失值以及小于 1 的值都视为缺失值；可重新编码值小于 1 的变量，或者给值小于 1 的变量加上一个常数，以使其成为非缺失值。

**假设。**数据必须至少包含三个有效个案。该分析基于正整数数据。离散化选项通过将其值分组为具有接近正态分布的类别，将自动分类小数值变量，并且将自动把字符串变量的值转换为正整数。可指定其他离散化方案。

**相关过程。**将所有变量调整为数值级别对应于标准主成份分析。在标准线性主成份分析中使用转换后的变量可获得交替绘图功能。如果所有变量都有多名义尺度级别，则分类主成份分析等同于多重对应分析。如果需要处理的是变量集，则应使用分类（非线性）典型相关性分析。

## 获取分类主成份分析

- ▶ 从菜单中选择：  
分析 > 降维 > 最优尺度...

图片 3-1  
“最优尺度”对话框



- ▶ 选择有些变量不是无序多分类。
- ▶ 选择一个集合。
- ▶ 单击定义。

图片 3-2  
“分类主成份”对话框



- ▶ 选择至少两个分析变量并指定解的维数。

- ▶ 单击确定。

也可以指定补充变量，它们适合于找到的解，或者为图标变量。

## 在 CATPCA 中定义度量和权重

可以为分析变量和补充变量设置最佳度量水平。缺省情况下，它们调整为带两个内部结点的二度单调性样条（有序）。此外，还可以设置分析变量的权重。

图片 3-3  
“定义度量和权重”对话框



**变量权重。**可以选择为每个变量定义权重。指定的值必须为正整数。缺省值为 1。

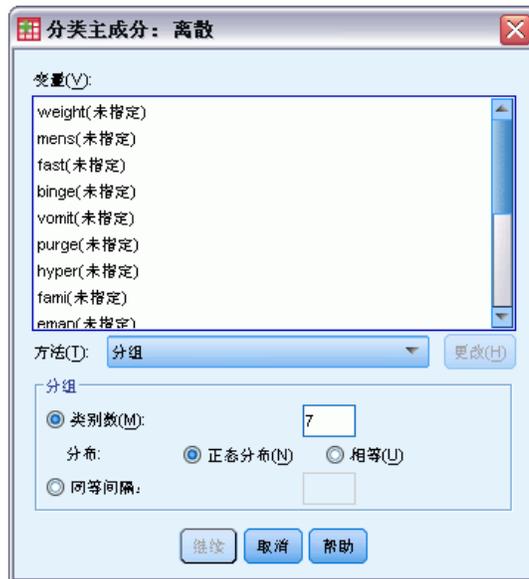
**最佳度量水平。**还可选择用于量化每个变量的尺度级别。

- **有序样条。**已观察到的变量的类别顺序保存在最优尺度化变量中。类别点将位于一条通过原点的直线上（矢量）。生成的转换是一个选定次数的平滑单调的分段多项式。这些段是由内部结点的用户指定的数量和过程决定的位置来指定的。
- **名义样条。**已观察到的变量中保留在最优尺度化变量中的唯一信息是类别中对象的分组。已观察到的变量的类别的顺序不会保留。类别点将位于一条通过原点的直线上（矢量）。生成的转换是一个选定次数的平滑的、可能非单调的分段多项式。这些段是由内部结点的用户指定的数量和过程决定的位置来指定的。
- **多标定。**已观察到的变量中保留在最优尺度化变量中的唯一信息是类别中对象的分组。已观察到的变量的类别的顺序不会保留。类别点将在特定类别中的对象的质心中。多表示为每一维获取不同的量化集。
- **有序。**已观察到的变量的类别顺序保存在最优尺度化变量中。类别点将位于一条通过原点的直线上（矢量）。生成的转换比有序样条转换拟合得好，但是平滑度较低。
- **名义。**已观察到的变量中保留在最优尺度化变量中的唯一信息是类别中对象的分组。已观察到的变量的类别的顺序不会保留。类别点将位于一条通过原点的直线上（矢量）。生成的转换比样条名义转换拟合得更好，但是平滑度较低。
- **数值。**类别被视为是有序的并且是等间距的（区间度）。已观察到的变量的类别顺序以及类别号之间的等间距保留在最优尺度化变量中。类别点将位于一条通过原点的直线上（矢量）。当所有变量都为数值级别时，分析与标准主成分分析类似。

## 分类主成份分析：离散化

使用“离散化”对话框可以选择对变量重新编码的方法。除非另有指定，否则小数变量将分组成具有近似正态分布的七个类别（如果变量的不同值的数目小于七，则将按此数目划分类别）。通过按照升序字母数值顺序分配类别指示符，字符串变量总是转换为正整数。字符串变量的离散化适用于这些整数。缺省情况下，其他变量保留原样。随后，分析中将使用离散化变量。

图片 3-4  
“离散化”对话框



**方法。**在分组、秩和乘之间选择。

- **分组。**重新编码为指定数量的类别或者按区间重新编码。
- **秩。**通过对个案排序来对变量进行离散化。
- **乘。**变量的当前值是乘以 10 且经过四舍五入的标准化值，并且加上了一个常数以使最低离散值为 1。

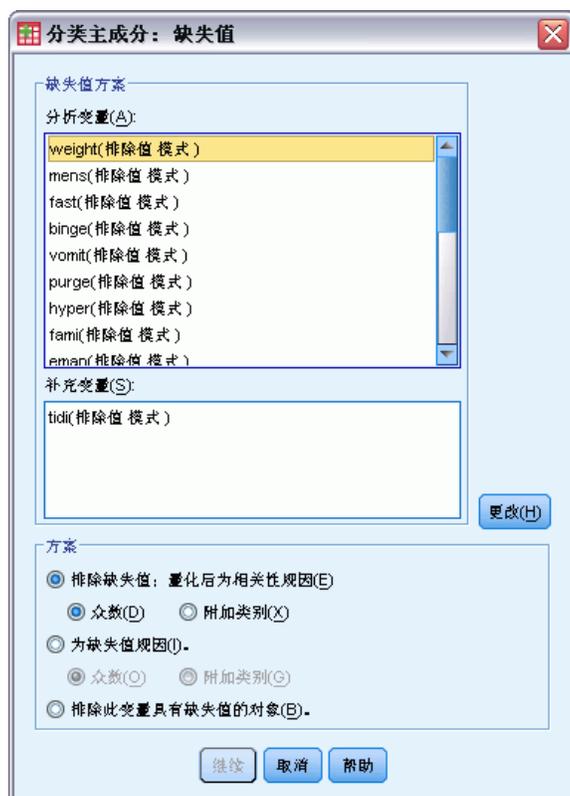
**分组。**在按分组离散化变量时，可以使用以下选项：

- **类别数。**指定一定数量的类别，并指定变量的值是否应该遵循在这些类别间的近似正态分布或均匀分布。
- **相等区间。**变量重新编码到由这些等大小的间隔所定义类别中。必须指定间隔的长度。

## 分类主成份分析：缺失值

使用“缺失值”对话框可以选择处理分析变量和补充变量中的缺失值的方案。

图片 3-5  
“缺失值”对话框



**方案。**选择排除缺失值（消极处理）、插补缺失值（积极处理）或者排除带缺失值的对象（列表删除）。

- **排除缺失值量化后为相关性归因。**选定的变量有缺失值的对象对于此变量的分析不起作用。如果消极处理所有变量，则所有变量都有缺失值的对象将视为补充对象。如果在“输出”对话框中指定了相关性，则（分析后）缺失值通过初始变量相关性变量的最频繁类别（即众数）进行归因。对于最优尺度化变量的相关性，可选择插补方法。选择众数将缺失值替换为最优尺度化变量的众数。选择附加类别将缺失值替换为附加类别的定量。这意味着此变量有缺失值的对象被视为属于同一（附加）类别。
- **插补缺失值。**选定的变量有缺失值的对象将对这些值归因。您可以选择插补方法。选择众数将缺失值替换为最频繁类别。当有多个众数时，将使用具有最小类别指示符的众数。选择附加类别将缺失值替换为相同的附加类别定量。这意味着此变量有缺失值的对象被视为属于同一（附加）类别。
- **排除此变量具有缺失值的对象。**选定的变量有缺失值的对象将从分析中排除。此方案不适用于补充变量。

## 分类主成份分析：选项

使用“选项”对话框可以选择初始配置、指定迭代和收敛标准、选择正态化方法、选择标记图的方法以及指定附加对象。

图片 3-6  
“选项”对话框



**附加对象。**指定要其成为附加对象的对象的个案编号，或者对象范围的第一个和最后一个个案编号，然后单击添加。继续操作，直到指定完所有附加对象。如果将某个对象指定为附加对象，则对于该对象将忽略个案权重。

**正态化方法。**可以指定用于标准化对象得分和变量的五个选项之一。给定分析中只能使用一种正态化方法。

- **主要变量。**此选项可优化变量之间的关联。对象空间中的变量坐标是成份载入（与主成分的相关性，如维和对象得分）。如果您主要对变量之间的相关性感兴趣，这将非常有用。
- **主要对象。**此选项可优化对象间的距离。如果您主要对对象之间的区别或相似性感兴趣，这将非常有用。
- **对称。**如果主要对对象和变量之间的关系感兴趣，则使用此标准化选项。
- **独立。**如果您想单独检查对象之间的距离和变量之间的相关性，请使用此标准化选项。
- **定制。**可指定封闭区间  $[-1, 1]$  中的任何实数值。值为 1 等同于“主要对象”方法；值为 0 等同于“对称”方法；值为 -1 等同于“主要变量”方法。通过指定大于 -1 小于 1 的值，可在对象和变量上分布特征值。此方法对于制作合适的双标图或三标图很有用。

**标准。**可以指定该过程可在其计算中执行的最大迭代次数。还可以选择收敛标准值。如果上两次迭代之间的总拟合之差小于收敛值，或者达到了最大迭代次数，则算法停止迭代。

**标注图。**可用于指定在图中将使用变量和值标签还是变量名称和值。还可指定标签的最大长度。

**图维数。**可用以控制在输出中显示的维数。

- **显示解中的所有维数。**解中的所有维数都显示在散点图矩阵中。
- **限制维数。**显示的维数限制为绘制的对。如果限制维数，则必须选择要绘制的最低和最高维数。最低维数的范围可从 1 到解中的维数减 1，并且针对较高维数绘制。最高维数值的范围可从 2 到解中的维数，表示要在绘制维数对时使用的最高维数。此指定项适用于所有请求的多维图。

**配置。**可从包含配置的坐标的文件中读取数据。文件中的第一个变量应包含第一维的坐标，第二个变量应包含第二维的坐标，依此类推。

- **初始。**指定的文件中的配置将用作分析起点。
- **固定。**指定的文件中的配置将用于拟合变量。拟合的变量必须选择作为分析变量，但是因为配置是固定的，所以它们视为补充变量（因此不需要选择它们作为补充变量）。

## 分类主成份分析： 输出

使用“输出”对话框可以为对象得分、成份载入、迭代历史、初始变量和转换后的变量的相关性、按每个变量和每个维解释的方差、选定的变量的类别量化以及选定的变量的描述统计生成表。

图片 3-7  
“输出”对话框



**对象得分。**显示对象得分并且有以下选项：

- **包括类别。**显示选定的分析变量的类别指示符。
- **标注对象得分。**从指定为标签变量的变量列表中，可选择一個用以标注对象的变量。

**成份载入。**显示未获得多标定尺度级别的所有变量的成份载入。

**迭代历史记录。**对于每次迭代，显示偏差考虑情况以及偏差考虑情况中的损失和增加项。

**初始变量的相关性。**显示初始变量的相关性矩阵以及该矩阵的特征值。

**转换变量的相关性。**显示已转换的（最优尺度）变量的相关性矩阵以及该矩阵的特征值。

**偏差考虑情况。**显示按质心坐标、矢量坐标以及每个变量和每个维总计（合并的质心坐标和矢量坐标）解释的方差量。

**类别量化。**给出选定的变量每一维的类别量化和坐标。

**描述统计。**显示选定的变量的频率、缺失值的数量以及众数。

## 分类主成份分析：保存

使用“保存”对话框可以将离散化数据、对象得分、转换后的值和近似值保存到当前会话中的外部 IBM® SPSS® Statistics 数据文件或数据集中。还可以将转换后的值、对象得分和近似值保存到活动数据集中。

- 数据集在当前会话期间可用，但在后续会话期间不可用，除非显式将其保存为数据文件。数据集名称必须符合变量命名规则。
- 文件名或数据集名称对于保存的每种数据类型都必须各不相同。
- 如果将对象得分或已转换的值保存到活动数据集中，则可以指定多名义维数的数量。

图片 3-8  
保存对话框



## 分类主成份分析：对象图

使用“对象和变量图”对话框可以指定所要的图类型以及将为其生成图的变量。

图片 3-9  
“对象和变量图”对话框



**对象点。**显示对象点的图。

**对象和变量（双标图）。**对象点是用您所选择的变量坐标（成份载入或变量质心）绘制的。

**对象、载入和质心（三标图）。**对象点是用多名义尺度级别变量的质心以及其他变量的成份载入绘制的。

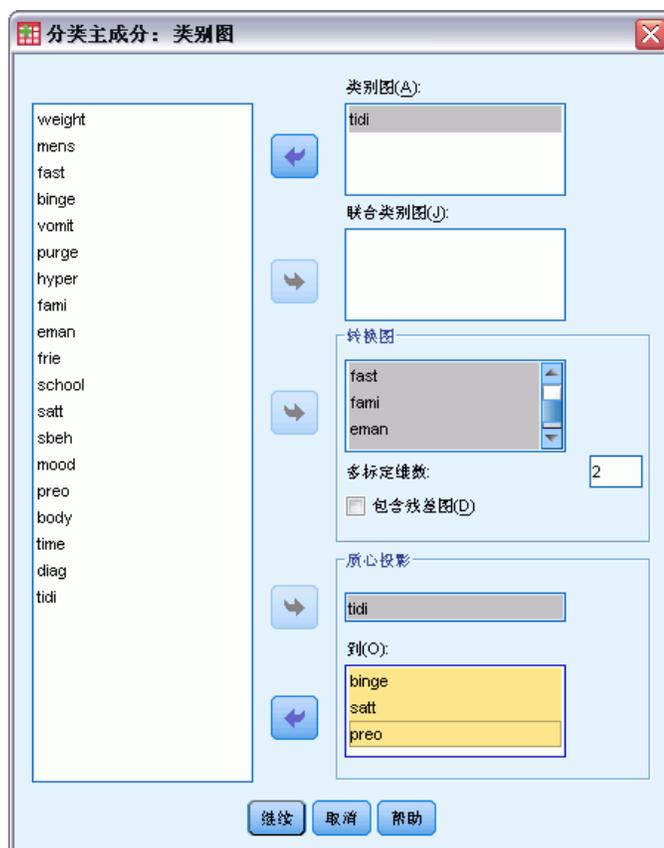
**双标图和三标图变量。**可选择将所有变量用于双标图和三标图或选择一个子集。

**标签对象。**可选择用选定的变量的类别（可在“选项”对话框中选择类别指示符值或值标签）或其个案号标注对象。如果选择变量，则为每个变量生成一个图。

## 分类主成份分析：类别图

使用“类别图”对话框可以指定所要的图类型以及将为其生成图的变量。

图片 3-10  
“类别图”对话框



**类别图。**对于选定的每个变量绘制质心和矢量坐标图。对于多名义尺度级别的变量，类别在该特定类别中的对象的质心中。对于所有其他尺度级别，类别在通过原点的矢量上。

**联合类别图。**这是每个选定的变量的质心和矢量坐标的单个图。

**变换图。**显示最优类别量化与类别指示符的比较图。可以为拥有多名义尺度级别的变量指定所要的维数；将为每一维分别生成一个图。还可以选择显示每个选定的变量的残差图。

**质心投影。**可选择一个变量并将其质心投影到选定的变量上。不能选择拥有多名义尺度级别的变量作为投影的对象。当请求此图时，含有投影的质心的坐标的表也会同时显示。

## 分类主成份分析： 载荷图

使用“载荷图”对话框可以指定将包含在图中的变量以及是否在图中包含质心。

图片 3-11  
“载荷图”对话框



**显示成份载入。**如果选中，将显示成份载荷图。

**载入变量。**可选择将所有变量用于成份载荷图或选择一个子集。

**包含质心。**拥有多名义尺度级别的变量没有成份载入，但是可以选择在图中包含这些变量的质心。可选择使用所有多名义变量或选择一个子集。

## CATPCA 命令附加功能

如果将选中的内容粘贴到语法窗口中，并编辑得到的 CATPCA 命令语法，则可以定制分类主成份分析。使用命令语法语言还可以：

- 将已转换的变量、对象得分和近似值保存到活动数据集时（使用 SAVE 子命令），为其指定根名称。
- 分别为每个图指定最大标签长度（使用 PLOT 子命令）。
- 为残差图指定单独的变量列表（使用 PLOT 子命令）。

请参见命令语法参考以获取完整的语法信息。

# 非线性典型相关性分析 (OVERALS)

非线性典型相关性分析对应于使用最优尺度的分类典型相关性分析。此过程的目的是确定分类变量集相互之间的相似程度。非线性典型相关性分析也用缩写词 OVERALS 来表示。

标准典型相关性分析是多重回归的扩展，其中第二个集不包含单响应变量，而是包含多响应变量。其目标是尽可能解释低维空间中两个数值变量集之间的关系中的方差。最初，每个集内的变量进行线性组合以使线性组合有最大的相关性。有了这些组合，就可以确定后续线性组合与前面的组合无关，并可确定其具有可能的最大相关性。

最优尺度方法在三个重要方面扩展了标准分析。首先，OVERALS 允许两个以上的变量集。其次，变量或者可调整为名义、有序，或者调整为数值。因此，可以分析变量间的非线性关系。最后，变量集与一个由对象得分定义的未知折中集进行比较，而不是使变量集之间的相关性最大化。

**示例。**使用最优尺度的分类典型相关性分析可用于以图形方式显示包含工作类别和教育年限的一个变量集与包含居住地区和性别的另一个变量集之间的关系。您可能会发现教育年限与居住地区的区别程度比其余变量高。您可能还会发现教育年限在第一维上区别最大。

**统计量和图。**频率、质心、迭代历史记录、对象得分、类别量化、权重、成份载入、单拟合和多拟合、对象得分图、类别坐标图、成份载荷图、类别质心图、转换图。

**数据。**使用整数来编码分类变量（名义或有序尺度级别）。要最小化输出，请使用从 1 开始的连续整数来编码每个变量。调整为数值级别的变量不应重新编码为连续整数。要最小化输出，对于调整为数值级别的每个变量，从每个值中减去最小观察值然后加上 1。小数表示的值则截去小数部分。

**假设。**变量可分成两个或更多变量集。分析中的变量调整为多名义、单名义、有序或数值。过程中使用的最大维数取决于变量的最佳度量水平。如果所有变量都指定为有序、单名义或数值，则最大维数是以下两个值中的较小者：观察次数减 1 或变量的总数。但是，如果只定义了两个变量集，则最大维数为较小集中的变量数。如果某些变量为多名义，则最大维数为多名义类别的总数加上非多名义变量的数目减去多名义变量的数目。例如，如果分析涉及五个变量，其中一个变量是带有四种类别的多名义变量，则最大维数为  $(4 + 4 - 1)$ ，即 7。如果指定了大于最大值的数，则会使用最大值。

**相关过程。**如果每个集只包含一个变量，则非线性典型相关性分析等效于使用最优尺度的主成分分析。如果所有这些都是多名义，则分析对应于多重对应分析。如果涉及两个变量集，并且其中一个仅包含一个变量，则分析等同于使用最优尺度的分类回归。

## 获取非线性典型相关性分析

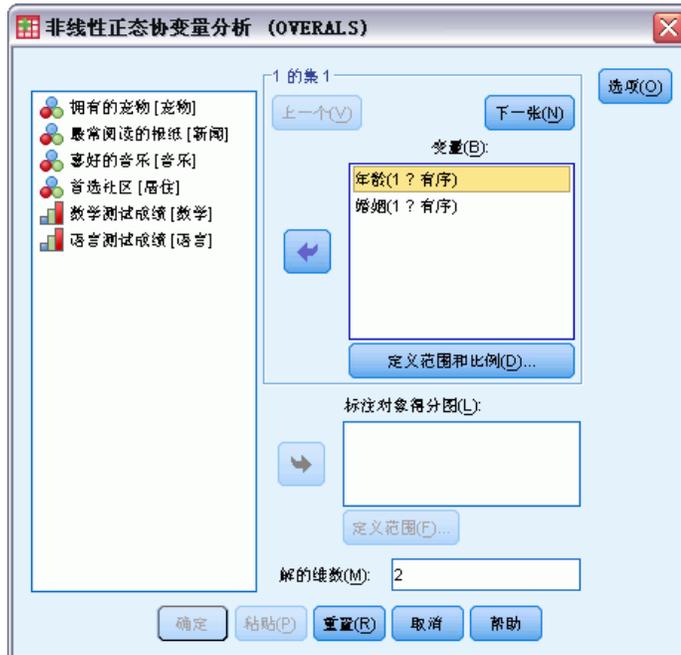
- ▶ 从菜单中选择：  
分析 > 降维 > 最优尺度...

图片 4-1  
“最优尺度”对话框



- ▶ 或者选择所有变量都是无序多分类，或者选择有些变量不是无序多分类。
- ▶ 选择多个集合。
- ▶ 单击定义。

图片 4-2  
“非线性正态协变量分析 (OVERALS)”对话框



- ▶ 定义至少两个变量集。选择要包含在第一个集内的变量。若要移到下一个集合，单击下一个，然后选择要包含在第二个集合中的变量。可根据需要添加更多集合。单击上一个返回到先前定义的变量集。
- ▶ 为每个选定的变量定义值范围和测量尺度（最佳度量水平）。
- ▶ 单击确定。
- ▶ 或者：
  - 选择一个或多个要为对象得分图提供点标签的变量。各个变量分别产生一个图，由该变量的值标注各个点。必须为各个图标签变量分别定义一个范围。在使用该对话框时，单个变量不能既在分析中使用，又作为标签变量使用。如果需要用在分析中使用的变量标注对象得分图，请使用“计算”工具（可从“转换”菜单上获取）创建该变量的副本。使用新变量标注图。或者，可使用命令语法。
  - 指定解中您所要的维数。通常，根据需要选择尽量少的维数来解释大多数变异。如果分析涉及两个以上的维，则将生成前三维的三维图。通过编辑图表可显示其他维。

## 定义范围和尺度

图片 4-3  
“定义范围和尺度”对话框



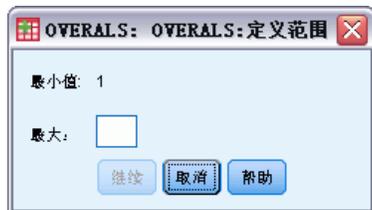
必须为每个变量定义范围。指定的最大值必须是整数。小数数据值会在分析中被截断。指定范围之外的类别值在分析中会被忽略。为了使输出最小，请使用“自动重新编码”工具（可从“转换”菜单上获取）为视作名义或有序的变量创建从 1 开始的连续类别。对于调整为数值级别的变量，建议不要重新编码为连续整数。要最小化视为数值的变量的输出，对于每个变量，从每个值中减去最小值然后加上 1。

还必须选择将用于量化每个变量的尺度。

- **有序**。观察到的变量的类别顺序保留在已量化的变量中。
- **单标定**。在已量化变量中，相同类别中的对象得到相同得分。
- **多标定**。每个维的定量可以不同。
- **离散数**。类别被视为有序且等间距。观察到的变量的类别号和类别顺序之间的差别保留在已量化的变量中。

## 定义范围

图片 4-4  
“定义范围”对话框



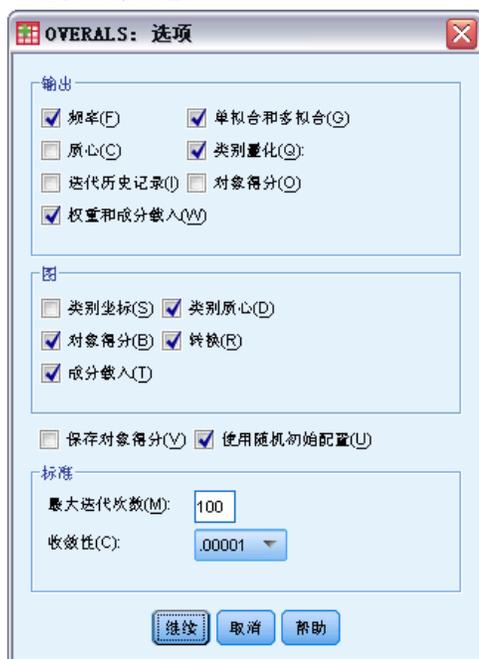
必须为每个变量定义范围。指定的最大值必须是整数。小数数据值会在分析中被截断。指定范围之外的类别值在分析中会被忽略。要最小化输出，请使用“自动重新编码”工具（可从“转换”菜单上获取）创建从 1 开始的连续类别。

还必须为用于标注对象得分图的每个变量定义范围。但是，数据值在为变量定义的范围之外的类别的标签确实显示在图中。

## 非线性典型相关性分析：选项

使用“选项”对话框可以选择可选统计量和图、将对象得分保存为活动数据集中的新变量、指定迭代和收敛标准以及为分析指定初始配置。

图片 4-5  
“选项”对话框



**显示。** 可用统计量包括边际频率（计数）、质心、迭代历史记录、权重和成份载入、类别量化、对象得分以及单拟合和多拟合统计量。

- **质心.** 类别量化, 对象得分的投影平均值和实际平均值, 其中的对象 (个案) 包含在属于相同变量类别的那些对象的每个集合中。
- **权重和成分载入.** 集合中每个已量化的变量的每个维度的回归系数 (其中, 在已量化的变量上对对象得分进行回归) 以及已量化的变量在对象空间中的投影。它指示每个变量对每个集合中的维度的贡献。
- **单拟合和多拟合.** 对于对象, 是对单和多类别坐标 / 类别量化的拟合优度的测量。
- **类别量化.** 分配给变量类别的最优刻度值。
- **对象得分.** 分配给特定维度中某个对象 (个案) 的最优得分。

**图.** 可生成类别坐标图、对象得分图、成份载荷图、类别质心图以及转换图。

**保存对象得分.** 可将对象得分保存为活动数据集中的新变量。对象得分针对在主对话框中指定的维数保存。

**使用随机初始配置.** 如果部分或全部变量为单名义, 则应使用随机初始配置。如果未选择此选项, 则使用嵌套初始配置。

**标准.** 可以指定非线性典型相关性分析可在其计算中执行的最大迭代次数。还可以选择收敛标准值。如果上两次迭代之间的总拟合之差小于收敛值, 或者达到了最大迭代次数, 则分析停止迭代。

## OVERALS 命令附加功能

如果将选中的内容粘贴到语法窗口中, 并编辑生成的 OVERALS 命令语法, 则可以定制非线性典型相关性分析。使用命令语法语言还可以:

- 指定要绘制的维对, 而不是绘制所有抽取的维 (在 PLOT 子命令上使用 NDIM 关键字)。
- 指定用于在图上标注点的值标签字符的数目 (使用 PLOT 子命令)。
- 指定五个以上的变量作为对象得分图的标记变量 (使用 PLOT 子命令)。
- 选择在分析中使用的变量作为对象得分图的标记变量 (使用 PLOT 子命令)。
- 选择变量以便为定量得分图提供点标签 (使用 PLOT 子命令)。
- 如果不想使用活动数据集中所有个案, 请指定要包含在分析中的个案数 (使用 NOOBSERVATIONS 子命令)。
- 为通过保存对象得分创建的变量指定根名称 (使用 SAVE 子命令)。
- 指定要保存的维数, 而不是保存所有抽取的维 (使用 SAVE 子命令)。
- 将类别量化写入矩阵文件 (使用 MATRIX 子命令)。
- 生成可能比常规高分辨率图更易于读取的低分辨率图 (使用 SET 子命令)。
- 仅对指定变量生成质心和转换图 (使用 PLOT 子命令)。

请参见命令语法参考以获取完整的语法信息。

# 对应分析

对应分析的目标之一是描述低维空间中对应表中的两个名义变量之间的关系，同时描述每个变量的类别之间的关系。对于每个变量，类别点在图中的距离反映了相似的类别绘制为相互靠近的类别间的关系。从原点到另一个变量的类别点的矢量上的一个变量的投影点描述了变量之间的关系。

列联表的分析通常包括检查行和列轮廓表以及通过卡方统计检验自变量。但是，轮廓表的数量可能非常大，并且卡方检验不会揭示因变量结构。交叉表过程提供了多个关联性测量和关联性检验，但是不能以图形方式表示变量间的任何关系。

因子分析是描述低维空间中的变量之间的关系的标准技术。但是，因子分析需要定距数据，并且观察数应为变量数的五倍。而对应分析采用名义变量，并且可以描述每个变量的类别之间的关系和变量之间的关系。此外，对应分析可用于分析任何正对应度量的表。

**示例。**对应分析可用于以图形的方式显示员工类别和吸烟习惯之间的关系。您可能会发现，就吸烟而言，初级经理人有别于秘书，但是秘书与高级经理人没差别。您可能还会发现重度烟瘾与初级经理人相关，而轻度烟瘾与秘书相关。

**统计量和图。**对应度量、行和列轮廓表、奇异值、行和列得分、惯量、质量、行和列得分置信统计量、奇异值置信统计量、转换图、行点图、列点图和双标图。

**数据。**要分析的分类变量调整为名义。对于汇总数据或除频率之外的对应度量，应使用具有正相似性值的加权变量。或者，对于表数据，使用语法读取表。

**假设。**该过程中使用的最大维数取决于活动行和列类别的数目以及等同性约束的数目。如果未使用等同性约束并且所有类别都是活动的，则最大维数比具有最少类别的变量的类别数少一。例如，如果一个变量有五个类别，另一个变量有四个类别，则最大维数为三。补充类别不是活动类别。例如，如果一个变量有五个类别，其中两个是补充类别，而另一个变量有四个类别，则最大维数为二。将所有约束为相等的类别集视为一个类别。例如，如果一个变量有五个类别，其中三个约束为相等，则在确定最大维数时，应将该变量视为有三个类别。其中两个类别未受约束，第三个类别对应于三个受约束的类别。如果指定大于最大值的维数，则会使用最大值。

**相关过程。**如果涉及两个以上的变量，则使用多重对应分析。如果变量应调整为有序，则使用分类主成份分析。

## 获取对应分析

- ▶ 从菜单中选择：  
分析 > 降维 > 对应分析...

图片 5-1  
“对应分析”对话框

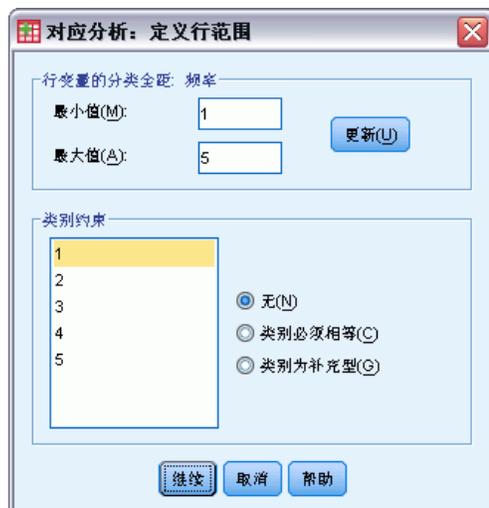


- ▶ 选择行变量。
- ▶ 选择列变量。
- ▶ 定义变量的范围。
- ▶ 单击确定。

## 在“对应分析”中定义行范围

必须为行变量定义范围。指定的最小值和最大值必须为整数。小数数据值会在分析中被截断。指定范围之外的类别值在分析中会被忽略。

图片 5-2  
“定义行范围”对话框



初始状态下，所有类别都不受约束且是活动的。可将行类别约束为等于其他行类别，或者可将行类别定义为补充类别。

- **类别必须相等。**类别必须具有相等的得分。如果所获得的类别顺序不理想或不直观，请使用等同性约束。可约束为相等的行类别的最大数量等于活动行类别总数减 1。若要对类别集施加不同的等同性约束，请使用语法。例如，使用语法，将类别 1 和 2 约束为相等，将类别 3 和 4 约束为相等。
- **类别为补充型。**补充类别不影响分析，但会出现在由活动类别定义的空间中。补充类别对定义维不起作用。补充行类别的最大数目为行类别总数减 2。

## 在“对应分析”中定义列范围

必须为列变量定义范围。指定的最小值和最大值必须为整数。小数数据值会在分析中被截断。指定范围之外的类别值在分析中会被忽略。

图片 5-3  
“定义列范围”对话框



初始状态下，所有类别都不受约束且是活动的。可将列类别约束为等于其他列类别，或者可将列类别定义为补充类别。

- **类别必须相等。**类别必须具有相等的得分。如果所获得的类别顺序不理想或不直观，请使用等同性约束。可约束为相等的列类别的最大数量等于活动列类别总数减 1。若要对类别集施加不同的等同性约束，请使用语法。例如，使用语法，将类别 1 和 2 约束为相等，将类别 3 和 4 约束为相等。
- **类别为补充型。**补充类别不影响分析，但会出现在由活动类别定义的空间中。补充类别对定义维不起作用。补充列类别的最大数目为列类别总数减 2。

## 对应分析：模型

使用“模型”对话框可以指定维数、距离测量、标准化方法以及正态化方法。

图片 5-4  
“模型”对话框



**解的维数。**指定维数。通常，根据需要选择尽量少的维数来解释大多数变异。最大维数取决于分析中使用的活动类别数以及等同性约束。最大维数是以下两项中的较小者：

- 活动行类别数减去约束为相等的行类别数，加上受约束的行类别集的数目。
- 活动列类别数减去约束为相等的列类别数，加上受约束的列类别集的数目。

**距离测量。**可选择对应表的行和列之间距离的测量。选择下列选项之一：

- **卡方。**使用加权轮廓表距离，其中权重是行或列的质量。此度量是标准对应分析所必需的。
- **欧几里德距离。**使用行对和列对之间平方差之和的平方根。

**标准化方法。**选择下列选项之一：

- **行和列均值已删除。**行和列都居中。此方法是标准对应分析所必需的。
- **行均值已删除。**只有行居中。
- **列均值已删除。**只有列居中。
- **使行总和相等，移去删除。**使行居中之前，使行边距相等。
- **使列总和相等，移去删除。**使列居中之前，使列边距相等。

**正态化方法。**选择下列选项之一：

- **对称。**对于每个维，行得分为列得分的加权平均值除以对应的奇异值，列得分为行得分的加权平均值除以对应的奇异值。如果想要检查两个变量的类别之间的差异或相似性，则使用此方法。

- **主要。**行点和列点之间的距离是对应表中对应于所选距离测量的距离的近似值。如果想要检查一个或两个变量的类别之间的差别，而非两个变量之间的差别，则使用此方法。
- **主要行。**行点之间的距离是对应表中对应于所选距离测量的距离近似值。行得分是列得分的加权平均值。如果想要检查行变量的类别之间的差异或相似性，则使用此方法。
- **主要列。**列点之间的距离是对应表中对应于所选距离测量的距离近似值。列得分是行得分的加权平均值。如果想要检查列变量的类别之间的差异或相似性，则使用此方法。
- **定制。**必须指定介于  $-1$  和  $1$  之间的值。值为  $-1$  对应于主要列。值为  $1$  对应于主要行。值为  $0$  对应于对称。所有其他值不同程度地将惯量分布于行得分和列得分上。此方法对于制作合适的双标图很有用。

## 对应分析： 统计量

使用“统计量”对话框可以指定生成的数值输出。

图片 5-5  
“统计量”对话框



**对应表。**带有行和列边际总计的输入变量的交叉表。

**行点概览。**对于每个行类别，有得分、质量、惯量、对维惯量的贡献，以及维对点惯量的贡献。

**列点概览。**对于每个列类别，有得分、质量、惯量、对维惯量的贡献，以及维对点惯量的贡献。

**行轮廓表。**对于每个行类别，为跨列变量类别的分布。

**列轮廓表。**对于每个列类别，为跨行变量类别的分布。

**对应表的排列。**对应表重新组织为：行和列根据第一维上的得分按递增顺序排列。或者，可指定将为其生成置换表的最大维数。将为从  $1$  到指定数字的每一维分别生成一个置换表。

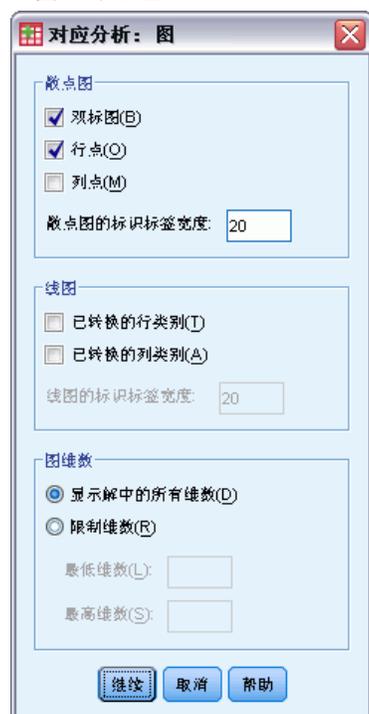
**行点的置信统计量。**包括所有非补充行点的标准差和相关性。

**列点的置信统计量。**包括所有非补充列点的标准差和相关性。

## 对应分析：图

使用“图”对话框可以指定生成哪些图。

图片 5-6  
“图”对话框



**散点图。**生成维的所有成对图矩阵。可用散点图包括：

- **双标图。**生成行点和列点的联合图矩阵。如果选择了主标准化，则双标图不可用。
- **行点。**生成行点图的矩阵。
- **列点。**生成列点图的矩阵。

或者，可指定在标注点时要使用多少个值标签字符。该值必须为小于或等于 20 的非负整数。

**线图。**为所选变量的每一维生成一个图。可用的线图包括：

- **已转换的行类别。**根据初始行类别值的对应行得分生成这些值的图。
- **已转换的列类别。**根据初始列类别值的对应列得分生成这些值的图。

或者，可指定在标注类别轴时要使用多少个值标签字符。该值必须为小于或等于 20 的非负整数。

**图维数。**可用以控制在输出中显示的维数。

- **显示解中的所有维数。**解中的所有维数都显示在散点图矩阵中。
- **限制维数。**显示的维数限制为绘制的对。如果限制维数，则必须选择要绘制的最低和最高维数。最低维数的范围可从 1 到解中的维数减 1，并且针对较高维数绘制。最高维数值的范围可从 2 到解中的维数，表示要在绘制维数对时使用的最高维数。此指定项适用于所有请求的多维图。

## CORRESPONDENCE 命令附加功能

如果将选定的内容粘贴到语法窗口中，并编辑生成的 CORRESPONDENCE 命令语法，则可以定制对应分析。使用命令语法语言还可以：

- 指定表数据作为输入，而不是使用个案情况数据（使用 TABLE = ALL 子命令）。
- 指定用于标注每种类型的散点图矩阵或双标图矩阵的点的值标签字符数（使用 PLOT 子命令）。
- 指定用于标注每种类型的线图的点的值标签字符数（使用 PLOT 子命令）。
- 将行和列得分的矩阵写入矩阵数据文件（使用 OUTFILE 子命令）。
- 将奇异值和得分的置信统计量（方差和协方差）的矩阵写入矩阵数据文件（使用 OUTFILE 子命令）。
- 指定多个类别集相等（使用 EQUAL 子命令）。

请参见命令语法参考以获取完整的语法信息。

# 多重对应分析

多重对应分析通过为个案（对象）和类别分配数值量化标定（分类）数据，以使同类别的对象紧靠在一起，不同类别的对象远远分开。每个对象都尽可能地接近应用于该对象的类别的类别点。这样，类别可将对象分成数个均一子组。当变量将相同类别中的对象分类到同一子组中时，这些变量被视为是均一的。

**示例。**多重对应分析可用以以图形方式显示工作类别、少数民族分类和性别之间的关系。您可能发现在人群之间少数民族分类和性别有区别，但是工作类别没有区别。您可能还会发现拉丁裔美国人和非洲裔美国人类别相似。

**统计量和图。**对象得分、区分测量、迭代历史记录、原始变量和转换后变量的相关性、类别量化、描述统计、对象点图、双标图、类别图、联合类别图、转换图和区分测量图。

**数据。**字符串变量值总是按升序字母数值顺序转换为正整数。用户定义的缺失值、系统缺失值以及小于 1 的值都视为缺失值；可重新编码值小于 1 的变量，或者给值小于 1 的变量加上一个常数，以使其成为非缺失值。

**假设。**所有变量都有多名义尺度级别。数据必须至少包含三个有效个案。该分析基于正整数数据。离散化选项通过将其值分组成具有接近正态分布的类别，将自动分类小数值变量，并且将自动把字符串变量的值转换为正整数。可指定其他离散化方案。

**相关过程。**对于两个变量，多重对应分析类似于对应分析。如果相信变量具有有序或数值属性，则应使用分类主成份分析。如果需要处理变量集，则应使用非线性典型相关性分析。

## 获取多重对应分析

- ▶ 从菜单中选择：  
分析 > 降维 > 最优尺度...

图片 6-1  
“最优尺度”对话框



- ▶ 选择所有变量都是无序多分类。
- ▶ 选择一个集合。
- ▶ 单击定义。

图片 6-2  
“多重对应分析”对话框



- ▶ 选择至少两个分析变量并指定解的维数。

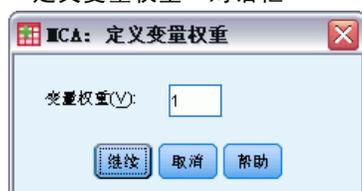
- ▶ 单击确定。

也可以指定补充变量，它们适合于找到的解，或者为图标记变量。

## 在“多重对应分析”中定义变量权重

可以设置分析变量的权重。

图片 6-3  
“定义变量权重”对话框

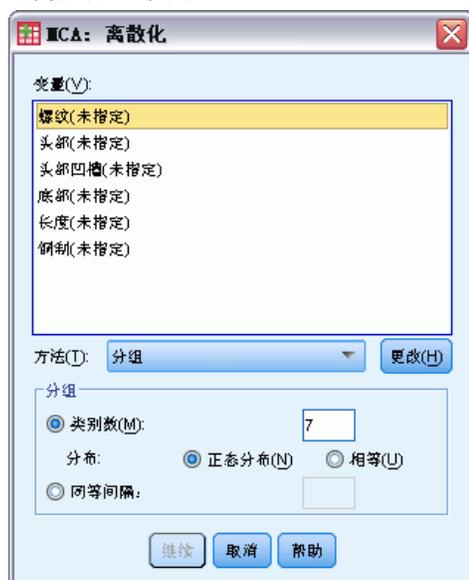


**变量权重。**可以选择为每个变量定义权重。指定的值必须为正整数。缺省值为 1。

## 多重对应分析：离散化

使用“离散化”对话框可以选择对变量重新编码的方法。除非另有指定，否则小数值变量分组为具有近似正态分布的七个类别（如果变量的相异值的数目小于七，则按此数目划分类别）。通过按照升序字母数值顺序分配类别指示符，字符串变量总是转换为正整数。字符串变量的离散化适用于这些整数。缺省情况下，其他变量保留原样。随后，分析中将使用离散化变量。

图片 6-4  
“离散化”对话框



**方法。**在分组、秩和乘之间选择。

- **分组。**重新编码为指定数量的类别或者按区间重新编码。
- **秩。**通过对个案排序来对变量进行离散化。
- **乘。**变量的当前值是乘以 10 且经过四舍五入的标准化值，并且加上了一个常数以使最低离散值为 1。

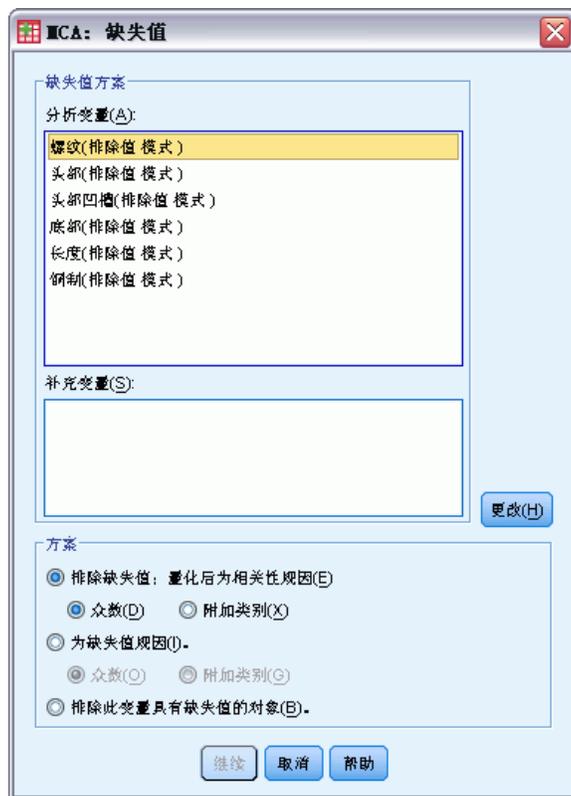
**分组。**在按分组离散化变量时，可以使用以下选项：

- **类别数。**指定一定数量的类别，并指定变量的值是否应该遵循在这些类别间的近似正态分布或均匀分布。
- **相等区间。**变量重新编码到由这些等大小的间隔所定义的类别中。必须指定间隔的长度。

## 多重对应分析： 缺失值

使用“缺失值”对话框可以选择处理分析变量和补充变量中的缺失值的方案。

图片 6-5  
“缺失值”对话框



**缺失值方案。**选择排除缺失值（消极处理）、插补缺失值（积极处理）或者排除带缺失值的对象（列表删除）。

- **排除缺失值；量化后为相关性规因。**选定的变量有缺失值的对象对于此变量的分析不起作用。如果消极处理所有变量，则所有变量都有缺失值的对象将视为补充对象。如果在“输出”对话框中指定了相关性，则（分析后）缺失值通过初始变量相关性变量

的最频繁类别（即众数）进行归因。对于最优尺度化变量的相关性，可选择插补方法。选择众数将缺失值替换为最优尺度化变量的众数。选择附加类别将缺失值替换为附加类别的定量。这意味着此变量有缺失值的对象被视为属于同一（附加）类别。

- **为缺失值归因。**选定的变量有缺失值的对象将对这些值归因。您可以选择插补方法。选择众数将缺失值替换为最频繁类别。当有多个众数时，将使用具有最小类别指示符的众数。选择附加类别将缺失值替换为相同的附加类别定量。这意味着此变量有缺失值的对象被视为属于同一（附加）类别。
- **排除此变量具有缺失值的对象。**选定的变量有缺失值的对象将从分析中排除。此方案不适用于补充变量。

## 多重对应分析：选项

使用“选项”对话框可以选择初始配置、指定迭代和收敛标准、选择正态化方法、选择标记图的方法以及指定附加对象。

图片 6-6  
“选项”对话框



**附加对象。**指定要其成为附加对象的对象的个案编号，或者对象范围的第一个和最后一个个案编号，然后单击添加。继续操作，直到指定完所有附加对象。如果将某个对象指定为附加对象，则对于该对象将忽略个案权重。

**正态化方法。**可以指定用于标准化对象得分和变量的五个选项之一。给定分析中只能使用一种正态化方法。

- **主要变量。**此选项可优化变量之间的关联。对象空间中的变量坐标是成份载入（与主成分的相关性，如维和对象得分）。如果您主要对变量之间的相关性感兴趣，这将非常有用。
- **主要对象。**此选项可优化对象间的距离。如果您主要对对象之间的区别或相似性感兴趣，这将非常有用。
- **对称。**如果主要对对象和变量之间的关系感兴趣，则使用此标准化选项。
- **独立。**如果您想单独检查对象之间的距离和变量之间的相关性，请使用此标准化选项。
- **定制。**可指定封闭区间  $[-1, 1]$  中的任何实数值。值为 1 等同于“主要对象”方法；值为 0 等同于“对称”方法；值为 -1 等同于“主要变量”方法。通过指定大于 -1 小于 1 的值，可在对象和变量上分布特征值。此方法对于制作合适的双标图或三标图很有用。

**标准。**可以指定该过程可在其计算中执行的最大迭代次数。还可以选择收敛标准值。如果上两次迭代之间的总拟合之差小于收敛值，或者达到了最大迭代次数，则算法停止迭代。

**标注图。**可用于指定在图中将使用变量和值标签还是变量名称和值。还可指定标签的最大长度。

**图维数。**可用以控制在输出中显示的维数。

- **显示解中的所有维数。**解中的所有维数都显示在散点图矩阵中。
- **限制维数。**显示的维数限制为绘制的对。如果限制维数，则必须选择要绘制的最低和最高维数。最低维数的范围可从 1 到解中的维数减 1，并且针对较高维数绘制。最高维数值的范围可从 2 到解中的维数，表示要在绘制维数对时使用的最高维数。此指定项适用于所有请求的多维图。

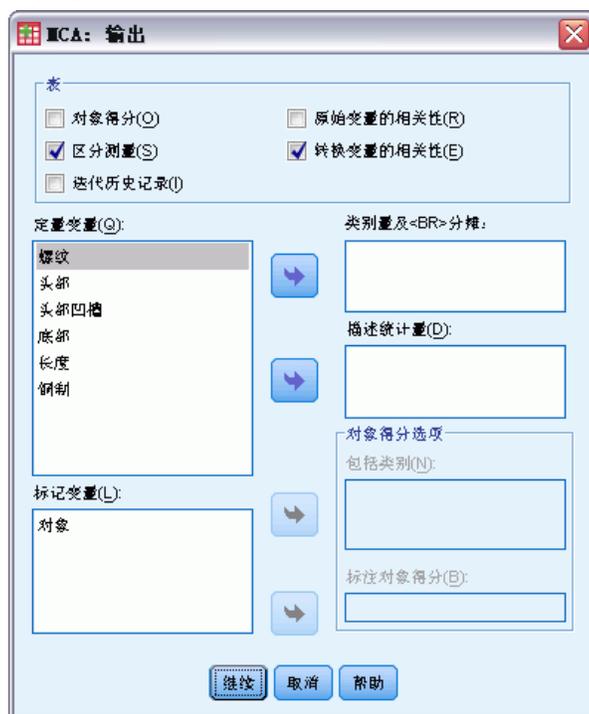
**配置。**可从包含配置的坐标的文件中读取数据。文件中的第一个变量应包含第一维的坐标，第二个变量应包含第二维的坐标，依此类推。

- **初始。**指定的文件中的配置将用作分析起点。
- **固定。**指定的文件中的配置将用于拟合变量。拟合的变量必须选择作为分析变量，但是，因为配置是固定的，所以它们被视为补充变量（因此不需要选择它们作为补充变量）。

## 多重对应分析： 输出

使用“输出”对话框可以为对象得分、区分测量、迭代历史、原始变量和转换后变量的相关性、选定的变量的类别量化和选定的变量的描述统计生成表。

图片 6-7  
“输出”对话框



**对象得分。**显示对象得分，包括质量、惯量和贡献，并且有以下选项：

- **包括类别。**显示选定的分析变量的类别指示符。
- **标注对象得分。**从指定为标签变量的变量列表中，可选择一個用以标注对象的变量。

**区分测量。**显示每个变量和每一维的区分测量。

**迭代历史记录。**对于每次迭代，显示偏差考虑情况以及偏差考虑情况中的损失和增加项。

**初始变量的相关性。**显示初始变量的相关性矩阵以及该矩阵的特征值。

**转换变量的相关性。**显示已转换的（最优尺度）变量的相关性矩阵以及该矩阵的特征值。

**类别量化和贡献。**针对选定的变量的每一维给出类别量化（坐标），包括质量、惯量和贡献。

注意：坐标和贡献（包括质量和惯量）显示在枢轴表输出的不同层上，缺省显示坐标。要显示贡献，在表上激活（双击）“层”下拉列表并从中选择“贡献”。

**描述统计。**显示选定的变量的频率、缺失值的数量以及众数。

## 多重对应分析：保存

使用“保存”对话框可以将离散化数据、对象得分和转换后的值保存到当前会话中的外部 IBM® SPSS® Statistics 数据文件或数据集中。还可以将转换后的值和对象得分保存到活动数据集中。

- 数据集在当前会话期间可用，但在后续会话期间不可用，除非显式将其保存为数据文件。数据集名称必须符合变量命名规则。
- 文件名或数据集名称对于保存的每种数据类型都必须各不相同。
- 如果将对象得分或已转换的值保存到活动数据集中，则可以指定多名义维数的数量。

图片 6-8  
保存对话框



## 多重对应分析：对象图

使用“对象图”对话框可以指定所要的图类型以及要绘制的变量

图片 6-9  
“对象图”对话框



**对象点。**显示对象点的图。

**对象和质心（双标图）。**对象点用变量质心绘制。

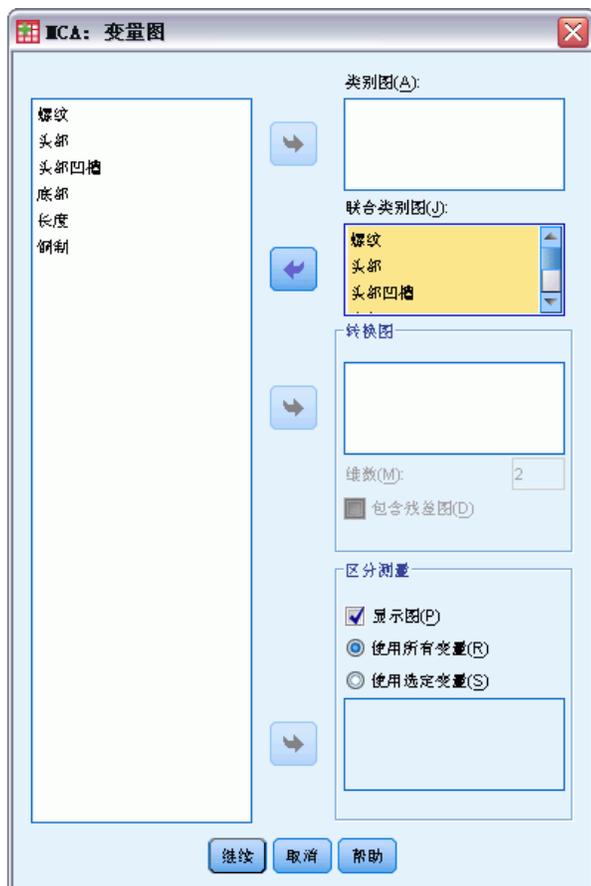
**双标图变量。**可选择将所有变量用于双标图或选择一个子集。

**标签对象。**可选择用选定的变量的类别（可在“选项”对话框中选择类别指示符值或值标签）或其个案号标注对象。如果选择变量，则为每个变量生成一个图。

## 多重对应分析： 变量图

使用“变量图”对话框可以指定所要的图类型和要绘制的变量。

图片 6-10  
“变量图”对话框



**类别图。**对于选定的每个变量绘制质心坐标图。类别在特定类别中的对象的质心中。

**联合类别图。**这是每个选定的变量的质心坐标的单个图。

**变换图。**显示最优类别量化与类别指示符的比较图。可指定所要的维数；将为每一维分别生成一个图。还可以选择显示每个选定的变量的残差图。

**区分测量。**为选定的变量生成区分测量的单个图。

## MULTIPLE CORRESPONDENCE 命令附加功能

如果将选中的内容粘贴到语法窗口中，并编辑生成的 MULTIPLE CORRESPONDENCE 命令语法，则可以定制多重对应分析。使用命令语法语言还可以：

- 将已转换的变量、对象得分和近似值保存到活动数据集时（使用 SAVE 子命令），为其指定根名称。
- 分别为每个图指定最大标签长度（使用 PLOT 子命令）。
- 为残差图指定单独的变量列表（使用 PLOT 子命令）。

请参见命令语法参考以获取完整的语法信息。

## 多维尺度 (PROXSCAL)

多维尺度尝试查找对象间的一组近似性测量值的结构。该过程是通过将观察值分配到概念低维度空间中的特定位置实现的，这样使空间中的点之间的距离尽可能与给定的（不）相似性相匹配。结果是一个该低维空间中的对象的最小平方表示形式，这在许多个案中将帮助您进一步理解您的数据。

**示例。**多维尺度对于确定感知关系很有用。例如，在考虑产品形象时，可进行调查以获取描述您的产品与竞争对手产品的感知相似性（或近似性）的数据集。通过使用这些近似性和自变量（如价格），您可尝试确定哪些变量对于人们如何看待这些产品至关重要，以对产品形象做出相应的调整。

**统计量和图。**迭代历史记录、应力度量标准、应力分解、公共空间坐标、最终配置中的对象距离、私有空间权重、私有空间、转换后的近似值、转换后的自变量、应力图、公共空间散点图、私有空间权重散点图、私有空间散点图、变换图、Shepard 残差图，以及自变量变换图。

**数据。**数据可以以相似性矩阵或转换为相似性矩阵的变量的形式提供。矩阵的格式可以为单列，也可以为多列。近似值可以在比率、区间、有序或样条尺度级别上进行处理。

**假设。**至少必须指定三个变量。维数不能超过对象数减一。如果与多随机起点组合，则省略维数减少。如果仅指定一个源，则所有模型都等同于恒等模型；因此，该分析缺省设置为恒等模型。

**相关过程。**调整所有数值级别变量对应于标准多维尺度分析。

## 获取多维尺度

- ▶ 从菜单中选择：  
分析 > 标度 > 多维尺度 (PROXSCAL)...

这将打开“数据格式”对话框。

图片 7-1  
“数据格式”对话框



- ▶ 指定数据的格式：

**数据格式。**指定您的数据是由相似性测量值组成，还是您要从数据中创建近似值。

**源的数目。**如果数据是近似值，则指定相似性测量值是单源还是多源。

**一个源。**如果有一个近似值源，则指定数据集是格式化为跨列矩阵中的近似值，还是单列中的近似值（单列中有两个单独的变量用于标识每个近似值的行和列）。

- **近似值在跨列矩阵中。**相似性矩阵跨与对象同数量的列分布。这生成“跨列矩阵”对话框中的近似值。
- **近似值位于单列中。**相似性矩阵拼并到单个列或变量中。需要两个附加变量，以识别每个单元的行和列。这生成“单列”对话框中的近似值。

**多个源。**如果有多个近似值源，则指定数据集是格式化为跨列堆积矩阵中的近似值（在多个列中，每列一个源）还是单列中的近似值。

- **近似值位于跨列的堆积矩阵中。**相似性矩阵跨与对象同数量的列分布，并跨行相互堆积（行数等于对象的数量乘以源的数量）。这生成“跨列矩阵”对话框中的近似值。
- **近似值位于多列中，每列一个源。**相似性矩阵拼并到多个列或变量中。需要两个附加变量，以识别每个单元的行和列。这生成“多列”对话框中的近似值。
- **近似值累积在单列中。**相似性矩阵拼并到单个列或变量中。需要三个附加变量，以识别行、列和每个单元的源。这生成“单列”对话框中的近似值。

- ▶ 单击定义。

## 矩阵中的跨列近似值

如果在“数据格式”对话框中为单个源或多个源选择了矩阵数据模型中的近似值，则会出现以下主对话框：

图片 7-2  
“矩阵中的跨列近似值”对话框



- ▶ 选择三个或更多近似值变量。（确保列表中的变量顺序与近似值的列顺序一致。）
- ▶ 或者，选择与近似值变量数相等的权重变量数。（确保权重的顺序与它们所加权的近似值的顺序一致。）
- ▶ 或者，如果存在多个源，则选择源变量。（每个近似值变量中的个案数应等于近似值变量的数目乘以源的数目。）

此外，还可以定义多维尺度模型、对公共空间施加约束、设置收敛标准、指定要使用的初始配置以及选择图和输出。

## 多列中的近似值

如果在“数据格式”对话框中为多个源选择了多列模型，则会出现以下主对话框：

图片 7-3  
“多列中的近似值”对话框



- ▶ 选择两个或更多近似值变量。（每个变量被假设为来自单独源的近似值矩阵。）
- ▶ 选择一个行变量以定义每个近似值变量中近似值的行位置。
- ▶ 选择一个列变量以定义每个近似值变量中近似值的列位置。（相似性矩阵中没有得到行/列指定项的单元被视为缺失。）
- ▶ 或者，选择与近似值变量数相等的权重变量数。

此外，还可以定义多维尺度模型、对公共空间施加约束、设置收敛标准、指定要使用的初始配置以及选择图和输出。

## 单列中的近似值

如果在“数据格式”对话框中为单个源或多个源选择了单列模型，则会出现以下主对话框：

图片 7-4  
“单列中的近似值”对话框



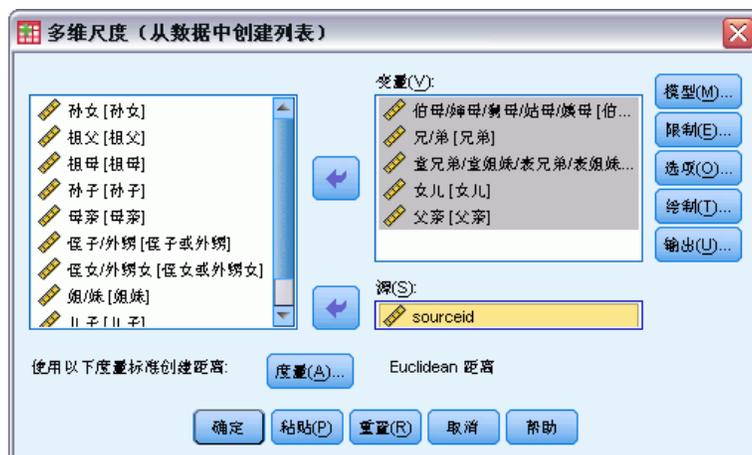
- ▶ 选择一个近似值变量。（它被视作一个或多个近似值矩阵。）
- ▶ 选择一个行变量以定义近似值在近似值变量中的行位置。
- ▶ 选择一个列变量以定义近似值在近似值变量中的列位置。
- ▶ 如果存在多个源，请选择源变量。（对于每个源，相似性矩阵中没有得到行/列指定项的单元被视为缺失。）
- ▶ 或者，选择一个权重变量。

此外，还可以定义多维尺度模型、对公共空间施加约束、设置收敛标准、指定要使用的初始配置以及选择图和输出。

## 从数据中创建近似值

如果在“数据格式”对话框中选择从数据中创建近似值，则会出现以下主对话框：

图片 7-5  
“从数据中创建近似值”对话框



- ▶ 如果创建两个变量之间的距离（请参见“从数据中创建度量”对话框），则选择至少三个变量。这些变量将用于创建相似性矩阵（如果有多个源，则创建多个矩阵）。如果创建个案之间的距离，则只需要一个变量。
- ▶ 如果存在多个源，请选择源变量。
- ▶ 或者，选择一个用于创建近似值的测量。

此外，还可以定义多维尺度模型、对公共空间施加约束、设置收敛标准、指定要使用的初始配置以及选择图和输出。

## 从数据中创建度量

图片 7-6  
“从数据中创建度量”对话框



多维尺度使用不相似性数据创建尺度分析解。如果您的数据为多变量数据（度量到的变量的值），就必须创建不相似性数据才能计算多维尺度解。可以指定从数据创建非相似性测量的详细信息。

**度量。**允许您指定进行分析的非相似性测量。从与您的数据类型相对应的“度量”组选择一个选项，然后从与该类度量相对应的下拉列表选择一种度量。可以使用的选项有：

- **区间。**欧氏距离、平方 Euclidean 距离、Chebychev、区组、Minkowski 或定制。
- **计数。**卡方统计量测量或 Phi 平方统计量测量。
- **二元。**欧氏距离、平方 Euclidean 距离、尺度差分、模式差分、方差或 Lance 和 Williams。

**创建距离矩阵。**可以选择要分析的单位。选项有“变量之间”或“个案之间”。

**转换值。**在某些情况下（例如，当用相差很大的尺度测量变量时），您会希望先将值标准化，然后再计算近似值（对二分类数据不可用）。从“标准化”下拉列表选择一种标准化方法（如果不需要标准化，请选择无）。

## 定义多维尺度模型

使用“模型”对话框可以指定尺度模型、其最大和最小维数、相似性矩阵的结构、要用于近似值的转换，以及近似值是在每个源中单独转换还是在源上无条件转换。

图片 7-7  
“模型”对话框



**尺度模型。**选择下列选项之一：

- **恒等。**所有源都具有相同的配置。
- **加权欧几里得。**该模型是个体差异模型。每个源都具有一个私有空间，在该空间中公共空间的每个维都有不同的加权。
- **广义欧几里得。**该模型是个体差异模型。每个源都具有私有空间（等于公共空间的一个旋转度，后跟各个维数的不同加权）。
- **减少的秩。**此模型是广义欧几里得模型，您可以为其指定私有空间的秩。您必须指定大于等于 1 且小于最大维数的秩。

**形状。**指定近似值应从相似性矩阵的下三角部分获取还是从上三角部分获取。可指定使用满矩阵，在此情况下，将分析上三角部分和下三角部分的加权和。虽然将只用到指定的部分，但是任何情况下，都应该指定完整的矩阵，包括对角线。

**近似值。**指定相似性矩阵是包含相似性测量还是非相似性测量。

**近似值转换。**选择下列选项之一：

- **比例。**转换后的近似值与原始近似值成比例。这只对于正值近似值是允许的。
- **区间。**转换后的近似值与原始近似值成比例，外加一个截距项。该截距确保所有转换后的近似值都是值。

- **有序**。转换后的近似值与初始近似值具有相同的顺序。指定结近似值应保持打结还是允许解开。
- **样条**。转换后的近似值是初始近似值的平滑非递减的分段多项式转换。指定多项式的次数以及内部结点数。

**应用转换**。指定是只有每个源中的近似值相互比较还是在源上进行无条件比较。

**维**。缺省情况下，在两个维中（最小值 = 2，最大值 = 2）计算解。可选择最小为 1 最大为对象数减一的整数（只要最小值小于等于最大值）。该过程计算最大维数中的解，然后逐步减去维数，直至达到最低值。

## 多维尺度：约束

使用“约束”对话框可以对公共空间施加约束。

图片 7-8  
“限制”对话框



**公共空间的约束**。指定所要的约束类型。

- **无约束**。对公共空间不施加约束。
- **某些固定坐标**。选定的第一个变量包含对象在第一维上的坐标，第二个变量对应于在第二维上的坐标，依此类推。缺失值表示维上的坐标是自由的。选定的变量的数量必须等于所请求的最大维数。
- **自变量的线性组合**。公共空间被限制为选定的变量的线性组合。

**约束变量**。选择定义公共空间的约束的变量。如果指定线性组合，则为约束变量指定区间、名义、有序或样条转换。无论哪种情况，每个变量的个案数必须等于对象数。

## 多维尺度：选项

使用“选项”对话框可以选择初始配置样式、指定迭代和收敛性准则，以及选择标准更新或不严格的更新。

图片 7-9  
“选项”对话框



**初始配置。**选择下列选项之一：

- **单形体。**在最大维上，对象相互之间等距离放置。执行一次迭代以改进这种高维配置，随后执行维数减少操作以获取具有您在“模型”对话框中指定的最大维数的初始配置。
- **Torgerson。**一个经典尺度解用作初始配置。
- **单随机起点。**随机选择配置。
- **多随机起点。**随机选择多个配置，并且标准化初始应力最低的配置用作初始配置。
- **定制。**选择包含您自己的初始配置的坐标的变量。选定的变量的数量应等于指定的最大维数，并且第一个变量与维 1 上的坐标对应，第二个变量与维 2 上的坐标对应，依此类推。每个变量中的个案数应等于对象数。

**迭代标准。**指定迭代标准值。

- **应力收敛。**当连续标准化初始应力值之间的差小于此处指定的数字（该数字必须介于 0.0 到 1.0 之间）时，算法将停止迭代。
- **最小应力。**当标准化初始应力小于此处指定的数字（该数字必须介于 0.0 到 1.0 之间）时，算法将停止。

- **最大迭代次数。**算法将执行指定的迭代次数（除非已先满足了上述某个条件）。
- **使用不严格的更新。**不严格的更新将加速算法；这些更新不能与除恒等模型之外的其他模型一起使用，也不能与约束一起使用。

## 多维尺度：图，版本 1

使用“图”对话框可以指定将生成哪些图。如果您使用“多列中的近似值”数据格式，则会显示下面的“图”对话框。对于私有空间权重、初始近似值对比转换后的近似值和转换后的近似值对比距离图，指定应为其生成图的源。可用源的列表为主对话框中的近似值变量列表。

图片 7-10  
“图”对话框，版本 1



**应力。**生成标准化初始应力与维数的关系图。仅当维数的最大值大于维数的最小值时，才能生成此图。

**公共空间。**显示公共空间坐标的散点图矩阵。

**私有空间。**对于每个源，私有空间的坐标显示在散点图矩阵中。仅当在“模型”对话框中指定其中某个个体差异模型时才可能具有此行为。

**私有空间权重。**生成私有空间权重的散点图。仅当在“模型”对话框中指定其中某个个体差异模型时才可能具有此行为。对于加权欧几里得模型，权重显示在图中，每个轴上有一维。对于广义欧几里得模型，每一维生成一个图，指示该维的旋转和权重。减少的秩模型生成与广义欧几里得模型相同的图，但是减少私有空间的维数。

**初始近似值与转换后的近似值。**生成初始近似值与转换后的近似值的关系图。

**转换后的近似值对比距离。**绘制转换后的近似值与距离。

**转换后的自变量。**为自变量生成变换图。

**变量和维数相关性。**显示自变量和公共空间维数之间的相关性的图。

## 多维尺度：图，版本 2

使用“图”对话框可以指定将生成哪些图。如果您的数据格式是除“多列中的近似值”之外的任何其他格式，则会显示下面的“图”对话框。对于私有空间权重、初始近似值对比转换后的近似值和转换后的近似值对比距离图，指定应为其生成图的源。输入的源号必须是在主对话框中指定的源变量的值，并且范围必须是从 1 到源的数目。

图片 7-11

“图”对话框，版本 2



## 多维尺度：输出

使用“输出”对话框可以控制显示的输出的数量并将其中的一部分保存到单独的文件中。

图片 7-12  
“输出”对话框



**显示。**选择以下一项或多项进行显示：

- **公共空间坐标。**显示公共空间的坐标。
- **私有空间坐标。**只有在模型不是恒等模型的情况下，才会显示私有空间的坐标。
- **私有空间权重。**只有在指定了个体差异模型之一的情况下，才会显示私有空间权重。根据模型的不同，空间权重被分解为旋转权重和维权重，这两种权重也会显示。
- **距离。**显示配置中对象之间的距离。
- **转换后的近似值。**显示配置中对象之间的转换后的近似值。
- **输入数据。**包括原始近似值，如果存在数据权重、初始配置和自变量的固定坐标，则还包括这些项。
- **随机起点的应力。**显示每个随机起点的随机数种子以及标准化初始应力值。
- **迭代历史记录。**显示主要算法的迭代历史记录。
- **多应力度量标准。**显示不同的应力值。该表包含标准化初始应力值、Stress-I 值、Stress-II 值、S-Stress 值、离散考虑情况 (DAF) 值和同余 Tucker' s 系数值。
- **应力分解。**显示对象和源的最终标准化初始应力的分解，包括每个对象的平均值和每个源的平均值。
- **转换后的自变量。**如果选择了线性组合约束，则会显示转换后的自变量和对应的回归权重。
- **变量和维数相关性。**如果选择了线性组合约束，则会显示自变量和公共空间维数之间的相关性。

**保存到新文件。**可将公共空间坐标、私有空间权重、距离、转换后的近似值以及转换后的自变量保存到单独的 IBM® SPSS® Statistics 数据文件中。

## PROXSCAL 命令附加功能

如果将选中的内容粘贴到语法窗口中，并编辑生成的 PROXSCAL 命令语法，则可以定制近似值分析的多维尺度。使用命令语法语言还可以：

- 为转换和残差图指定单独的变量列表（使用 PLOT 子命令）。
- 为私有空间权重、转换和残差图指定单独的源列表（使用 PLOT 子命令）。
- 指定要显示的自变量变换图的子集（使用 PLOT 子命令）。

请参见命令语法参考以获取完整的语法信息。

# 多维展开 (PREFSCAL)

“多维展开”过程尝试查找允许您直观地检查两组对象之间关系的公共定量尺度。

**示例。**您已经让 21 个人按照各自的偏好对 15 种早餐食品进行排序，顺序为 1 到 15。使用多维展开，您可以确定他们从以下两个主要方面区分早餐食品：软面包和硬面包之间，以及增肥食品和不增肥食品之间。

或者，您让一组驾驶员就 10 种属性对 26 款车型进行评分，分数为 6 分制，从 1=“完全不真实”到 6=“非常真实”。对个体取平均值后，得到的值被视为相似性。通过使用多维展开，您将找到相似模型的聚类以及与这些模型最密切相关的属性。

**统计量和图。**多维展开过程可生成迭代历史记录、应力度量标准、应力分解、公共空间坐标、最终配置中的对象距离、私有空间权重、私有空间、转换后的近似值、应力图、公共空间散点图、私有空间权重散点图、私有空间散点图、变换图和 Shepard 残差图。

**数据。**数据以矩形相似性矩阵的形式提供。每个列被视为一个单独的列对象。相似性矩阵的每一行被视为独立的行对象。当存在多个近似值源时，矩阵堆积。

**假设。**至少必须指定两个变量。解中的维数不能超过对象数减一。如果仅指定一个源，则所有模型都等同于恒等模型；因此，该分析缺省设置为恒等模型。

## 获取多维展开

- ▶ 从菜单中选择：  
分析 > 标度 > 多维展开 (PREFSCAL)...

图片 8-1  
“多维展开”主对话框



- ▶ 选择标识矩形相似性矩阵中的列的两个或更多变量。每个变量表示独立的列对象。

- ▶ 或者，选择与列对象变量的数目相等的权重变量数。权重变量的顺序应与它们加权的列对象的顺序一致。
- ▶ 或者，选择一个行变量。此变量的值（或值标签）用于在输出中标记行对象。
- ▶ 如果存在多个源，也可以选择源变量。数据文件中的个案数应等于行对象的数目乘以源的数目。

此外，还可以定义多维展开模型、对公共空间施加约束、设置收敛标准、指定要使用的初始配置以及选择图和输出。

## 定义多维展开模型

使用“模型”对话框可以指定尺度模型、其最大和最小维数、相似性矩阵的结构、用于近似值的转换，以及近似值是在行上进行有条件转换、在源上进行有条件转换，还是在源上进行无条件转换。

图片 8-2  
“模型”对话框



**尺度模型。**选择下列选项之一：

- **恒等。**所有源都具有相同的配置。
- **加权欧几里得。**该模型是个体差异模型。每个源都具有一个私有空间，在该空间中公共空间的每个维都有不同的加权。
- **广义欧几里得。**该模型是个体差异模型。每个源都具有私有空间（等于公共空间的一个旋转度，后跟各个维数的不同加权）。

**近似值。**指定相似性矩阵是包含相似性测量还是非相似性测量。

**维。**缺省情况下，在两个维中（最小值 = 2，最大值 = 2）计算解。只要最小值小于等于最大值，就可以选择最小为 1 最大为对象数减 1 的整数。该过程计算最大维数中的解，然后逐步减去维数，直至达到最低值。

**近似值转换。**选择下列选项之一：

- **无。**不转换近似值。或者选择**包括截距**，在这种情况下，近似值可通过一个常数项移位。
- **线性。**转换后的近似值与原始近似值成比例；即转换函数估算一个斜率，并且截距修正为 0。这又称为比率转换。或者选择**包括截距**，在这种情况下，近似值可通过一个常数项移位。这又称为区间转换。
- **样条。**转换后的近似值是初始近似值的平滑非递减的分段多项式转换。可指定多项式的次数和内部结点数。或者选择**包括截距**，在这种情况下，近似值可通过一个常数项移位。
- **平滑。**转换后的近似值与原始近似值有着相同的顺序，包括将后继值之间的差值考虑在内的约束。结果为“平滑有序”转换。可以指定结近似值应保持打结还是允许解开。
- **有序。**转换后的近似值与初始近似值具有相同的顺序。可以指定结近似值应保持打结还是允许解开。

**应用转换。**指定是只有每行中的近似值相互比较，还是只有每个源中的近似值相互比较，或者是在行或源上无条件进行比较；即，转换是逐行执行、逐源执行，还是一次性对所有近似值执行。

## 多维展开：约束

使用“约束”对话框可以对公共空间施加约束。

图片 8-3  
“限制”对话框



**公共空间的约束。**可选择固定行和/或列对象在公共空间中的坐标。

**行/列约束变量。**选择包含约束的文件，然后选择定义对公共空间的约束的变量。选定的第一个变量包含在第一维上的对象的坐标，第二个变量对应于在第二维上的坐标，依此类推。缺失值表示维上的坐标是自由的。选定的变量的数量必须等于所请求的最大维数。每个变量的个案数必须等于对象数。

## 多维展开：选项

使用“选项”对话框可以选择初始配置样式，指定迭代和收敛标准以及为应力设置惩罚项。

图片 8-4  
“选项”对话框



**初始配置。**选择下列选项之一：

- **经典。**矩形相似性矩阵用于补充完整对称 MDS 矩阵的内部块（行之间和列之间的值）。完整矩阵一旦形成，经典尺度解将用作初始配置。可使用三角不等性或 Spearman 距离通过插补法填充内部块。
- **Ross-Cliff。**Ross-Cliff 起点将双中心和正方形相似性矩阵的奇异值分解的结果用作行和列对象的初始值。
- **对应。**对应起点使用对反向数据（相似性而非不相似性）的对应分析的结果，且对行和列的得分进行对称的标准化。
- **质心。**该过程是通过使用特征值分解将行对象定位在配置中来启动的。然后列对象定位在指定选择项的质心。对于选择项的数目，请指定介于 1 和近似值变量的数目之间的正整数。
- **多随机起点。**为多个随机选择的初始配置计算解，并且惩罚强度最低的解作为最佳解显示。
- **定制。**可选择包含您自己的初始配置的坐标的变量。选定的变量的数量应等于指定的最大维数，并且第一个变量与维 1 上的坐标对应，第二个变量与维 2 上的坐标对应，依此类推。每个变量中的个案数应等于行和列对象合并的数目。行和列坐标应该是堆积的，列坐标在行坐标后面。

**迭代标准。**指定迭代标准值。

- **应力收敛。**当连续惩罚强度值中的相对差值小于此处指定的数字（必须为非负）时，算法将停止迭代。
- **最小应力。**当惩罚强度小于此处指定的数字（必须为非负）时，算法将停止。
- **最大迭代次数。**算法将执行此处指定的迭代次数（除非已先满足了上述某一条件）。

**惩罚项。**该算法尝试最小化惩罚强度，该惩罚强度是一个拟合优度测量，等于 Kruskal 的 Stress-I 和基于转换后的近似值的变异系数的惩罚项的乘积。使用这些控件可以设置惩罚项的强度和范围。

- **强度。**强度参数的值越小，惩罚越强。请指定介于 0.0 和 1.0 之间的值。
- **范围。**此参数设置惩罚有效的起始时刻。如果设置为 0.0，则惩罚无效。增加该值将使算法搜索转换后的近似值之间变差更大的解。请指定一个非负值。

## 多维展开： 图

使用“图”对话框可以指定将生成哪些图。

图片 8-5  
“图”对话框



图。以下图可用：

- **多起点。**显示惩罚强度的堆积直方图，该直方图同时显示应力和惩罚。
- **初始公共空间。**显示初始公共空间的坐标的散点图矩阵。
- **每维的应力。**生成惩罚强度对比维数的线图。仅当维数的最大值大于维数的最小值时，才能生成此图。
- **最终公共空间。**显示公共空间坐标的散点图矩阵。
- **空间权重。**生成私有空间权重的散点图。仅当在“模型”对话框中指定其中某个个体差异模型时才可能具有此行为。对于加权欧几里得模型，所有源的权重显示在一张图中，并且每个轴上一维。对于广义欧几里得模型，每一维生成一个图，指示该维对每个源的旋转和权重。

- **私有空间**。显示每个源的私有空间坐标的散点图矩阵。仅当在“模型”对话框中指定其中某个个体差异模型时才可能具有此行为。
- **转换图**。生成原始近似值对比转换后的近似值的散点图。根据应用转换的方式，将为每行或每个源分配不同的颜色。无条件转换将生成单色。
- **Shepard 图**。原始近似值对比转换后的近似值和距离。距离由点表示，转换后的近似值由线表示。根据应用转换的方式，将为每行或每个源生成不同的线。无条件转换将生成一条线。
- **拟合散点图**。显示转换后的近似值对比距离的散点图。如果指定多个源，则会为每个源指定单独的颜色。
- **残差图**。显示转换后的近似值对比残差（转换后的近似值减去距离）的散点图。如果指定多个源，则会为每个源指定单独的颜色。

**行对象样式**。这些样式使您可以进一步控制行对象在图中的显示方式。可选颜色变量的值用于循环选择所有颜色。可选标记变量的值用于循环选择所有可能的标记。

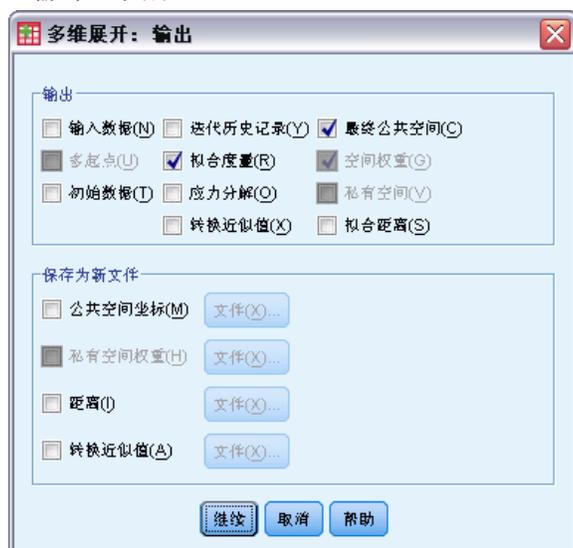
**源图**。对于私有空间、拟合散点图和残差图（如果转换由源应用，则还包括变换图一和 Shepard 图一），可指定应为其生成图的源。输入的源号必须是在主对话框中指定的源变量的值，范围为从 1 到源的数目。

**行图**。如果转换由行应用，则对于变换图和 Shepard 图，可指定应为其生成图的行。输入的行号必须在从 1 到行数的范围之内。

## 多维展开：输出

使用“输出”对话框可以控制显示的输出的数量并将其中的一部分保存到单独的文件中。

图片 8-6  
“输出”对话框



**显示**。选择以下一项或多项进行显示：

- **输入数据**。包括原始近似值，如果存在数据权重、初始配置和固定坐标，则还包括这些项。

- **多起点**。显示每个随机起点的随机数种子以及惩罚强度值。
- **初始数据**。显示初始公共空间的坐标。
- **迭代历史记录**。显示主要算法的迭代历史记录。
- **拟合测量**。显示不同的测量。该表包含多个拟合优度、拟合劣度、相关性、方差和非退化度量。
- **应力分解**。显示惩罚强度的对象、行和源分解，包括行、列和源均值以及标准差。
- **转换后的近似值**。显示转换后的近似值。
- **最终公共空间**。显示公共空间的坐标。
- **空间权重**。显示私有空间权重。仅当指定其中某个个体差异模型时，该选项才可用。根据模型的不同，空间权重被分解为旋转权重和维权重，这两种权重也会显示。
- **私有空间**。显示私有空间的坐标。仅当指定其中某个个体差异模型时，该选项才可用。
- **拟合距离**。显示配置中对象之间的距离。

**保存到新文件**。可将公共空间坐标、私有空间权重、距离以及转换后的近似值保存到不同的 IBM® SPSS® Statistics 数据文件中。

## PREFSCAL 命令附加功能

如果将选中的内容粘贴到语法窗口中，并编辑生成的 PREFSCAL 命令语法，则可以定制近似值分析的多维展开。使用命令语法语言还可以：

- 在有多个源可用的情况下，为私有空间、拟合散点图和残差图指定多源列表，——并且在矩阵有条件转换的情况下，还可以为变换图和 Shepard 图——指定多源列表（使用 PLOT 子命令）。
- 在行有条件转换的情况下，为变换图和 Shepard 图指定多行列表（使用 PLOT 子命令）。
- 指定多行而不是行 ID 变量（使用 INPUT 子命令）。
- 指定很多源而不是源 ID 变量（使用 INPUT 子命令）。

请参见命令语法参考以获取完整的语法信息。

# 部分 II:

## 示例

# 分类回归

使用最优尺度进行分类回归的目标是描述一个响应变量和一组预测变量之间的关系。通过量化这种关系，可以为任意预测变量组合预测响应值。

在本章中，将使用两个示例来阐释最优尺度回归中涉及的分析。第一个示例使用一个小型数据集来阐释基本概念。第二个示例使用一个大型变量集和观察值集来阐释实际应用。

## 示例：地毯清洁用品数据

在一个常用示例 (Green 和 Wind, 1973) 中，一家公司非常重视一种新型地毯清洁用品的市场营销，希望检验以下五种因素对消费者偏好的影响—包装设计、品牌名称、价格、优秀家用品标志和退货保证。包装设计有三个因子水平，每个因子水平因刷体位置而不同；有三个品牌名称 (K2R、Glory 和 Bissell)；有三个价格水平；最后两个因素各有两个级别 (有或无)。下表显示了在地毯清洁用品调查中所使用的变量以及这些变量的变量标签和变量值。

表 9-1  
地毯清洁用品调查中的解释性变量

变量名	变量标签	值标签
package	Package design	A*、B*、C*
brand	Brand name	K2R、Glory、Bissell
price	Price	\$1.19, \$1.39, \$1.59
seal	Good Housekeeping seal	No、yes
money	Money-back guarantee	No、yes

十名消费者对这些因素所定义的 22 个特征进行了排序。变量优选包含对每个特征的平均等级的排序。低排序与高偏好相对应。此变量反映了对每个特征的偏好的总体度量。下面，您将使用分类回归研究这五个因素与偏好的相关程度。此数据集可在 carpet.sav 中找到。有关详细信息，请参阅附录 A 中的样本文件中的 IBM SPSS Categories 21。

## 标准线性回归分析

- ▶ 要生成标准线性回归输出，请从菜单中选择：  
分析 > 回归 > 线性...

注意：此功能需要“Statistics Base”选项。

图片 9-1  
“线性回归”对话框



- ▶ 选择优选作为因变量。
- ▶ 选择包装设计到退款保证作为自变量。
- ▶ 单击图。

图片 9-2  
“图”对话框



- ▶ 选择 \*ZRESID 作为 y 轴变量。
- ▶ 选择 \*ZPRED 作为 x 轴变量。
- ▶ 单击继续。

- ▶ 在“线性回归”对话框中单击保存。

图片 9-3  
“保存”对话框



- ▶ 在“残差”组中选择标准化。
- ▶ 单击继续。
- ▶ 在“线性回归”对话框中单击确定。

## 模型摘要

图片 9-4  
标准线性回归的模型摘要

模型	R	R 方	调整的 R 方	估计的标准误差
1	.841 <sup>a</sup>	.707	.615	3.99610

a. 预测变量: (常量), 退款保证, 价格, 优秀家用品奖章, 品牌名称, 包装设计。

线性回归是描述此问题中关系的标准方法。指示回归模型与数据的拟合度的最常用度量是  $R^2$ 。此统计量表示由预测变量的加权组合所解释的响应变异量。 $R^2$  越接近于 1, 该模型的拟合度就最高。根据五个预测变量对优选进行回归, 会得到  $R^2$  值 0.707, 这表示线性回归中的预测变量解释了偏好等级中大约 71% 的变异。

## 系数

下表显示了标准化系数。系数的符号指示在某一预测变量增加（而所有其他预测变量均保持不变）时预测响应是增加还是减少。对于分类数据，类别编码确定了预测变量增加的含义。例如，退款保证、包装设计或优秀家庭用品奖章增加将会导致预测偏好等级减少。对于退款保证，将无退款保证和退款保证分别编码为 1 和 2。退款保证变量增加对应于无退款保证变为退款保证。因此，增加退款保证会减少预测偏好等级，这与增加预测偏好相对应。

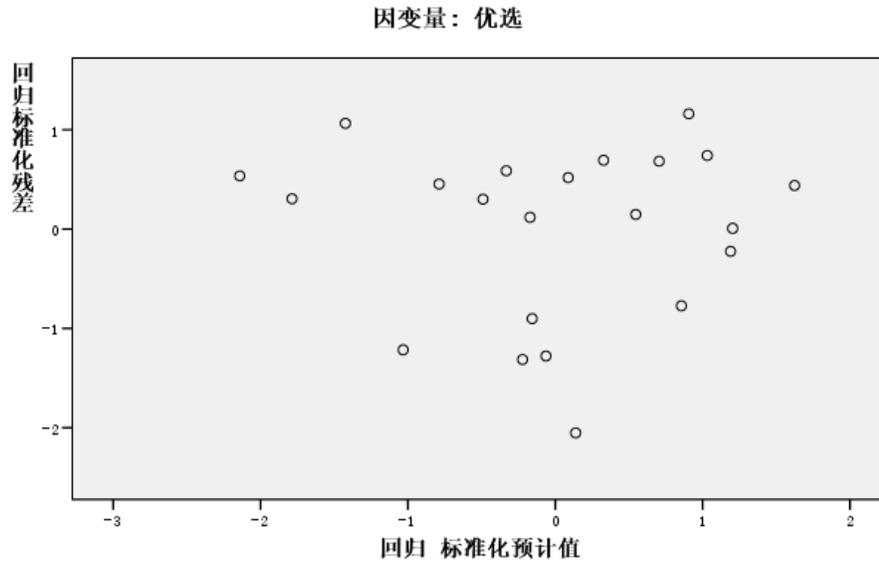
图片 9-5  
回归系数

模型		非标准化系数		标准化系数	t	显著性
		B	标准误	Beta		
1	(常量)	22.529	5.177		4.352	.000
	包装设计	-4.159	1.036	-.560	-4.015	.001
	品牌名称	.429	1.054	.056	.407	.689
	价格	2.703	1.009	.366	2.681	.016
	优秀家庭用品奖章	-4.314	1.780	-.330	-2.423	.028
	退款保证	-2.779	1.921	-.197	-1.447	.167

系数值反映了预测偏好等级的改变量。使用标准化系数时，解释基于变量的标准差。每个系数表示根据某一预测变量中的一个标准差变化（而所有其他预测变量均保持不变）而产生的预测响应变化的标准差数量。例如，变量品牌名称中的一个标准差变化会导致预测偏好增加 0.056 个标准差。变量优选的标准差为 6.44，因此优选的增量为  $0.056 \times 6.44 = 0.361$ 。变量包装设计中的变化使预测偏好中产生的变化最大。

## 残差散点图

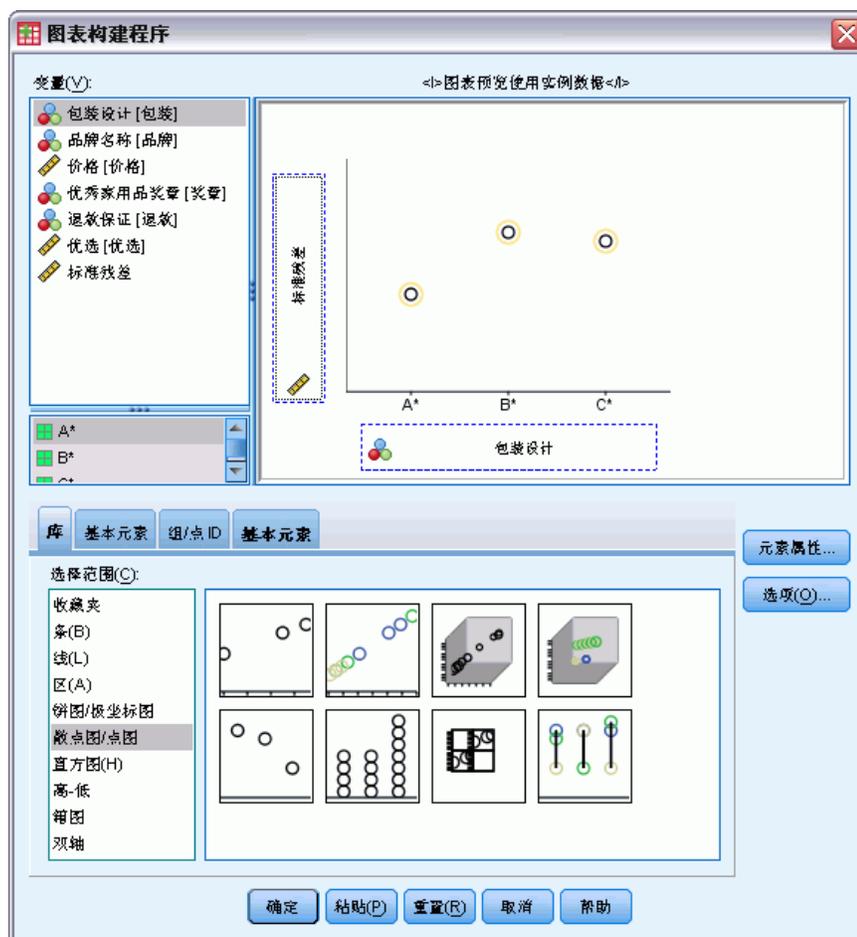
图片 9-6  
残差与预测值



根据标准化预测值绘制标准残差。如果该模型拟合度较高，则不应呈现出任何模式。此处可以看到有一个 U 形，其中高和低标准化预测值都具有正残差。接近 0 的标准预测值趋向于具有负残差。

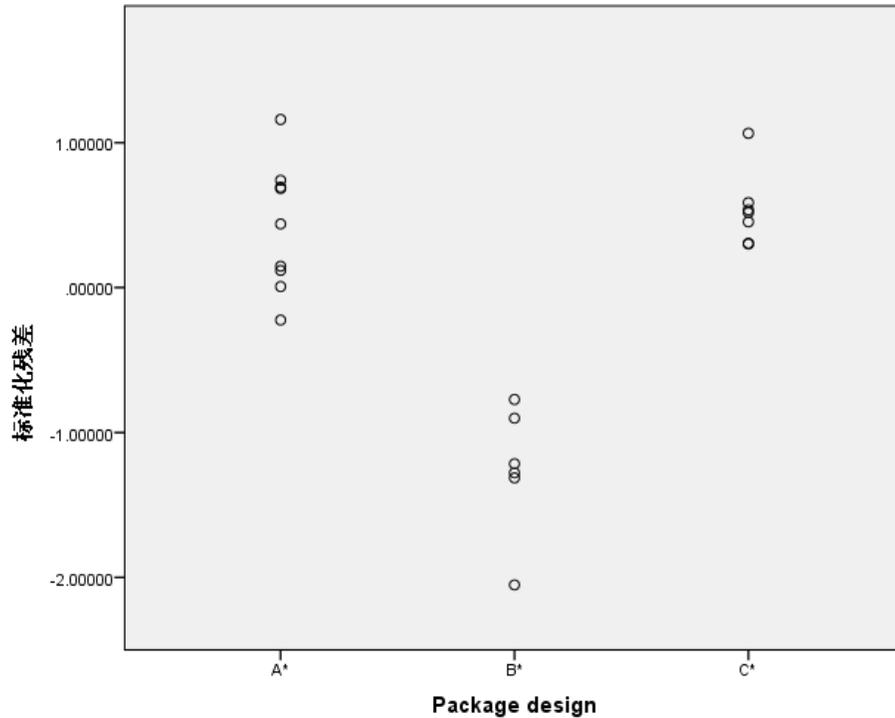
- ▶ 要按预测变量包装设计生成残差散点图，请从菜单中选择：  
图形 > 图表构建程序...

图片 9-7  
图表构建程序



- ▶ 选择散点图/点图库并选择“简单散点”。
- ▶ 选择标准化残差作为y轴变量，选择包装设计作为x轴变量。
- ▶ 单击确定。

图片 9-8  
残差与包装设计



在包装的标准化残差图中，会更明显地呈现出 U 形。“设计 B\*”的所有残差都是负值；而另两种设计只有一个残差为负值，而其他所有残差都是正值。由于线性回归模型将一个参数与每个变量拟合，因此不能使用此标准方法捕捉该关系。

## 分类回归分析

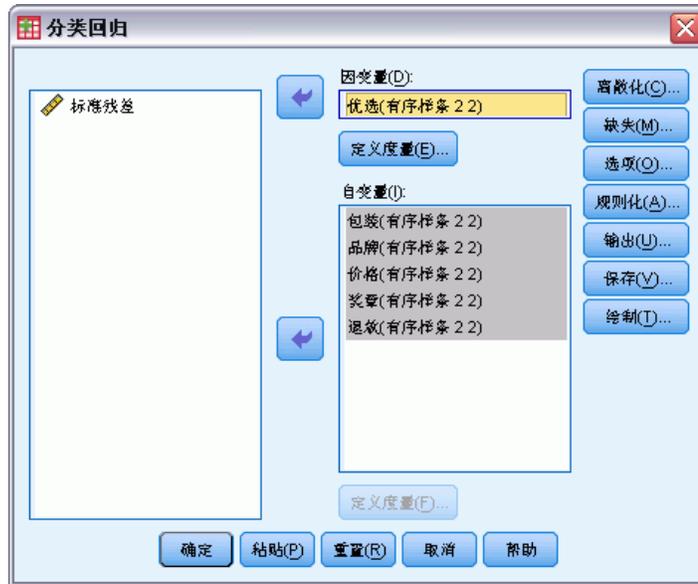
变量的分类性质以及优选和包装设计之间的非线性关系表明最优得分回归可能比标准回归具有更好的效果。残差图中显现的 U 形表明应对包装设计使用名义处理。所有其他预测变量都将在数值尺度级别处理。

响应变量可确保得到特别注意。由于要预测优选的值，因此要在量化过程中尽可能地恢复其类别的多个属性。使用有序或名义尺度级别会忽略响应类别之间的差异。但是，线性转换响应类别会保留类别差异。因此，通常首选采用数值来标定响应，下面就将使用该方法。

## 运行分析

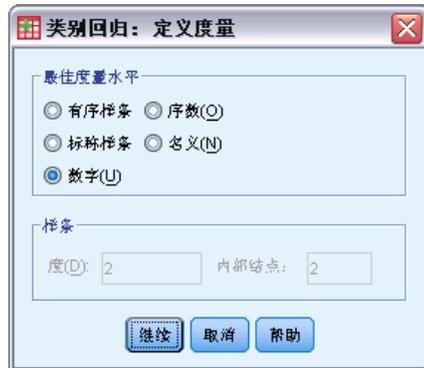
- ▶ 要运行“分类回归”分析，请从菜单中选择：  
分析 > 回归 > 最佳尺度 (CATREG)...

图片 9-9  
“分类回归”对话框



- ▶ 选择优选作为因变量。
- ▶ 选择包装设计到退款保证作为自变量。
- ▶ 选择优选，然后单击定义尺度。

图片 9-10  
“定义尺度”对话框



- ▶ 选择数值作为最佳度量水平。
- ▶ 单击继续。
- ▶ 选择包装设计，然后在“分类回归”对话框中单击定义尺度。

图片 9-11  
“定义尺度”对话框



- ▶ 选择名义作为最佳度量水平。
- ▶ 单击继续。
- ▶ 选择从品牌名称到退款保证的变量，然后在“分类回归”对话框中单击定义尺度。

图片 9-12  
“定义尺度”对话框



- ▶ 选择数值作为最佳度量水平。
- ▶ 单击继续。
- ▶ 在“分类回归”对话框中单击输出。

图片 9-13  
“输出”对话框



- ▶ 选择初始变量的相关性和转换变量的相关性。
- ▶ 取消选择 ANOVA。
- ▶ 单击继续。
- ▶ 在“分类回归”对话框中单击保存。

图片 9-14  
“保存”对话框



- ▶ 选择将残差保存到活动数据集。
- ▶ 选择“已转换的变量”组中的将已转换的变量保存到活动数据集。
- ▶ 单击继续。
- ▶ 在“分类回归”对话框中单击图。

图片 9-15  
“图”对话框



- ▶ 选择为包装设计和价格创建转换图。
- ▶ 单击继续。
- ▶ 在“分类回归”对话框中单击确定。

## 互相关性

预测变量之间的互相关性可用于识别回归中的多重共线性。高度相关的变量会导致回归估计值不稳定。但是，由于这些变量高度相关，因此从模型中忽略其中一个变量仅会最低限度地影响预测。可由被忽略变量解释的响应变异仍由其余相关变量解释。但是，零阶相关系数对离群值很敏感，而且还不能识别出因某一预测变量与其他预测变量组合之间的高度相关而产生的多重共线性。

图片 9-16  
初始预测变量相关系数

	包装设计	品牌名称	价格	优秀家用 品奖章	退款保证
包装设计	1.000	-.189	-.126	.081	.066
品牌名称	-.189	1.000	.065	-.042	-.034
价格	-.126	.065	1.000	.000	.000
优秀家用 品奖章	.081	-.042	.000	1.000	-.039
退款保证	.066	-.034	.000	-.039	1.000
维	1	2	3	4	5
特征值	1.291	1.038	.980	.905	.785

图片 9-17  
已转换的预测变量相关系数

	包装设计	品牌名称	价格	优秀家用 品奖章	退款保证
包装设计	1.000	-.156	-.089	.032	.102
品牌名称	-.156	1.000	.065	-.042	-.034
价格	-.089	.065	1.000	.000	.000
优秀家用 品奖章	.032	-.042	.000	1.000	-.039
退款保证	.102	-.034	.000	-.039	1.000
维	1	2	3	4	5
特征值	1.248	1.043	.983	.905	.821

上面显示了未转换和已转换预测变量的相关系数。所有值都接近于 0，这表示各变量之间的多重共线性无关。

您会看到只有涉及包装设计的相关系数发生了变化。因为所有其他预测变量都以数值处理，所以将为这些变量保留类别之间的差异和类别的顺序。因此，这些变量的相关系数不会发生变化。

## 模型拟合度和系数

分类回归过程会得到  $R^2$  值 0.948，这表明已转换偏好等级中大约 95% 的方差可由最优转换预测变量的回归来解释。相比标准方法，转换预测变量提高了拟合度。

图片 9-18  
分类回归的模型摘要

多 R	R 方	调整的 R 方
.974	.948	.927

因变量：优选

预测值：包装设计 品牌名称 价格 优秀家用 品奖章 退款保证

下表显示了标准化回归系数。分类回归对变量进行标准化，因此只报告标准化系数。这些值除以其相应的标准误，即得到每个变量的 F 检验。但对每个变量的检验取决于其他预测变量在模型中的存在。换句话说，这些检验确定忽略模型中的一个预测变量（保留所有其他变量）是否会严重降低模型的预测能力。对于后续模型，这些值不能用于一次忽略多个变量。此外，交替使用最小平方可优化量化过程，这表明必须保守解释这些检验。

图片 9-19  
已转换预测变量的标准化系数

	标准系数		df	F	显著性
	Beta	标准误			
包装设计	-.748	.060	2	155.289	.000
品牌名称	.045	.060	1	.578	.459
价格	.371	.059	1	39.312	.000
优秀家用 品奖章	-.350	.059	1	35.299	.000
退款保证	-.159	.059	1	7.175	.017

因变量：优选

包装设计所得的系数最大。在包装设计中的一个标准差增加会在预测偏好等级中产生一个 0.748 标准差降低。但是，包装设计采用名义方式处理，因此定量中的增加并不一定与初始类别代码的增加对应。

标准化系数通常解释为反映每个预测变量的重要性。但是，回归系数不能完全描述预测变量的影响或预测变量之间的关系。必须将其他统计量与标准化系数结合使用，才能全面地研究预测变量效果。

## 相关性和重要性

要解释预测变量对回归的贡献，仅检查回归系数是不够的。除此之外，还应检查相关系数、偏相关系数和部分相关系数。下表为每个变量提供了这些相关性测量。

零阶相关系数表示已转换预测变量和已转换响应之间的相关性。就此数据而言，包装设计的相关系数最大。但是，如果可以解释任一预测变量或响应中的一些变异，则可以得到更好的表示形式来指出预测变量效果的好坏。

图片 9-20  
零阶、部和偏相关系数（已转换的变量）

	相关性			重要性	容差	
	零阶	偏	部分		转换后	转换前
包装设计	-.816	-.955	-.733	.644	.959	.942
品牌名称	.206	.193	.045	.010	.971	.961
价格	.440	.851	.369	.172	.989	.982
优秀家庭用品奖章	-.370	-.838	-.349	.137	.996	.991
退款保证	-.223	-.569	-.158	.037	.987	.993

因变量：优选

模型中的其他变量可能会在预测响应期间影响给定预测变量的性能。偏相关系数会从预测变量和响应中去除其他预测变量的线性影响。此度量相当于按其他预测变量对某一预测变量进行回归的残差与按其他预测变量对响应进行回归的残差之间的相关性。偏相关系数的平方表示根据去除其他变量影响后剩余的响应的残差方差所解释的方差比例。例如，包装设计的偏相关系数为  $-0.955$ 。去除了其他变量的影响后，包装设计解释了偏好等级中  $(-0.955)^2 = 0.91 = 91\%$  的变异。如果去除了其他变量的影响，则价格和优秀家庭用品奖章也解释了大部分方差。

作为一种从响应和预测变量去除变量影响的替代方法，可以只从预测变量去除影响。部分相关系数表示按其他预测变量对某一变量进行回归的响应和残差之间的相关性。将此值平方会得到根据总响应方差解释的方差比例。如果从包装设计中去除了品牌名称、优秀家庭用品奖章、退款保证和价格的影响，则包装设计解释了偏好等级中  $(-0.733)^2 = 0.54 = 54\%$  的变异。

## 重要性

除了回归系数和相关系数之外，还可以使用 Pratt 的相对重要性度量 (Pratt, 1987) 帮助解释预测变量对回归的贡献。较高的个体重要性（相对于其他重要性）表明预测变量对回归十分重要。此外，某一变量的重要性较低但其系数大小与重要预测变量的系数大小类似，表明存在抑制变量。

与回归系数相反，此度量以加法方式定义预测变量的重要性，即一组预测变量的重要性为各预测变量的重要性之和。Pratt 的测量等于预测变量的回归系数和零阶相关系数之积。这些积构成了  $R^2$ ，因此除以  $R^2$  所得的结果为 1。例如，包装设计和

品牌名称这组预测变量的重要性为 0.654。包装设计的重要性最大，包装设计、价格和优秀家庭用品奖章占此预测变量组合重要性的 95%。

### 多重共线性

如果预测变量之间的相关系数较大，将会大大降低回归模型的稳定性。相关预测变量会导致参数估计值不稳定。容差反映了自变量与其他自变量的线性相关度。此度量是变量的方差中不能归因于方程中的其他自变量的比例。如果去除某个预测变量后，其他预测变量可以解释该预测变量的大部分方差，则表明模型中不需要该预测变量。容差值接近于 1 表明不能通过其他预测变量很好地预测某一变量。相反，容差很低的变量对模型贡献的信息很少，可能导致计算问题。此外，Pratt 的重要性度量的负值较大则表明存在多重共线性。

所有容差度量都非常高。所有预测变量都不能通过其他预测变量准确地预测，因此不存在多重共线性。

### 转换图

根据初始类别值的对应定量对其进行绘图，可揭示在定量列表中隐藏的趋势。此类图通常称为变换图。对于具有类似定量的类别，应特别注意。这些类别以相同的方式影响预测响应。不过，该图的基本外观由变换类型确定。

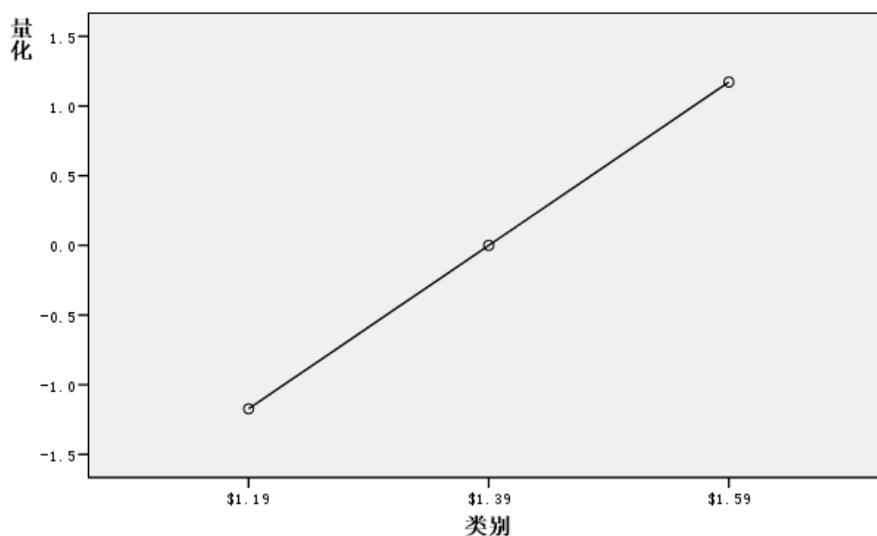
作为数值处理的变量会得到定量和初始类别之间的线性关系，这与变换图中的一条直线相对应。在定量中保留了初始类别的顺序和初始类别之差。

作为有序处理的变量的定量的顺序与初始类别的顺序相对应。但是，不保留类别之差。因此，变换图是非递减的，但不一定是一条直线。如果连续类别与类似定量相对应，则可能不需要区分类别，并且可以合并类别。这样的类别会在变换图上呈现为一个平台。但是，对应作为名义变量处理的变量施加有序结构也会产生此模式。如果变量的后续名义处理呈现出相同的模式，合并类别就有了保证。此外，如果以有序处理的变量的定量呈一条直线，则可能更适合采用数值变换。

对于以名义处理的变量，类别沿水平轴的顺序与用于表示类别的代码顺序相对应。对类别顺序或类别之间距离的解释缺乏根据。该图可以采用任何非线性或线性形式。如果呈现增加趋势，则应尝试有序处理。如果名义变换图呈现线性趋势，则可能更适合采用数值变换。

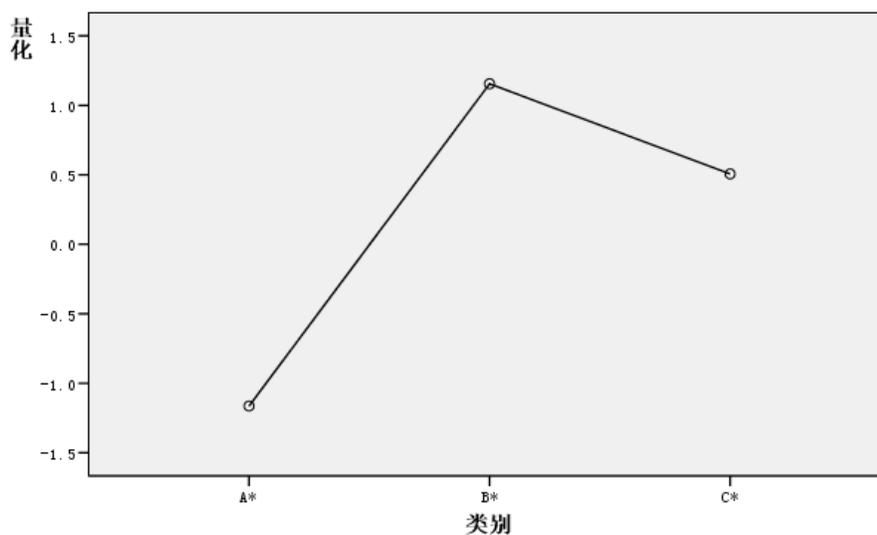
下图显示了以数值处理的价格的变换图。您会看到类别沿直线的顺序与初始类别的顺序相对应。此外，\$1.19 和 \$1.39 的定量（分别为 -1.173 和 0）之差与 \$1.39 和 \$1.59 的定量（分别为 0 和 1.173）之差相同。在定量中保留了类别 1 和 3 与类别 2 距离相同的这一事实。

图片 9-21  
价格（数值）的变换图



对包装设计的名义转换所得的变换图如下所示。您会看到与上图截然不同的非线性形状，其中第二个类别的定量最大。就回归而言，第二个类别降低了预测偏好等级，而第一个和第三个类别的效果与之正好相反。

图片 9-22  
包装设计（名义）的变换图

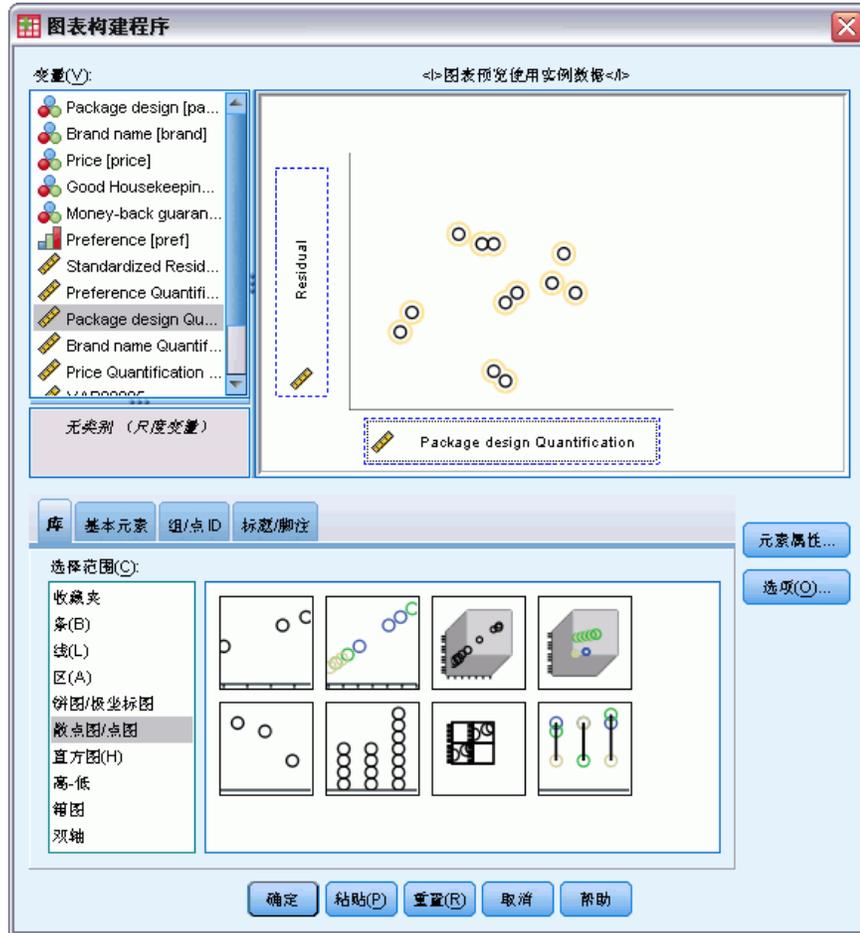


## 残差分析

通过使用保存到活动数据集中的已转换数据和残差，可以按包装设计的已转换值创建一个预测值的散点图。

要创建此散点图，请调用图表生成器，然后单击**重置**以清除以前的选项并恢复默认选项。

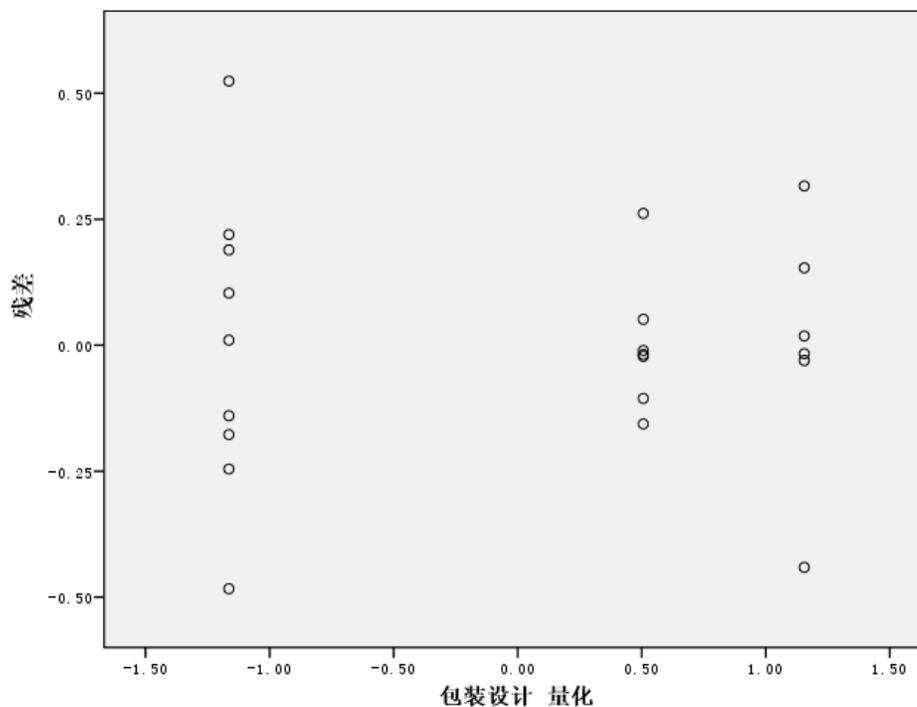
图片 9-23  
图表构建程序



- ▶ 选择散点图/点图库并选择“简单散点”。
- ▶ 选择残差作为 y 轴变量。
- ▶ 选择包装设计量化作为 x 轴变量。
- ▶ 单击确定。

该散点图将显示按包装设计的最优得分绘制的标准残差。所有残差都在两个标准差 0 范围内。在标准线性回归的散点图中呈现的 U 形被随机散点取代。通过最优类别量化，增强了预测能力。

图片 9-24  
分类回归的残差



## 示例：臭氧数据

在此示例中，将使用一个大型数据集来阐释最优尺度转换的选择和效果。这些数据包含了 Breiman 和 Friedman (Breiman 和 Friedman(F), 1985) 以及 Hastie 和 Tibshirani (Hastie 和 Tibshirani, 1990) 等人以前分析的六个气象变量的 330 个观察值。下表描述了这些初始变量。分类回归将尝试根据其余变量预测臭氧浓度。以前的研究人员发现了这些变量之间的非线性，这妨碍了标准回归方法。

表 9-2  
初始变量

变量	描述
ozon	每日臭氧水平；分为 38 个类别
ibh	逆温层底高度
dpg	压力梯度（毫米汞柱）
vis	能见度（英里）
temp	温度（华氏温度）
doy	一年中某天

此数据集可在 ozone.sav 中找到。[有关详细信息，请参阅附录 A 中的样本文件中的 IBM SPSS Categories 21。](#)

## 离散化变量

如果变量的类别数量超过了实际可解释的数量，则应使用“离散化”对话框修改这些类别以将其范围缩小至一个更可控的数量。

变量一年中某天的最小值为 3，最大值为 365。在分类回归中使用此变量相当于使用一个具有 365 个类别的变量。同样，能见度（英里）的值范围为 0 至 350。要简化分析解释，需将这些变量离散化为长度为 10 的相等区间值。

变量逆温层底高度的值范围为 111 至 5000。具有如此多类别的变量会导致关系非常复杂。但是，如果将此变量离散化为长度为 100 的相等区间值，则会大致得到 50 个类别。用 50 个类别的变量来代替 5000 个类别的变量会大大简化解释。

压力梯度（毫米汞柱）的值范围为 -69 到 107。该过程忽略了分析中用负数编码的所有类别，而且将此变量离散化为长度为 10 的相等区间值会大致得到 19 个类别。

温度（华氏温度）的值范围为 25 到 93（华氏温标）。为了用摄氏温标分析这些数据，将此变量离散化长度为 1.8 的相等区间值。

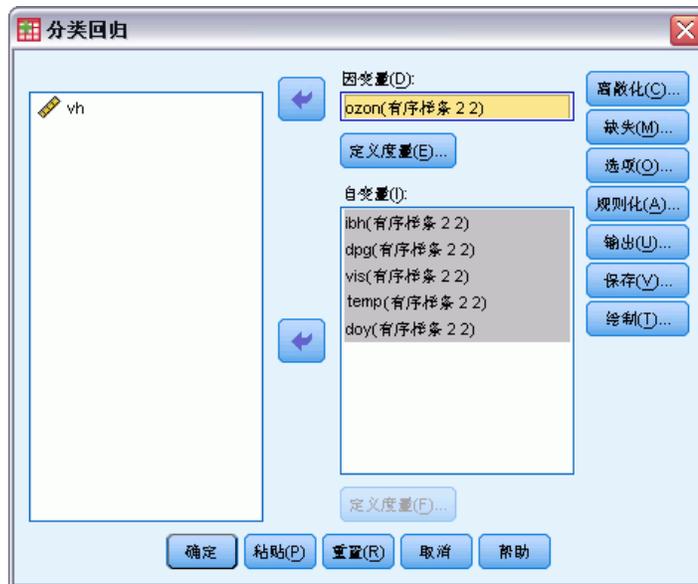
可能需要为变量进行不同的离散化。此处所使用的选项完全是个人决定。如果希望较少的类别，可以选择较大的间隔。例如，一年中某天可以按一年或一季度中的月份进行划分。

## 转换类型选择

可以按多个不同的级别来分析每个变量。但是，因为以响应预测作为目标，所以应采用数值最优尺度级别“按原样”标定响应。这样，将已转换的变量中将保留类别的顺序和类别之差。

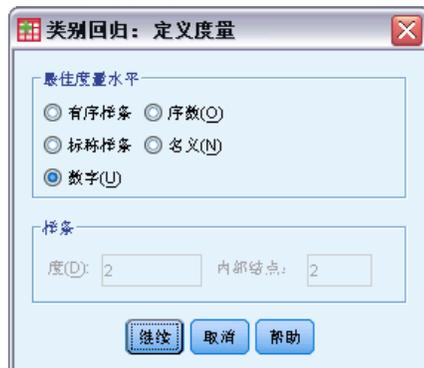
- ▶ 要运行“分类回归”分析，请从菜单中选择：  
分析 > 回归 > 最佳尺度 (CATREG)...

图片 9-25  
“分类回归”对话框



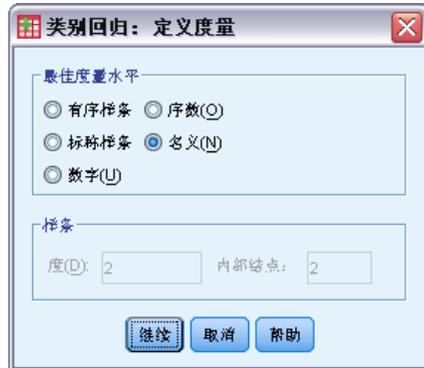
- ▶ 选择每日臭氧水平作为因变量。
- ▶ 选择从逆温层底高度到一年中某天的变量作为自变量。
- ▶ 选择每日臭氧水平，然后单击定义度量。

图片 9-26  
“定义尺度”对话框



- ▶ 选择数值作为最佳度量水平。
- ▶ 单击继续。
- ▶ 选择从逆温层底高度到一年中某天的变量，然后在“分类回归”对话框中单击定义度量。

图片 9-27  
“定义尺度”对话框



- ▶ 选择名义作为最佳度量水平。
- ▶ 单击继续。
- ▶ 在“分类回归”对话框中单击离散化。

图片 9-28  
“离散化”对话框



- ▶ 选择 ibh。
- ▶ 选择相等区间，然后键入 100 作为间隔长度。
- ▶ 单击更改。
- ▶ 选择 dpg、vis 和 doy。

- ▶ 键入 10 作为区间长度。
- ▶ 单击更改。
- ▶ 选择 temp。
- ▶ 键入 1.8 作为区间长度。
- ▶ 单击更改。
- ▶ 单击继续。
- ▶ 在“分类回归”对话框中单击图。

图片 9-29  
“图”对话框



- ▶ 选择从逆温层底高度到一年中某天的变换图。
- ▶ 单击继续。
- ▶ 在“分类回归”对话框中单击确定。

图片 9-30  
模型汇总

	多 R	R 方	调整 R 方	明显预测误差
标准数据	.938	.880	.785	.120

因变量: ozon  
预测变量: ibh dpg vis 温度 doy

将所有预测变量作为名义变量处理会得到  $R^2$  值 .880。贡献如此大一部分方差并不奇怪，因为作为名义值处理对定量没有施加限制。但解释结果可能会非常困难。

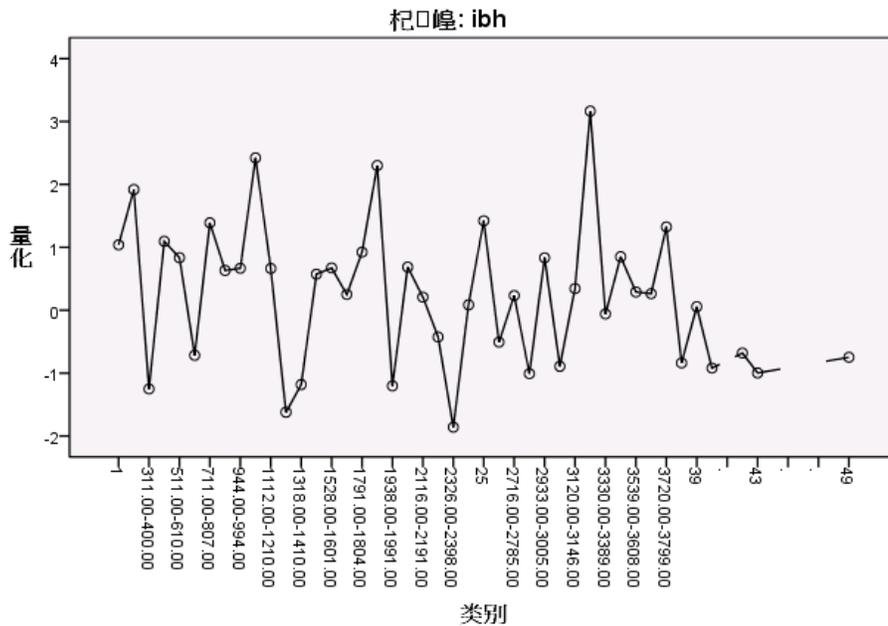
图片 9-31  
回归系数（所有预测变量均以名义处理）

	标准系数		df	F	Sig.
	Beta	标准误差的 Bootstrap (1000) 估计			
ibh	.297	.053	42	31.047	.000
dpg	.326	.055	16	34.793	.000
vis	.229	.050	17	21.465	.000
温度	.577	.091	35	40.562	.000
doy	.420	.070	36	36.171	.000

因变量: ozon

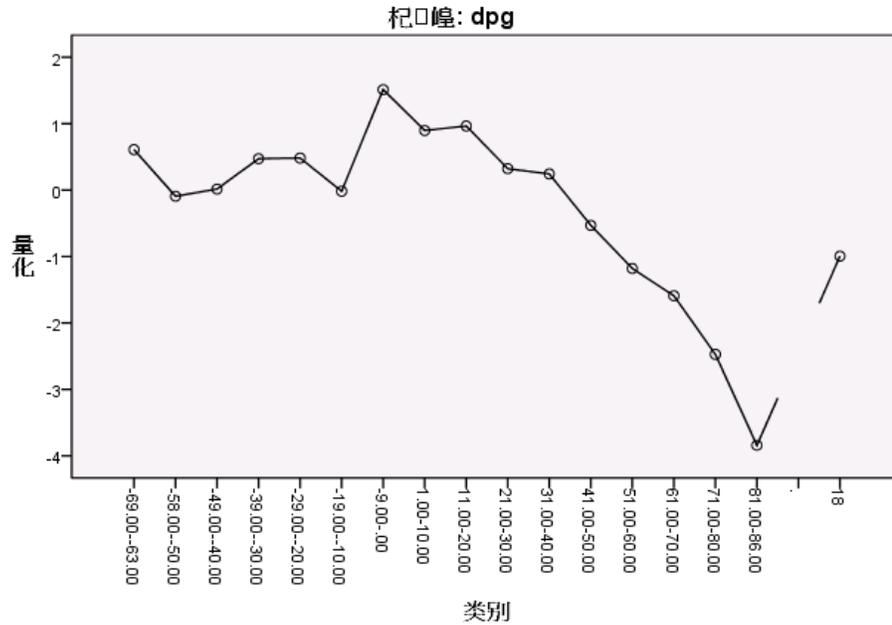
下表显示了这些预测变量的标准化回归系数。在解释这些值时往往会犯这样一个错误：只关注系数而忽略了定量。因此，举例来说，不能简单认为逆温层底高度的正值就意味着预测变量增加时，预测的臭氧水平也会相应增加。所有解释必须相对于转换的变量，因此当逆温层底高度的定量增加时，预测的臭氧水平也增加。要检查初始变量的影响，必须将这些变量的类别与定量相关。

图片 9-32  
逆温层底高度（名义）的变换图



逆温层底高度的变换图未显现出明显的模式。从该图的锯齿特性表明，从低类别移动到高类别会导致定量上下波动。因此，若要描述此变量的影响，需要单独着重分析各个类别。对此变量的定量施加有序或线性限制可能会显著降低拟合度。

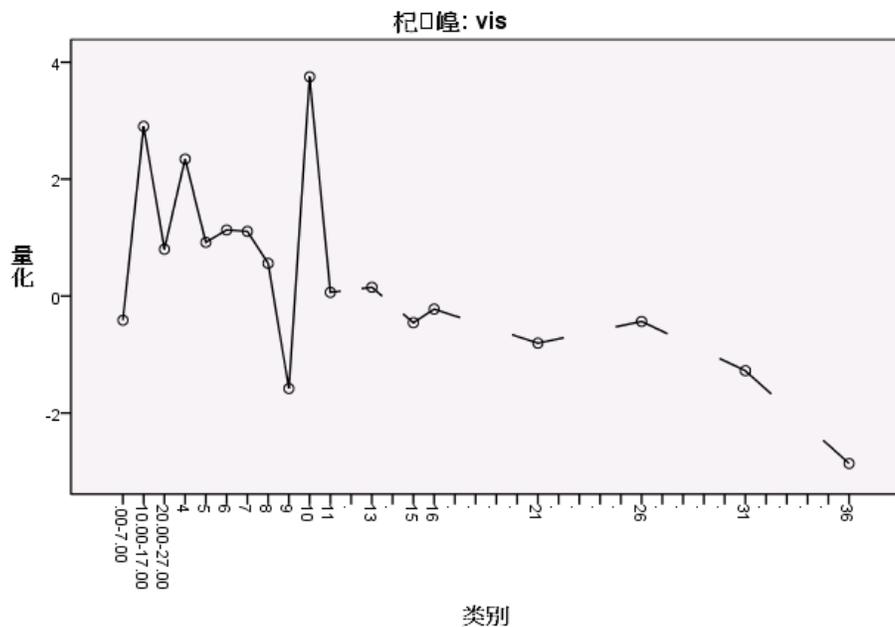
图片 9-33  
压力梯度（名义）的变换图



下图显示了压力梯度的变换图。初始离散化类别（1 到 6）所具有的定量较小，因此对预测响应的贡献最小。接下来的三个类别具有略大的正值，导致预测臭氧水平适度增加。

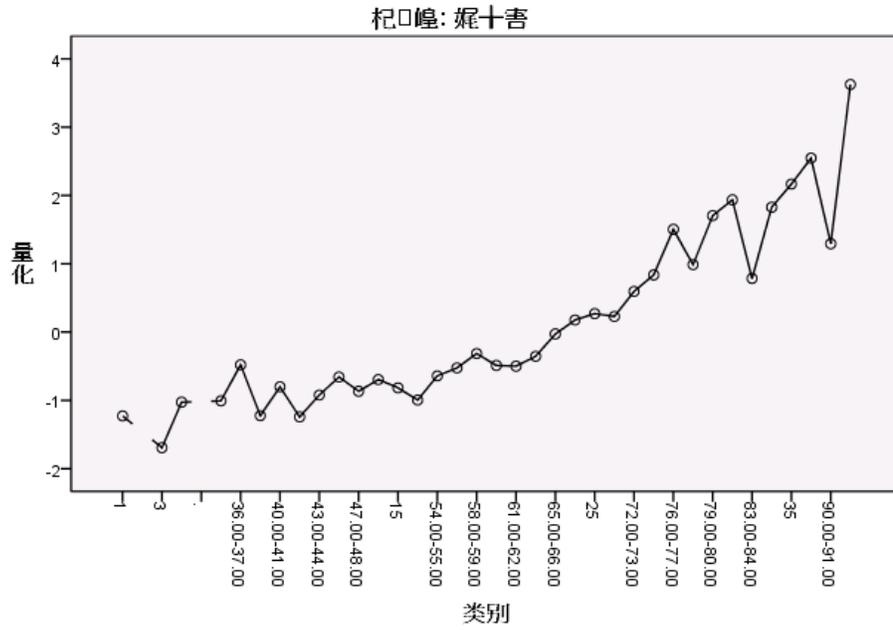
定量在类别 16 处降到了最低值，这表明此处的压力梯度对预测臭氧水平的降低作用最大。虽然在此类别后线图有所上升，但使用压力梯度的有序尺度级别在简化效果解释的同时，不会大幅度降低拟合度。不过，压力梯度重要性度量 0.04 和回归系数表明此变量在回归中不是十分有用。

图片 9-34  
能见度（名义）的变换图



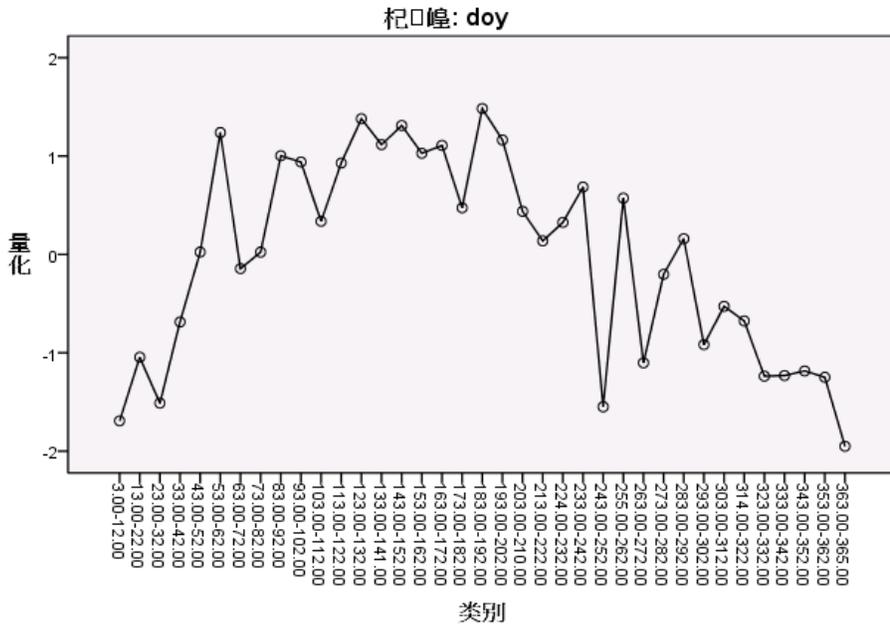
能见度的变换图与逆温层底高度的变换图类似，未呈现出明显的模式。对此变量的定量施加有序或线性限制可能会显著降低拟合度。

图片 9-35  
温度（名义）的变换图



温度的变换图呈现出另一种模式。随着类别的增加，定量也趋向于增加。因此，随着温度的增加，预测臭氧水平也趋向于增加。此模式表明应在有序级别标定温度。

图片 9-36  
一年中某天的变换图（名义）



此图显示一年中某天的变换图。定量在图表中点处增加到最高值，接着定量又在该点开始降低，呈现一个倒置的 U 形。考虑一年中某天的回归系数符号，初始类别的量化结果对预测臭氧水平的影响逐渐减少。对于中间类别，量化结果对预测臭氧水平的影响逐渐增加，在图表中点处到达最高值。

在该点后，定量又趋向于降低预测臭氧水平。虽然线型呈锯齿形，但图的总体形状仍可看出一定的趋势。因此，该变换图表明应在有序级别标定温度，而在标定级别标定所有其他预测变量。

要重新计算在有序级别标定温度的情况下的回归，请调用“分类回归”对话框。

图片 9-37  
“定义尺度”对话框



- ▶ 选择温度，然后单击定义度量。
- ▶ 选择有序作为最佳度量水平。

- ▶ 单击继续。
- ▶ 在“分类回归”对话框中单击保存。

图片 9-38  
“保存”对话框



- ▶ 选择“已转换的变量”组中的将已转换的变量保存到活动数据集。
- ▶ 单击继续。
- ▶ 在“分类回归”对话框中单击确定。

图片 9-39  
使用温度（有序）的回归的模型摘要

模型汇总				
	多 R	R 方	调整 R 方	明显预测误差
标准数据	.934	.872	.787	.128

因变量: ozon  
预测变量: ibh dpg vis 温度 doy

此模型得到  $R^2$  值 0.872，因此在将温度定量限制为经排序的变量时，做出贡献的方差量略有减少。

图片 9-40  
使用温度（有序）的回归系数

	标准系数		df	F	Sig.
	Beta	标准误差的 Bootstrap (1000) 估计			
ibh	.298	.042	42	51.376	.000
dpg	.301	.047	16	41.882	.000
vis	.224	.044	17	25.659	.000
温度	.609	.084	21	52.113	.000
doy	.373	.053	36	50.330	.000

因变量: ozon

下表显示了将温度标定为有序变量的模型的系数。通过将上一模型的系数与将温度标定为名义变量的模型的系数进行比较，可以发现没有大的变化。

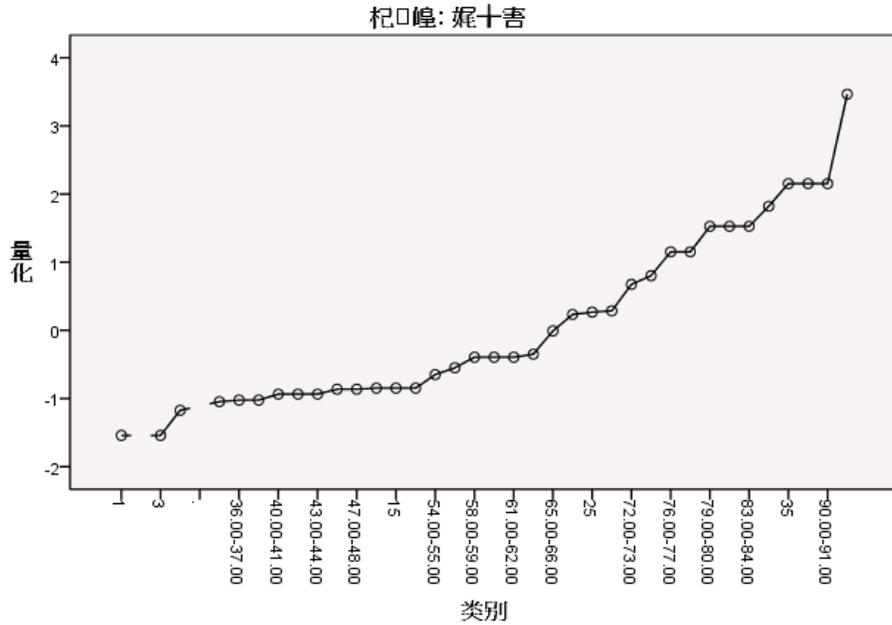
图片 9-41  
相关性、重要性和容差

	相关性			重要性	容差	
	零阶	偏	部分		转换后	转换前
ibh	.438	.627	.288	.150	.930	.596
dpg	.128	.606	.272	.044	.815	.858
vis	.365	.518	.216	.094	.933	.752
温度	.804	.843	.559	.562	.842	.580
doy	.352	.677	.329	.151	.777	.802

因变量: ozon

此外，重要性度量表明温度对于回归而言仍比其他变量重要得多。然而，现在从温度的有序尺度级别和正回归系数的结果来看，可以认为温度增加时，预测臭氧水平也会随之增加。

图片 9-42  
温度（有序）的变换图



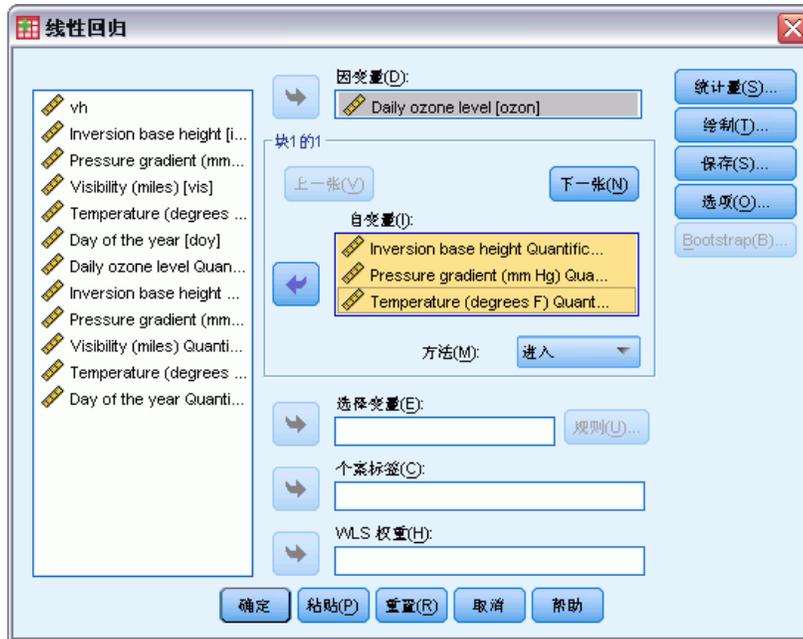
该变换图显示了对温度的定量的有序限制。原来在名义变换图中呈现的锯齿线在此图中变为一条上升线。此外，不再呈现出任何长的平台，这表明不需要拼并类别。

## 定量最优化

分类回归中的已转换变量可用于标准线性回归，这会产生相同的结果。但是，定量仅对于产生这些定量的模型是最优的。在线性回归中使用预测变量的子集并不对应于同一子集上的最优尺度回归。

例如，已计算的分类回归具有值为 0.875 的  $R^2$ 。由于保存了已转换的变量，因此为了拟合只使用温度、压力梯度和逆温层底高度作为预测变量的线性回归，请从菜单中选择：  
分析 > 回归 > 线性...

图片 9-43  
“线性回归”对话框



- ▶ 选择每日臭氧水平定量作为因变量。
- ▶ 选择逆温层底高度定量、压力梯度（毫米汞柱）定量和温度（华氏温度）作为自变量。
- ▶ 单击确定。

图片 9-44  
使用最优尺度预测变量子集的模型摘要

**模型汇总**

模型	R	R方	调整 R方	标准估计的误差
1	.856 <sup>a</sup>	.732	.729	4.16711

a. 预测变量: (常量), 温度 量化, Dpg 量化, lbh 量化。

通过对响应使用定量，在标准线性回归中温度、压力梯度和逆温底层高度得到的拟合度为 .732。要将此拟合度与只使用这三个预测变量的分类回归的拟合度进行比较，请调用“分类回归”对话框。

图片 9-45  
“分类回归”对话框



- ▶ 取消选择能见度（英里）和一年中某天作为自变量。
- ▶ 单击确定。

图片 9-46  
使用三个预测变量的分类回归的模型摘要

	多 R	R 方	调整 R 方	明显预测误差
标准数据	.892	.796	.735	.204

因变量: ozon  
预测变量: ibh dpg 温度

分类回归分析所得的拟合度为 .796，这高于拟合度 .732。这表明，只有当模型包含所有五个变量时，定量在初始回归中获取的尺度属性是最优的。

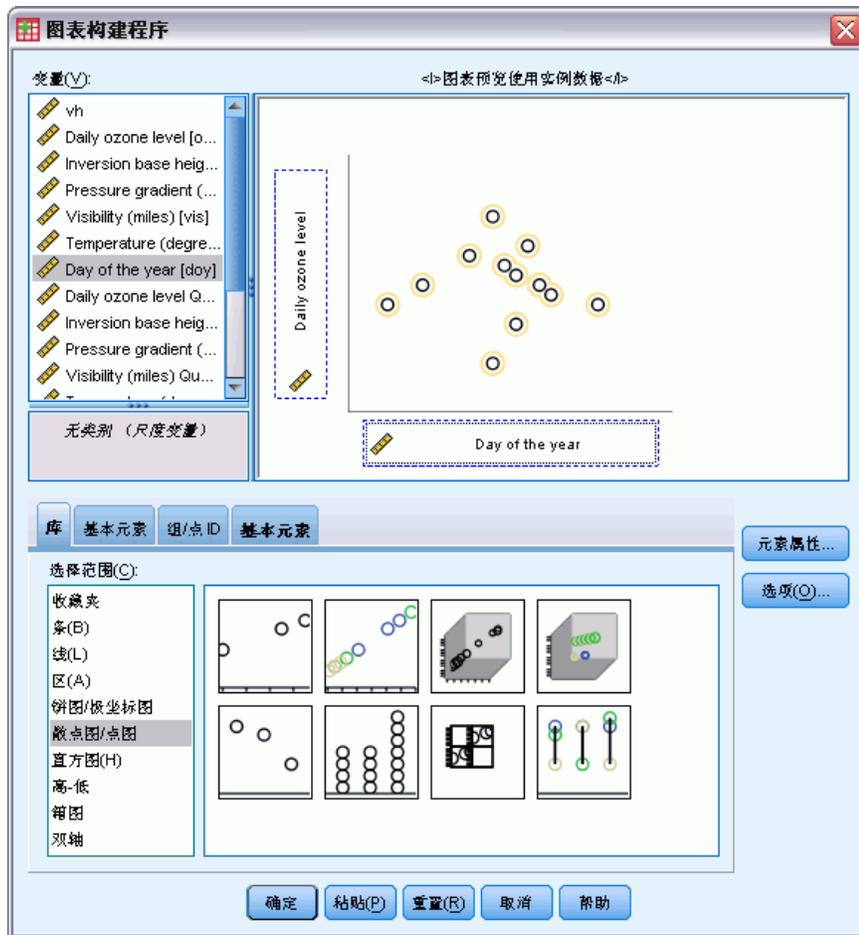
## 转换效果

对变量进行转换会在初始响应与已转换变量的线性初始预测变量集之间产生非线性关系。但如果存在多个预测变量，则模型中的其他变量会对成对关系造成妨碍。

要着重分析每日臭氧水平和一年中某天之间的关系，请首先查看散点图。从菜单中选择：

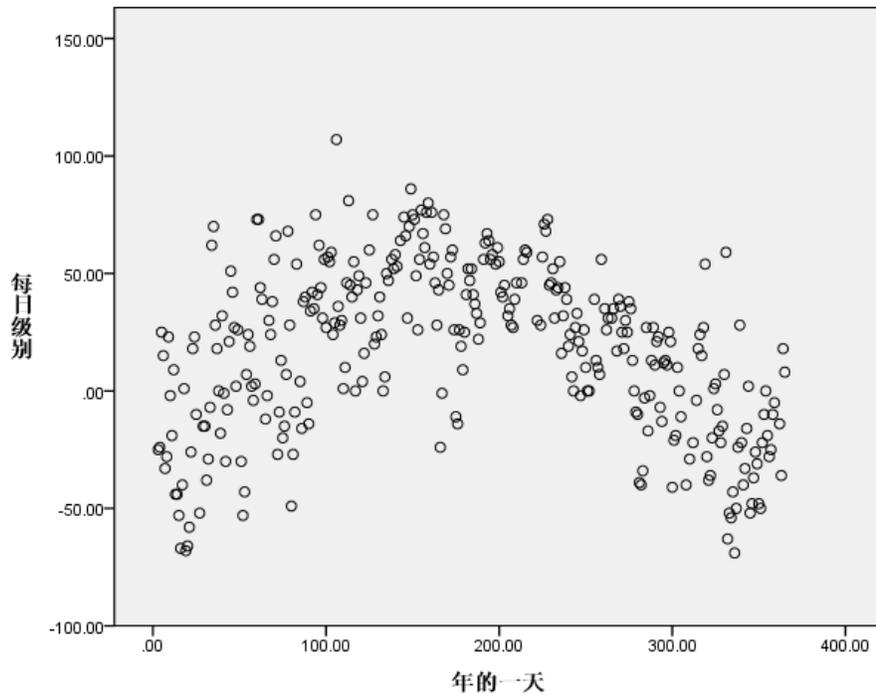
图形 > 图表生成器...

图片 9-47  
“图表生成器”对话框



- ▶ 选择散点图/点图库并选择“简单散点”。
- ▶ 选择每天臭氧水平作为 y 轴变量，选择一年中某天作为 x 轴变量。
- ▶ 单击确定。

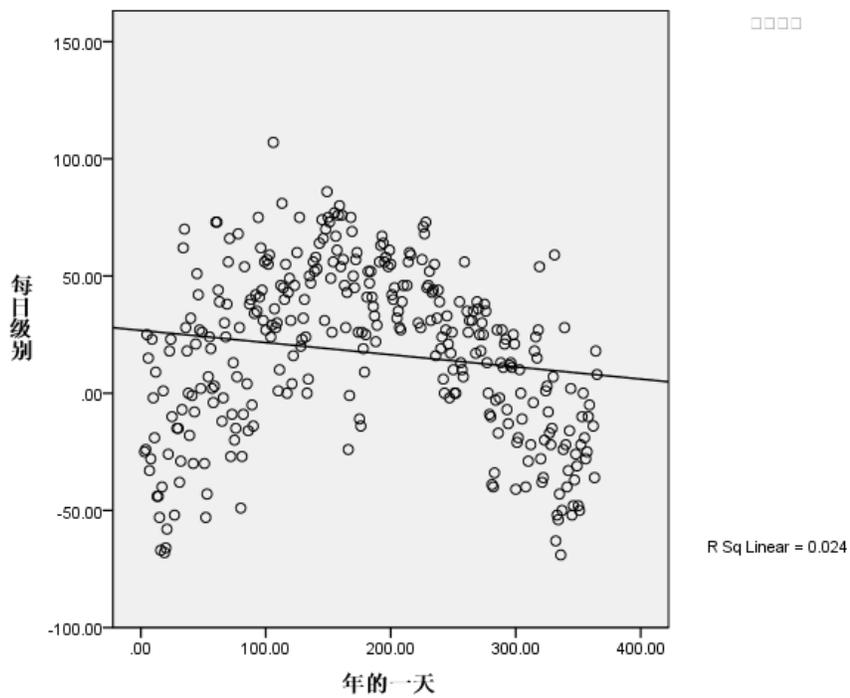
图片 9-48  
每日臭氧水平和一年中某天的散点图



下图演示了每日臭氧水平与一年中某天之间的关系。当一年中某天大约增加到 200 时，每日臭氧水平将会增加。但当一年中某天值大于 200 时，每日臭氧水平将会减少。这一倒置的 U 形表明了这两个变量之间存在二次关系。而线性回归无法捕捉到这种关系。

- ▶ 要在散点图上显示覆盖点的最佳拟合线，请双击激活该图。
- ▶ 在图表编辑器中选择一个点。
- ▶ 单击添加总计拟合线工具，然后关闭图表编辑器。

图片 9-49  
显示最佳拟合线的散点图



根据一年中某天对每日臭氧水平进行线性回归会得到  $R^2$  值 0.004。此拟合度表明一年中某天没有每日臭氧水平的预测值。考虑到此图所呈现的模式，这并没有什么奇怪的。但通过使用最优尺度，可以线性化二次关系并使用已转换的一年中某天变量来预测响应。

图片 9-50  
“分类回归”对话框



要根据一年中某天获取对每日臭氧水平的分类回归，请调用“分类回归”对话框。

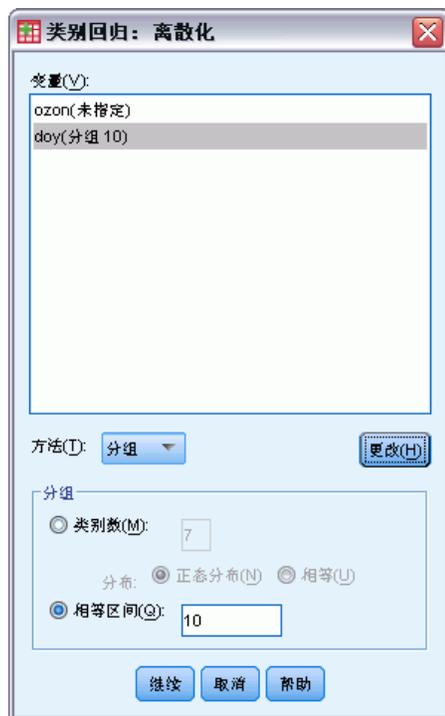
- ▶ 取消选择从逆温层底高度到温度（华氏温度）的变量作为自变量。
- ▶ 选择一年中某天作为自变量。
- ▶ 单击定义尺度。

图片 9-51  
“定义尺度”对话框



- ▶ 选择名义作为最佳度量水平。
- ▶ 单击继续。
- ▶ 在“分类回归”对话框中单击离散化。

图片 9-52  
“离散化”对话框



- ▶ 选择 doy。
- ▶ 选择相等区间。
- ▶ 键入 10 作为区间长度。
- ▶ 单击更改。
- ▶ 单击继续。
- ▶ 在“分类回归”对话框中单击图。

图片 9-53  
“图”对话框



- ▶ 选择 doy 的转换图。
- ▶ 单击继续。
- ▶ 在“分类回归”对话框中单击确定。

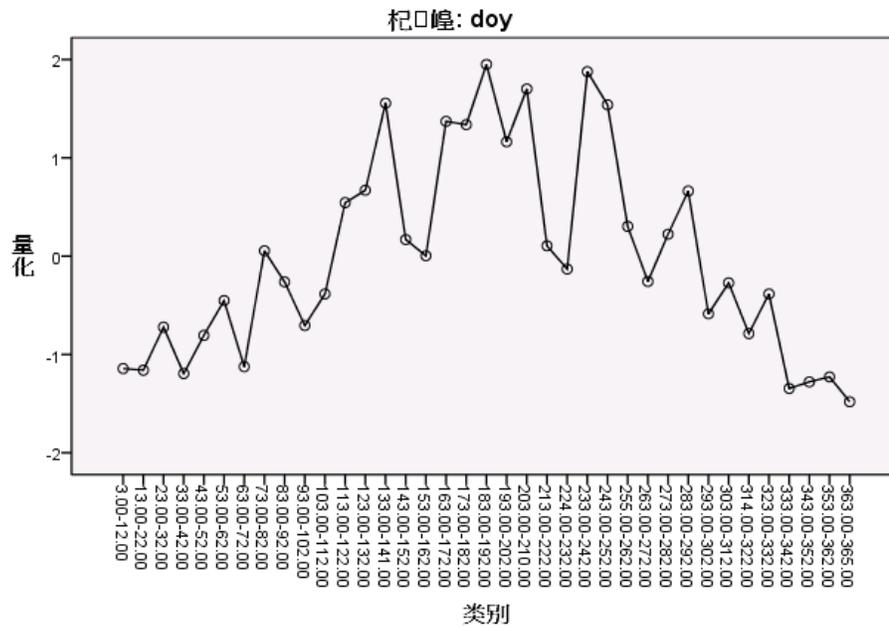
图片 9-54  
按“一年中某天”对“每日臭氧水平”进行分类回归的模型摘要

	多 R	R 方	调整 R 方	明显预测误差
标准数据	.741	.549	.494	.451

因变量: ozon  
预测值: doy

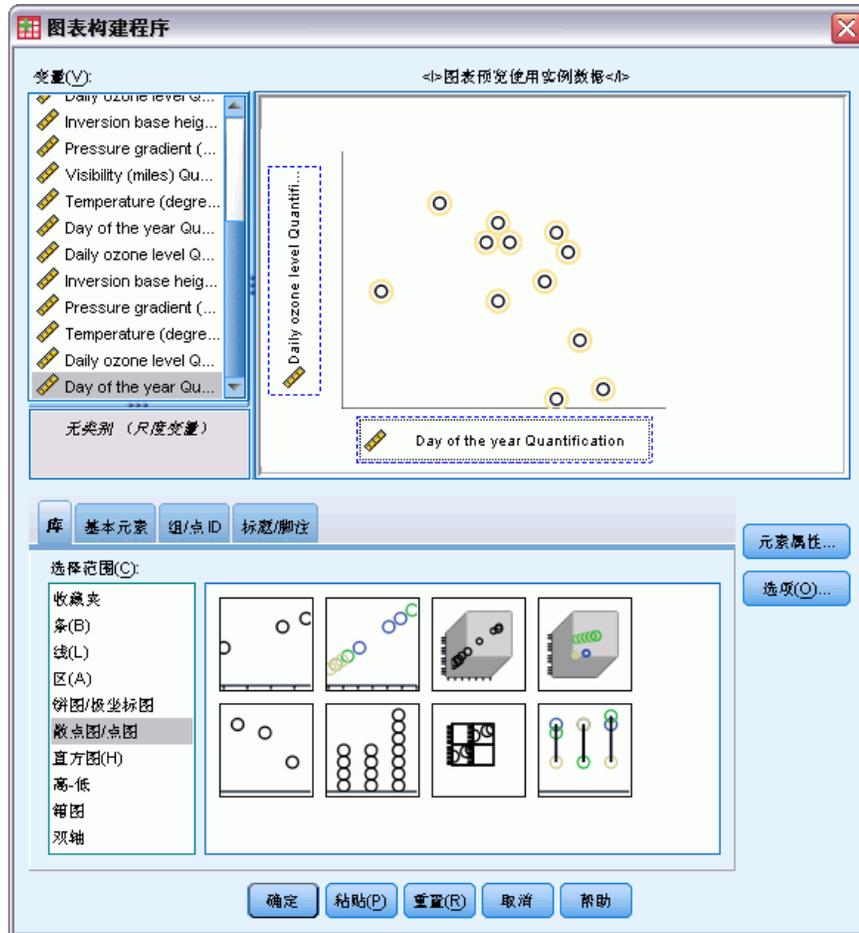
最优尺度回归按数值处理每日臭氧水平，按名义处理一年中某天。这会得到  $R^2$  值 0.549。虽然分类回归只贡献了每日臭氧水平中 55% 的变异，但这仍比初始回归有了实质的改进。通过转换一年中某天可以预测每日臭氧水平。

图片 9-55  
一年中某天的变换图（名义）



下图显示了一年中的变换图。一年中某天的极值都具有负定量，而中间值都具有正定量。应用了此变换后，一年中某天的高低值对预测变量每日臭氧水平产生了类似的影响。

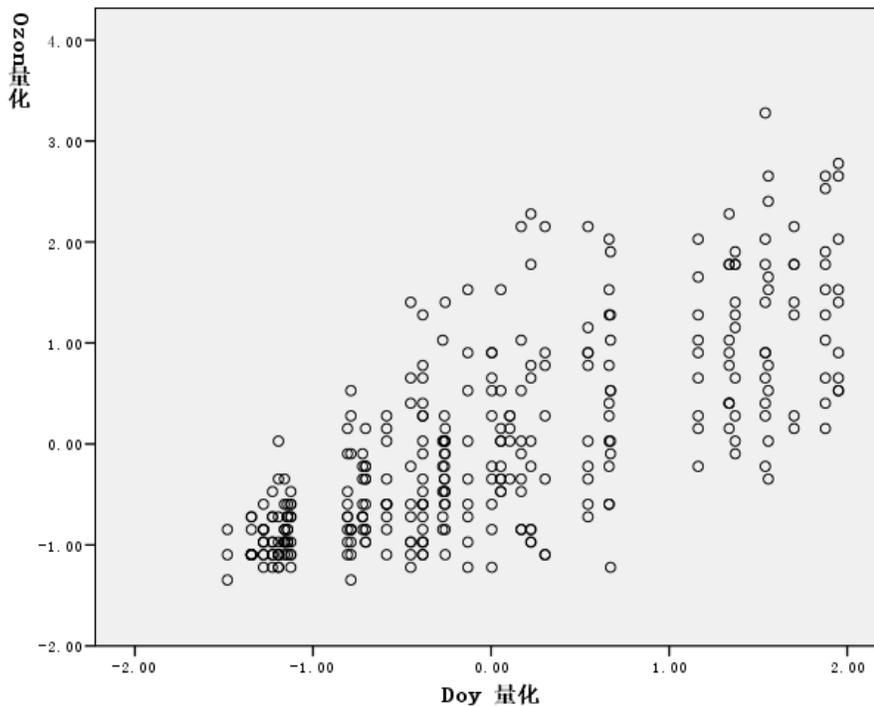
图片 9-56  
图表构建程序



要查看已转换变量的散点图，请调用图表生成器，然后单击重置以清除以前的选项。

- ▶ 选择散点图/点图库并选择“简单散点”。
- ▶ 选择每日臭氧水平定量 [TRA1\_3] 作为 y 轴变量，选择一年中某天定量 [TRA2\_3] 作为 x 轴变量。
- ▶ 单击确定。

图片 9-57  
已转换变量的散点图



下图描绘了已转换变量之间的关系。上升趋势取代了原来的倒置 U 形。回归线具有正斜率，这表明随着已转换变量一年中某天的增加，预测变量每日臭氧水平也将增加。使用最优尺度可线性化关系并可以提供原来被忽视的解释。

## 推荐参考

有关分类回归的更多信息，请参见以下内容：

Buja, A. 1990. Remarks on functional canonical variates, alternating least squares methods and ACE. *Annals of Statistics*, 18, .

Hastie, T., R. Tibshirani, 和 A. Buja. 1994. Flexible discriminant analysis. *Journal of the American Statistical Association*, 89, .

Hayashi, C. 1952. On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics*, 2, .

Kruskal, J. B. 1965. Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society Series B*, 27, .

Meulman, J. J. 2003. Prediction and classification in nonlinear data analysis: Something old, something new, something borrowed, something blue. *Psychometrika*, 4, .

- 
- Ramsay, J. O. 1989. Monotone regression splines in action. *Statistical Science*, 4, .
- Van der Kooij, A. J., 和 J. J. Meulman. 1997. MURALS: Multiple regression and optimal scaling using alternating least squares. 输入: *Softstat '97*, F. Faulbaum, 和 W. Bandilla, eds. Stuttgart: Gustav Fisher.
- Winsberg, S., 和 J. O. Ramsay. 1980. Monotonic transformations to additivity using splines. *Biometrika*, 67, .
- Winsberg, S., 和 J. O. Ramsay. 1983. Monotone spline transformations for dimension reduction. *Psychometrika*, 48, .
- Young, F. W., J. De Leeuw, 和 Y. Takane. 1976. Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 41, .

# 分类主成份分析

分类主成份分析可以被视为一种降低维数的方法。该方法对一组变量进行分析，从而揭示变差的主要维度。这样，就可以用较小的新数据集替换初始数据集，同时将信息损失程度降到最低。该方法揭示了变量之间、个案之间以及变量与个案之间的关系。

分类主成份分析用于量化观测数据的标准是对象得分（成份得分）应与每个已量化的变量具有很大的相关性。解的好坏将取决于满足此标准的程度。

下面提供了分类主成份分析的两个示例。第一个示例用一个比较小的数据集来阐释该过程所涉及的基本概念和解释。第二个示例检验实际的应用情况。

## 示例: 检验社会体系的相关性

此示例检验经 Guttman (Guttman, 1968) 改编的 Bell 表 (Guttman 等., 1968)。Lingoes 也讨论过这些数据 (Lingoes, 1968)。

Bell 创建了一个表，用来阐释可能的社会群体。Guttman 引用了该表的一部分，其中包括五个变量，用于描述以下七个理论社会群体的社会交往、对群体的归属感、成员的物理亲近度以及关系正式性：观众（比如在足球比赛现场的人们）、听众（比如在剧院或听课堂讲座的人们）、公众（比如报纸或电视观众）、组织群体（与观众类似但具有紧密的关系）、初级群体（关系密切）、次级群体（自发组织）及现代社区（因在物理上亲近而导致关系松散并需要专业化服务）。

下表显示了因将 Guttman-Bell 数据分为七个社会群体而在该数据集中所产生的变量，这些变量的变量标签和值标签（类别）分别与每个变量级别相关联。此数据集可在 guttman.sav 中找到。[有关详细信息，请参阅附录 A 中的样本文件中的 IBM SPSS Categories 21。](#)除了选择要包含在分类主成份分析计算中的变量之外，还可以选择用于在图中标记对象的变量。在此示例中，数据中的前五个变量包含在分析中，而聚类专门用作标记变量。在指定分类主成份分析时，必须为每个分析变量指定最佳度量水平。在此示例中，为所有分析变量指定了顺序级别。

表 10-1  
Guttman-Bell 数据集中的变量

变量名	变量标签	值标签
intnsity	交往强度	轻微、低、中、高
frquency	交往频率	轻微、一次、不经常、经常
blonging	归属感	无、轻微、不定、高
proximity	物理亲近度	远、近
formlity	关系正式性	没有关系、正式、不正式
聚类		观众、听众、公众、组织群体、初级群体、次级群体、现代社区

## 运行分析

- ▶ 要生成此数据集的分类主成份输出，请从菜单中选择：  
分析 > 降维 > 最优尺度...

图片 10-1  
“最优尺度”对话框



- ▶ 在“最佳度量水平”组中，选择有些变量不是无序多分类。
- ▶ 单击定义。

图片 10-2  
分类主要成分对话框



- ▶ 选择从交往强度到关系正式性作为分析变量。
- ▶ 单击定义尺度和权重。

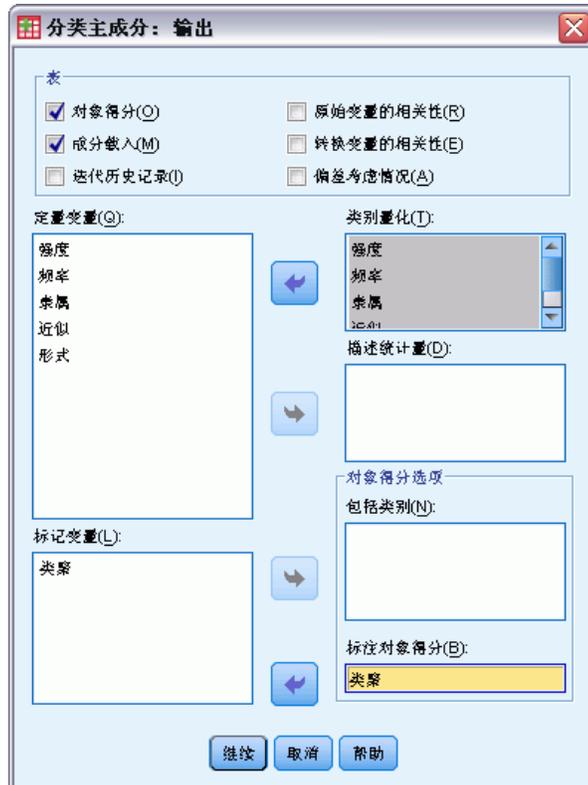
图片 10-3  
“定义尺度和权重”对话框



- ▶ 在“最佳度量水平”组中选择有序。
- ▶ 单击继续。
- ▶ 在“分类主成份”对话框中选择聚类作为标注变量。

- ▶ 单击输出。

图片 10-4  
“输出”对话框



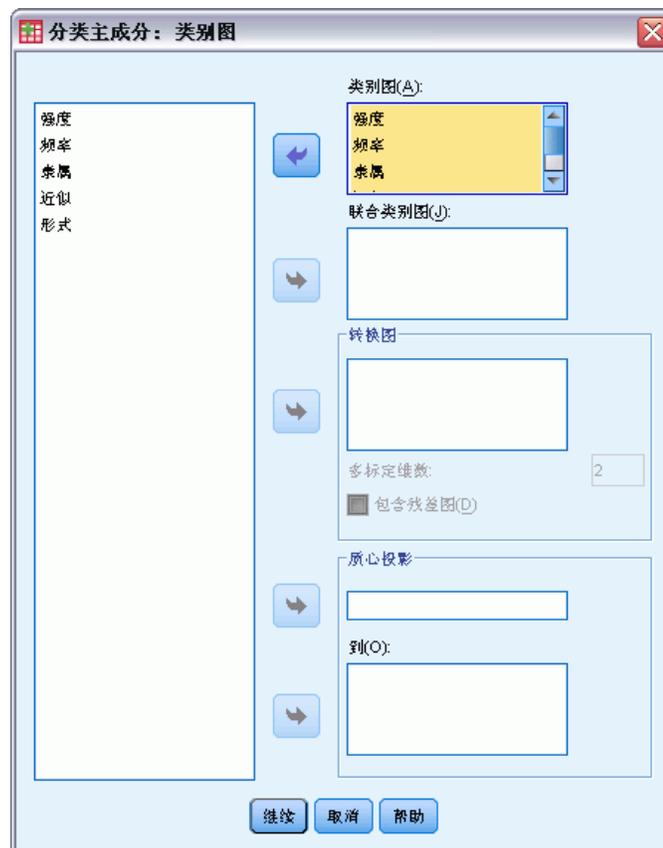
- ▶ 在“表格”组中选择对象得分并取消选择转换变量的相关性。
- ▶ 选择为从强度（交往强度）到正式性（关系正式性）的变量生成类别量化。
- ▶ 选择按聚类标注对象得分。
- ▶ 单击继续。
- ▶ 在“分类主成份”对话框的“图”组中单击对象。

图片 10-5  
“对象和变量图”对话框



- ▶ 在“图”组中选择对象和变量（双标图）。
- ▶ 在“标签对象”组中选择按变量标记对象，然后选择聚类作为要按其标注对象的变量。
- ▶ 单击继续。
- ▶ 在“分类主成份”对话框的“图”组中单击类别。

图片 10-6  
“类别图”对话框



- ▶ 选择为从强度（交往强度）到正式性（关系正式性）的变量生成联合类别图。
- ▶ 单击继续。
- ▶ 在“分类主成份”对话框中单击确定。

## 维数

以下各图显示了分类主成份分析的一些初始输出。在显示算法的迭代历史记录后，将显示模型摘要，其中包括每个维度的特征值。这些特征值相当于经典主成份分析的那些特征值。它们是计算每个维度的方差贡献率的测量。

图片 10-7  
迭代历史记录

迭代次数	方差考虑情况		损失		
	总计	增量	总计	质心坐标	矢量坐标的质心限制值
0 <sup>a</sup>	4.515315	.000000	5.484685	4.075583	1.409101
31 <sup>b</sup>	4.726009	.000008	5.273991	4.273795	1.000196

- a. 迭代 0 显示了所有变量的解的统计量，除了被当作数字的具有最佳刻度水平（多定类）的变量。
- b. 因为获得收敛的检验值，所以迭代过程停止。

图片 10-8  
模型汇总

维	Cronbach's Alpha	方差考虑情况	
		总计(特征值)	方差的 %
1	.881	3.389	67.774
2	.315	1.337	26.746
总计	.986 <sup>a</sup>	4.726	94.520

- a. 总 Cronbach's Alpha 基于总特征值。

特征值可用于指示所需的维数。在此示例中，使用默认维数 2。此维数正确吗？一般来说，如果所有变量均为单名义、有序或数值时，维度的特征值应大于 1。由于二维解贡献了 94.52% 的方差，因此第三个维度可能不会增加更多的信息。

对于多名义变量，没有用于确定适当维数的简单既定规则。如果变量数被类别总数减去变量数之差所取代，则上述规则仍适用。但可能只有此规则允许使用超过需要的维数。在选择维数时，最有用的准则是使维数足够小以便可以作出有意义的解释。该模型摘要表还显示了由该过程最大化的 Cronbach's alpha（可靠性度量）。

## 定量

对于每个变量，将提供每个维度的定量、矢量坐标和质心坐标。定量是赋予每个类别的值。质心坐标是同一类别中对象的平均对象得分。矢量坐标是类别需要位于一条线上时的坐标，表示对象空间中的变量。具有有序尺度级别和数值尺度级别的变量需要此坐标。

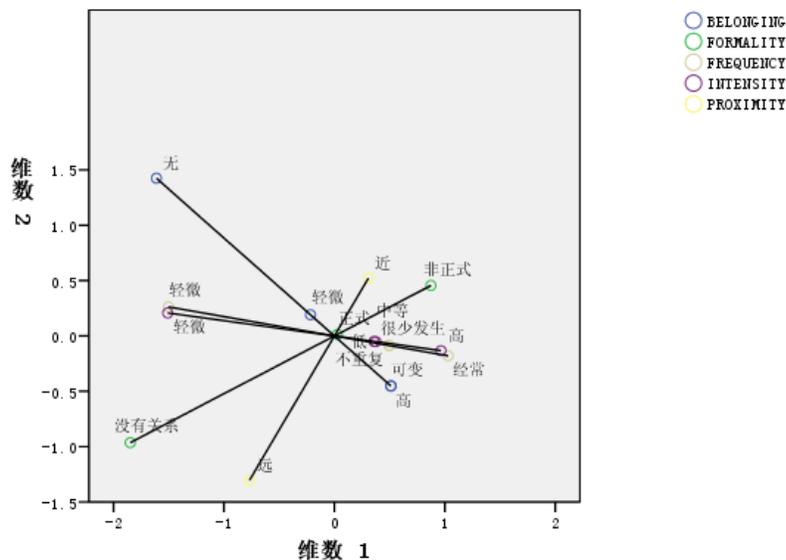
图片 10-9  
交往强度的定量

类别	频率	量化	质心坐标		矢量坐标	
			维		维	
			1	2	1	2
轻微	2	-1.530	-1.496	.308	-1.510	.208
低	2	.362	.392	.202	.358	-.049
中等	1	.379	.188	-1.408	.374	-.051
高	2	.978	1.010	.194	.965	-.133

主要变量标准化。

查看类别点联合图中的定量时，可以看到在级别为真正的有序值时，有些变量的一些类别并没有像预期的那样被分类主成份分析明确界定。例如，对于交往强度和交往频率这两个变量，其两个中间类别的定量相同或几乎相同。这种结果可能表明应尝试其他分类主成份分析，可以用一些合并的类别，也可以用其他分析级别，如（多个）标定。

图片 10-10  
联合图类别点



类别点联合图与成份载荷图类似，但它还显示了与最低定量（例如，交往强度的轻和归属感的无）相对应的端点位置。交往强度和交往频率这两个度量关系的变量非常接近地一起出现，并在维度 1 中贡献了大部分方差。此外，关系正式性和物理亲近度也非常接近。

通过着重于类别点，可以更清楚地查看这些关系。交往强度和交往频率不仅接近，而且其刻度的方向也十分类似，也就是说交往强度“轻微”与交往频率“轻微”接近，而交往强度“高”与交往频率“经常”接近。此外，还可以看出物理亲近度“近”好像与关系“不正式”类型总是一同出现，而物理亲近度“远”与关系“无”相关。

## 对象得分

您还可以请求对象得分表和对象得分图。对象得分图可用于检测离群值、检测典型对象组或揭示一些特殊模式。

对象得分表列出了按 Guttman-Bell 数据的社会群体标记的对象得分。通过查看对象点的值，可以在图中确定特定对象。

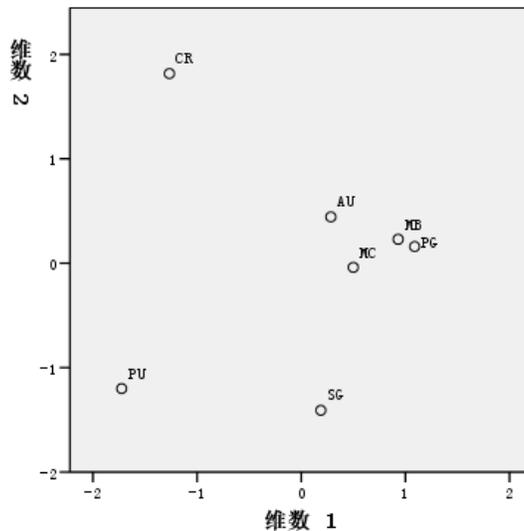
图片 10-11  
对象得分

类聚	维	
	1	2
CR	-1.266	1.816
AU	.284	.444
PU	-1.726	-1.201
MB	.931	.229
PG	1.089	.159
SG	.188	-1.408
MC	.500	-.039

主要变量标准化。

第一个维度显示将观众和公共（具有相对较大的负得分）与组织群体和初级群体（具有相对较大的正得分）区分开。第二个维度分为以下三块：公共和次级群体（具有较大的负值）、观众（具有较大正值）以及它们之间的其他社会群体。通过查看对象得分图，很容易看出这一点。

图片 10-12  
对象得分图

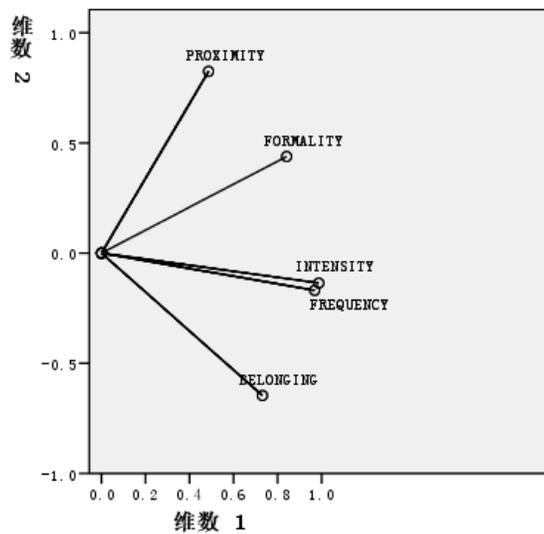


在该图中，可以看到公共和次级群体位于底部，观众位于顶部，而其他社会群体则位于中间。检查各对象之间的模式要取决于分析单元的其他可用信息。在此个案中，事先知道对象的分类。在其他个案中，可以使用补充变量来标记对象。此外，还可以看出分类主成份分析未将组织群体与初级群体区分开。在所使用的变量中，虽然大多数人通常不会将其家庭视为组织群体，但这两个群体在五个变量中的四个中的得分相同！显而易见，您可能希望找出所使用的变量和类别的可能缺陷。例如，“高”交往强度与“不正式”关系对于这两组来说可能意味着完全不同的内容。或者，您可以考虑一个更高维数的解。

## 成份载荷

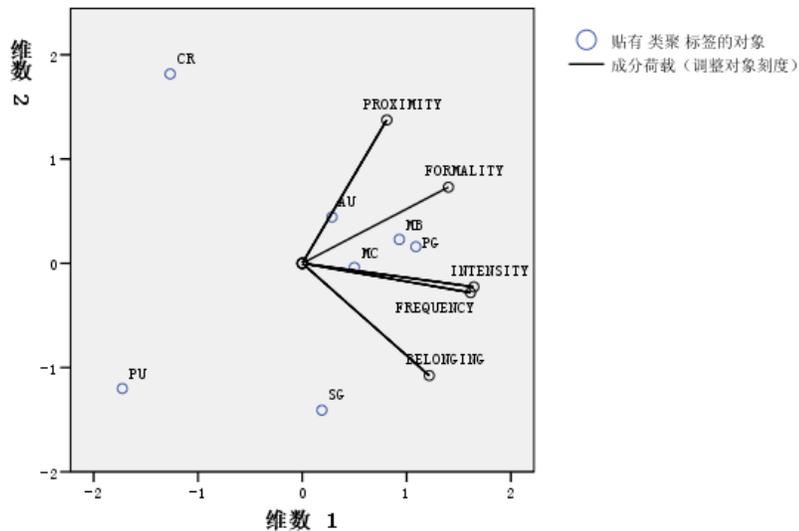
此图显示了成份载荷图。该图中的矢量（线）相对较长，这表示前两个维度贡献了所有已量化变量的大部分方差。在第一个维度上，所有变量都具有高（正）成份载荷。第二个维度主要与已量化变量归属感和物理亲近度相关，但方向相反。这意味着在第二个维度中具有较大负得分的对象在归属感变量中具有高分，而在物理亲近度变量中则具有低分。因此，第二个维度揭示了这两个变量之间的对立关系，同时表明这两个变量与已量化的变量交往强度和交往频率几乎没有关系。

图片 10-13  
成分载入



要检查对象与变量之间的关系，请查看对象与成份载荷的双标图。变量的矢量指向该变量的最高类别的方向。例如，对于物理亲近度和归属感变量，其最高类别分别为近和高。因此，观众的特征为物理亲近度“近”和归属感“无”，而次级群体的特征为物理亲近度“远”和归属感“高”。

图片 10-14  
双标图



## 附加维度

增大维数将会增加所贡献的变差量，并可以揭示低维数解中所隐藏的差异。在前面提到过，在二维中无法将组织群体与初级群体区分开。不过，如果增加维数，就可以区分这两个群体。

### 运行分析

- ▶ 要获取三维解，请调用“分类主成份”对话框。
- ▶ 键入 3 作为解的维数。
- ▶ 在“分类主成份”对话框中单击确定。

## 模型摘要

图片 10-15  
模型汇总

维	Cronbach's Alpha	方差考虑情况	
		总计(特征值)	方差的 %
1	.885	3.424	68.480
2	-.232	.844	16.871
3	-.459	.732	14.649
总计	1.000 <sup>a</sup>	5.000	99.999

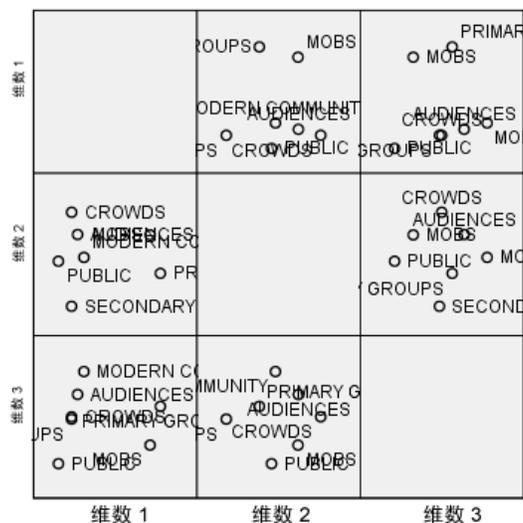
a. 总 Cronbach's Alpha 基于总特征值。

三维解具有特征值 3.424、0.844 和 0.732，贡献了几乎全部方差。

## 对象得分

用散点图矩阵绘制三维解的对象得分。在散点图矩阵中，每个维度分别按照二维散点图序列中的其他所有维度进行绘制。注意，三个维度中的前两个特征值与二维解中的特征值并不相同；也就是说，三维解不是嵌套的。由于目前维度 2 和维度 3 的特征值小于 1（给出的 Cronbach 的 alpha 为负），因此应首选二维解。此处提供三维解仅是为了进行说明。

图片 10-16  
三维对象得分散点图矩阵



图的第一排显示第一个维度将初级群体和组织群体与其他团体区分开。您会注意到，在第一排中的任一图中，对象沿垂直轴的顺序未改变；其中每个图都将维度 1 作为 y 轴。

使用中间一排图可以解释维度 2。第二个维度与两维解略有不同。在以前的两维解中，第二个维度分为三个不同的块；但在现在的三维解中，对象沿轴的展开程度更大。

第三个维度可帮助将组织群体与初级群体区分开，而在二维解中则未能区分。

进一步深入观察维度 2 与维度 3 以及维度 1 与维度 2 图。在维度 2 和维度 3 所定义的象限，对象大致构成了一个矩形，观众、现代社区、次级群体和公共是矩形的顶点。在此象限中，组织群体和初级群体分别是公众-观众和次级群体-现代社区的

凸组合。但前面提到过，它们在维度 1 上与其他群体区分开。听众与其他群体在维度 1 上没有区分开，并且是观众和现代社区的组合。

## 成份载荷

图片 10-17  
三维成份载荷

	维		
	1	2	3
INTENSITY	.980	-.005	-.201
FREQUENCY	.521	-.643	.561
BELONGING	.980	-.002	-.197
PROXIMITY	.519	.656	.549
FORMALITY	.981	.004	-.193

仅了解对象的区分方式并不能揭示变量与维度之间的对应关系。这需要使用成份载荷。第一个维度主要与归属感、交往强度和关系正式性相对应；第二个维度与交往频率和物理接近度区分开，而第三个维度则将这些变量与其他变量区分开。

## 示例：进食障碍症

进食障碍是耗弱心神的疾病，其症状表现为进食机能紊乱、严重的体形焦虑和体重控制强迫症。每年有数以百万人会患上此疾病，尤其是青少年更是深受其害。此疾病是可以治疗的，但大多数疗法需早期确诊才能取得疗效。

医疗人员可以尝试通过心理和体格检查评估来对进食障碍进行诊断。但时，由于目前没有确定厌食/暴食行为的标准症状参照，因此很难确诊患者属于多种不同进食障碍症中的哪一种。是否可以将患者症状明确分为四类？患者具有哪些共同的症状？

为了回答这些问题，研究人员 (Van der Ham, Meulman, Van Strien, 和 Van Engeland, 1997) 对 55 名已知存在进食障碍的青少年进行了调查，其中的数据如下表所示。

表 10-2  
患者诊断

诊断	患者数
神经性厌食症	25
神经性暴食症	9
神经性厌食后暴食症	14
非典型进食障碍症	7
总计	55

其中每名患者每年都将进行四次检查，因此总观测数为 220。在每次观测期间，将对这些患者按下表所列出的 16 种症状逐项评分。但 71 号和 76 号患者的症状得分均在时间点 2 缺失，47 号患者的症状得分在时间点 3 缺失，因此有效观测数为 217。此数据可在 anorectic.sav 中找到。有关详细信息，请参阅附录 A 中的样本文件中的 IBM SPSS Categories 21。

表 10-3  
度量健康情况的改进 Morgan-Russell 尺度分量

变量名	变量标签	下限 (得分 1)	上限 (得分 3 或 4)
weight	Body weight	Outside normal range	Normal
mens	Menstruation	Amenorrhea	Regular periods
fast	Restriction of food intake (fasting)	Less than 1200 calories	Normal/regular meals
binge	Binge eating	Greater than once a week	No bingeing
vomit	Vomiting	Greater than once a week	No vomiting
purge	Purging	Greater than once a week	No purging
hyper	Hyperactivity	Not able to be at rest	No hyperactivity
fami	Family relations	Poor	Good
eman	Emancipation from family	Very dependent	Adequate
frie	Friends	No good friends	Two or more good friends
school	School/employment record	Stopped school/work	Moderate to good record
satt	Sexual attitude	Inadequate	Adequate
sbeh	Sexual behavior	Inadequate	Can enjoy sex
mood	Mental state (mood)	Very depressed	Normal
preo	Preoccupation with food and weight	Complete	No preoccupation
body	Body perception	Disturbed	Normal

由于该调查的目的是为了确定症状与不同的进食障碍种类之间的关系，因此适合采用主成分分析方法。此外，由于是按有序刻度对症状进行评分，因此相对经典主成分分析，更适合使用分类主成份分析。

## 运行分析

为正确观察每个诊断的病程结构，需要将投影质心表的结果作为散点图的数据提供。可以使用输出管理系统完成此操作。

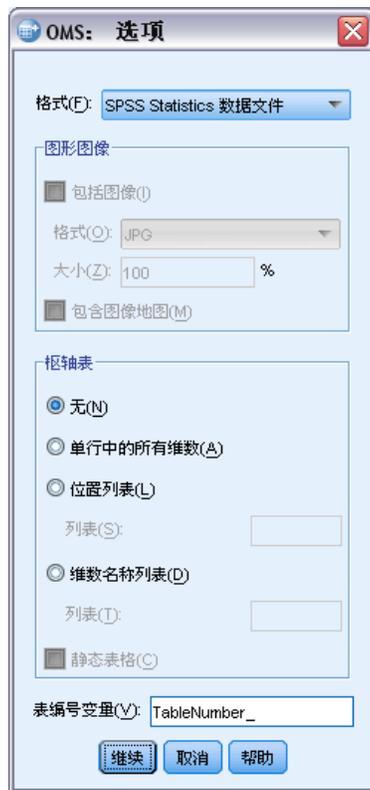
- ▶ 要启动 OMS 请求，请从菜单中选择：  
实用程序 > OMS 控制面板...

图片 10-18  
输出管理系统控制面板



- ▶ 选择表作为输出类型。
- ▶ 选择 CATPCA 作为命令。
- ▶ 选择 Projected Centroids 作为表类型。
- ▶ 在“输出目标”组中选择文件，然后键入 projected\_centroids.sav 作为文件名。
- ▶ 单击选项。

图片 10-19  
“选项”对话框



- ▶ 选择 IBM SPSS Statistics 数据文件作为输出格式。
- ▶ 键入 TableNumber\_1 作为表编号变量。
- ▶ 单击继续。

图片 10-20  
输出管理系统控制面板



- ▶ 单击添加。
- ▶ 单击确定, 然后再单击确定以确认 OMS 会话。  
输出管理系统随即设置为将 Projected Centroids 表的结果写入文件 projected\_centroids.sav。
- ▶ 要生成此数据集的分类主成份输出, 请从菜单中选择:  
分析 > 降维 > 最优尺度...

图片 10-21  
“最优尺度”对话框



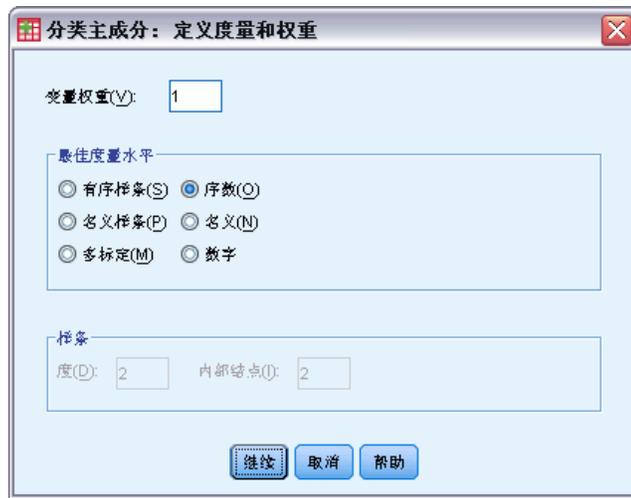
- ▶ 在“最佳度量水平”组中，选择有些变量不是无序多分类。
- ▶ 单击定义。

图片 10-22  
分类主要成分对话框



- ▶ 选择从 Body weight 到 Body perception 的变量作为分析变量。
- ▶ 单击定义尺度和权重。

图片 10-23  
“定义尺度和权重”对话框



- ▶ 选择有序作为最佳度量水平。
- ▶ 单击继续。
- ▶ 选择 Time/diagnosis interaction 作为补充变量，然后在“分类主成份”对话框中单击定义度量。

图片 10-24  
“定义尺度”对话框



- ▶ 选择多标定作为最佳度量水平。
- ▶ 单击继续。

图片 10-25  
分类主要成分对话框



- ▶ 选择从 Time of interview 到 Patient number 的变量作为标记变量。
- ▶ 单击选项。

图片 10-26  
“选项”对话框



- ▶ 选择按变量名称或值标注图。
- ▶ 单击继续。
- ▶ 在“分类主成份”对话框中单击输出。

图片 10-27  
“输出”对话框



- ▶ 在“表”组中选择对象得分。
- ▶ 请求 tidi 的类别量化。
- ▶ 选择以下包括的类别：time、diag 和 number。
- ▶ 单击继续。
- ▶ 在“分类主成份”对话框中单击保存。

图片 10-28  
“保存”对话框



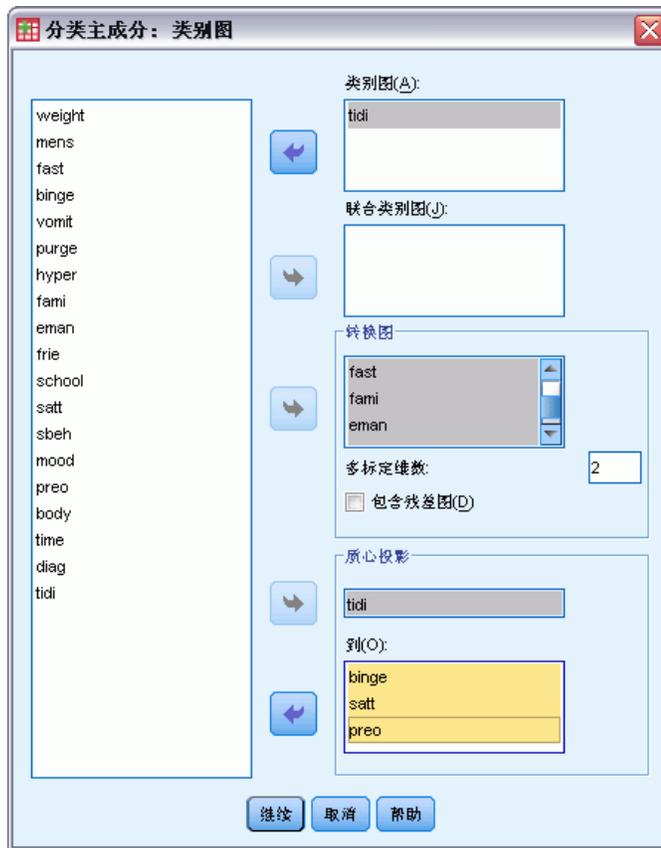
- ▶ 在“已转换的变量”组中，选择保存到活动数据集。
- ▶ 单击继续。
- ▶ 在“分类主成份”对话框中单击对象。

图片 10-29  
“对象和变量图”对话框



- ▶ 选择按照变量标记对象。
- ▶ 选择 time 和 diag 作为用于标记对象的变量。
- ▶ 单击继续。
- ▶ 在“分类主成份”对话框中单击类别。

图片 10-30  
“类别图”对话框



- ▶ 请求 tidi 的类别图。
- ▶ 请求 weight 到 body 变量的转换图。
- ▶ 选择将 tidi 的质心投影到 binge、satt 和 preo 上。
- ▶ 单击继续。
- ▶ 在“分类主成份”对话框中单击确定。

该过程将生成主体得分（均值 0 和单位方差）和类别量化（使主体得分和已转换的变量的平方均值相关系数最大化）。在现有分析中，类别量化仅限于反映有序信息。

最后，要将投影质心表信息写入 projected\_centroids.sav 中，需要结束 OMS 请求。调用 OMS 控制面板。

图片 10-31  
输出管理系统控制面板

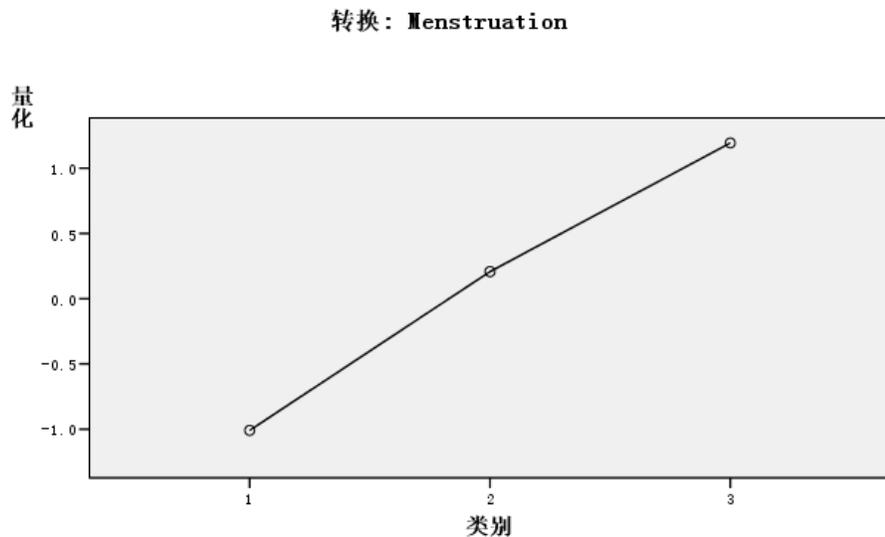


- ▶ 单击终止。
- ▶ 单击确定，然后再单击确定进行确认。

## 转换图

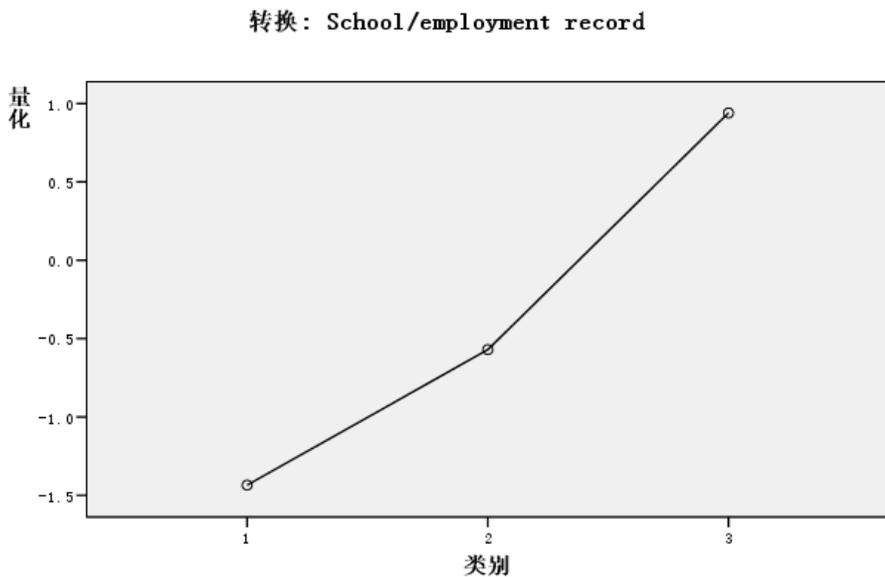
变换图在水平轴上显示初始类别编号；在垂直轴上给出最优定量。

图片 10-32  
Menstruation 变换图



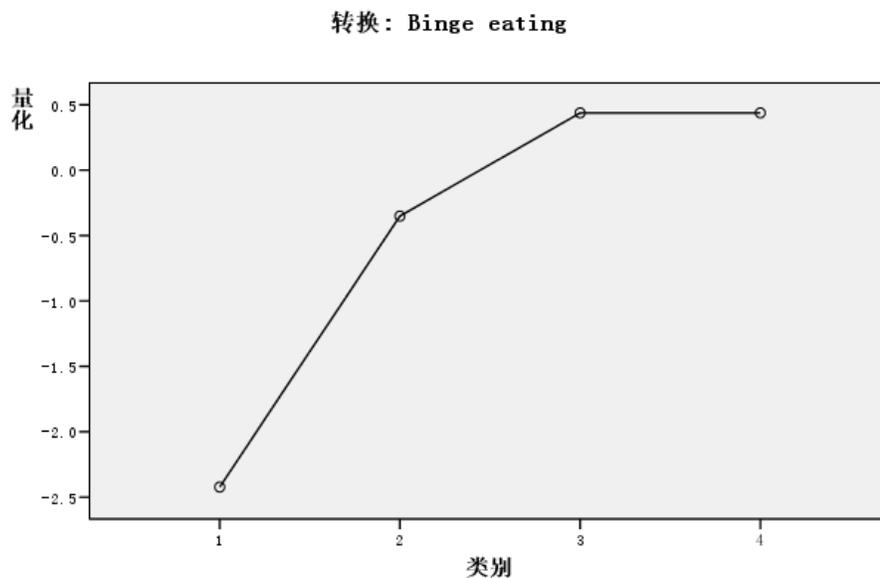
有些变量（如 Menstruation）获得了接近线性变换，因此在此分析中可将这些变量视为数值尺度级别。

图片 10-33  
School/employment record 的变换图



有些变量（如 School/employment record）的定量未获得线性变换，因此应视为有序尺度级别。类别 2 与类别 3 之间的差异比类别 1 与类别 2 之间的差异要显著的多。

图片 10-34  
Binge eating 的变换图



在 Binge eating 变量的量化过程中生成了一个有趣的个案。类别 1 到类别 3 所获得的变换是线性的；但类别 3 和类别 4 的量化值相等。此结果表明患者的类别 3 和类别 4 得分没有差异，并表明应通过将类别 4 重新编码为类别 3 在两成分解中使用数值尺度级别。

## 模型摘要

图片 10-35  
模型汇总

维	Cronbach's Alpha	方差考虑情况	
		总计(特征值)	方差的 %
1	.874	5.550	34.690
2	.522	1.957	12.234
总计	.925 <sup>a</sup>	7.508	46.924

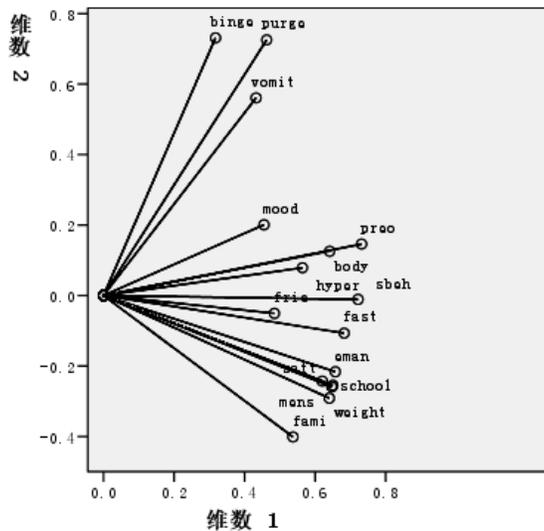
a. 总 Cronbach's Alpha 基于总特征值。

要了解模型与数据拟合的程度，请查看模型摘要。大约 47% 的总方差由两成分模型解释，其中 35% 归因于第一个维度，另 12% 归因于第二个维度。因此，各对象级别上近一半的变差由两成份模型解释。

## 成份载荷

要着手解释解的两个维度，请查看成份载荷。所有变量在第一个维度都具有正的成份载荷，这意味着有一个公共因子与所有变量正相关。

图片 10-36  
成份载荷图



第二个维度区分各变量。变量 Binge eating、Vomiting 和 Purging 在第二个维度中构成了具有较大正负荷的变量束。这些症状通常被视为暴食行为的代表。

变量 Emancipation from family、School/employment record、Sexual attitude、Body weight 和 Menstruation 构成了另一变量束，而且还可以将变量 Restriction of food intake (fasting) 和 Family relations 包含在此变量束中，因为这些变量的矢量与主聚类很接近，并且这些变量实际上被视为厌食症状 (fasting、weight、menstruation) 或心理症状 (emancipation、school/work record、sexual attitude、family relations)。此变量束中的各矢量相应地与暴食、呕吐和腹泻这些变量的矢量正交 (垂直)，这意味着此变量组与暴食变量组无关。

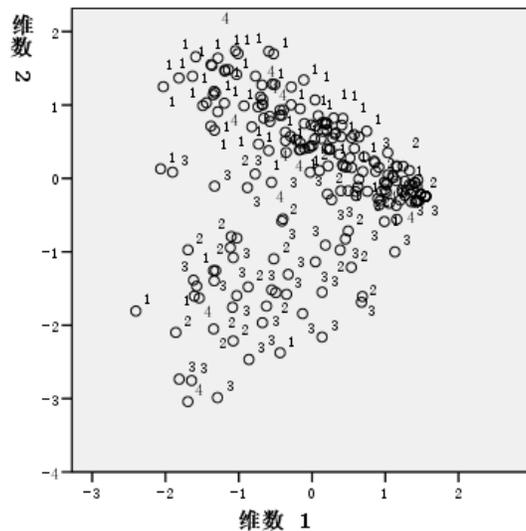
变量 Friends、Mental state (mood) 和 Hyperactivity 看起来与该解不是十分拟合。通过观察图中每个矢量的长度，可以看出这一点。由于给定变量的矢量长度与其拟合度相对应，因此这些变量的矢量最短。根据两成份解，可能会将这些变量从所提出的进食障碍症状中删除。但是，这些变量在更高维度的解中可能会有较好地拟合度。

变量 Sexual behavior、Preoccupation with food and weight 和 Body perception 在理论上构成了另一个症状组，这与患者自身的身体体验有关。虽然这些变量与前两个正交变量束相关，但这些变量的矢量很长，并且与第一个维度紧密相关，因此可以提供有关“公共”因子的一些有用信息。

## 对象得分

下图显示了对象得分图，在该图中用主体的诊断类别来标记主体。

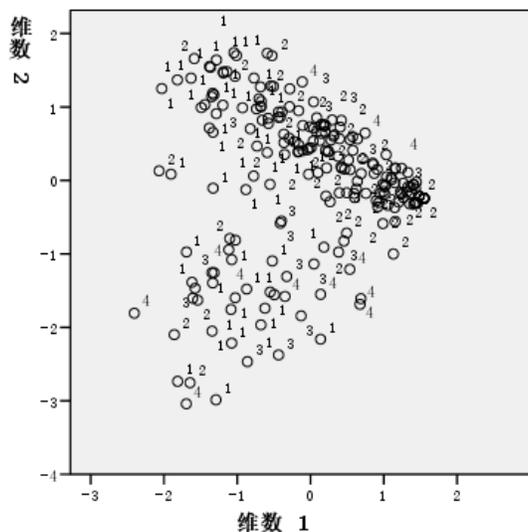
图片 10-37  
按诊断标注的对象得分图



该图不适于解释第一个维度，因为在该维度上诊断未将患者区分开。但是，该图中包含有关第二个维度的信息。厌食主体（1）和患有非典型进食障碍症的患者（4）构成了一组，位于一些暴食类主体（2 和 3）的上方。因此，第二个维度将暴食患者与其他患者区分开，这与前面部分所看到的一样（暴食变量束中的变量在第二个维度具有较大的正成份载荷）。如果通常与暴食相关的症状的成份载荷在第二个维度具有较大值，则这会很有意义。

下图显示了一个对象得分图，其中的主体用其诊断时间标记。

图片 10-38  
按问诊时间标记的对象得分



按时间标记对象得分揭示了第一个维度与时间相关，因为诊断时间通常是从时间点 1 向左推进，而其他对象则是向右推进。注意，可以通过以下方法连接该图中的时间点：保存这些对象得分并创建一个在 x 轴上使用维度 1 得分、在 y 轴上使用维度 2 得分的散点图，然后使用患者编号设置标记。

通过将按时间标记的对象得分图与按诊断标记的对象得分图进行比较，可以更深入地了解异常对象。例如，在按时间标记的图中，有一个其诊断位于时间点 4 的患者位于该图中的所有其他点的左侧。此对象是异常对象，因为点的常规趋势是：时间越长，越偏向右侧。有趣的是，在时间上处于不当位置的该点似乎也具有异常诊断，因为该患者患有厌食症，而其得分却位于暴食聚类中。通过查看对象得分表，可以确定此对象是 43 号患者，其诊断为神经性厌食症，其对象得分如下表所示。

表 10-4  
43 号患者的对象得分

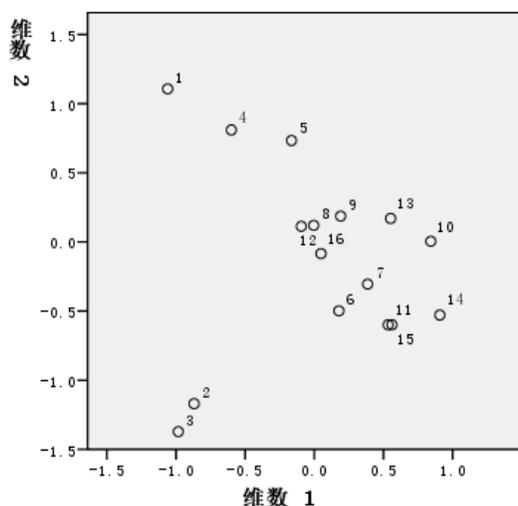
时间	维度 1	维度 2
1	-2.031	1.250
2	-2.067	0.131
3	-1.575	-1.467
4	-2.405	-1.807

该患者在时间点 1 的得分表明他患有典型厌食症，因为他在维度 1 中具有较大的负得分（这与较差体形相对应），而在维度 2 中具有正得分（这与厌食症状或较差的心理行为相对应）。不过，与多数患者不同的是，在维度 1 中没有或几乎没有渐近。在维度 2 中，明显有一些渐近逼近“正常”（在 0 左右，即介于厌食和暴食行为之间），但该患者却转为表现暴食症状。

## 检查病程结构

要了解有关两个维度如何与四个诊断类别和四个时间点相关的更多信息，请通过 Patient diagnosis 的四个类别和 Time of interview 的四个类别的交叉分类创建一个补充变量 Time/diagnosis interaction。这样，Time/diagnosis interaction 将具有 16 个类别，其中第 1 个类别表示第 1 次接受诊断的神经性厌食症患者。第 5 个类别表示第 2 次接受诊断的神经性厌食症，依此类推……，第 16 个类别即表示第 4 次接受诊断的非典型进食障碍症患者。通过使用变量 Time/diagnosis interaction，可以研究不同组随时间变化的病程。为该变量指定了一个多名义尺度级别，并用下图显示类别点。

图片 10-39  
时间/诊断交叉组合的类别点



在此图中，有些结构是显而易见的：位于时间点 1 的诊断类别明确地将神经性厌食症和非典型进食障碍症与第二个维度中的神经性暴食症和神经性厌食后暴食症区分开。除此之外，图中较难再看出其他模式了。

不过，通过基于定量创建散点图，可以很容易地看出一些模式。为此，请从菜单中选择：

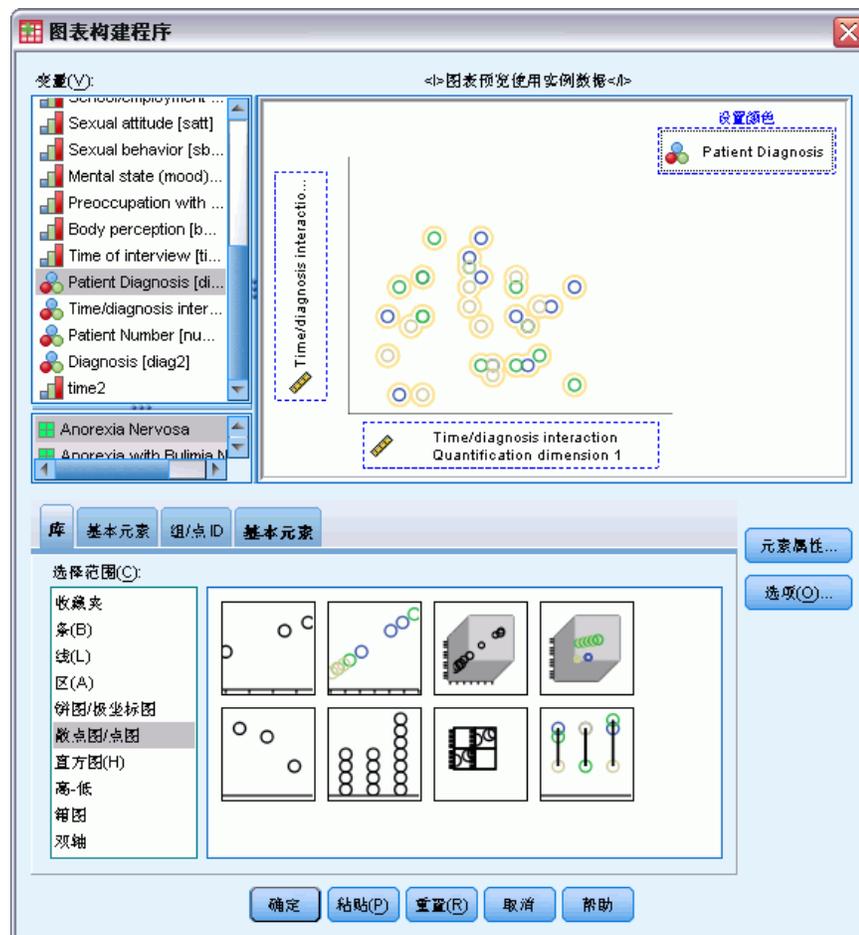
图形 > 图表构建程序...

图片 10-40  
“散点图/点图”库



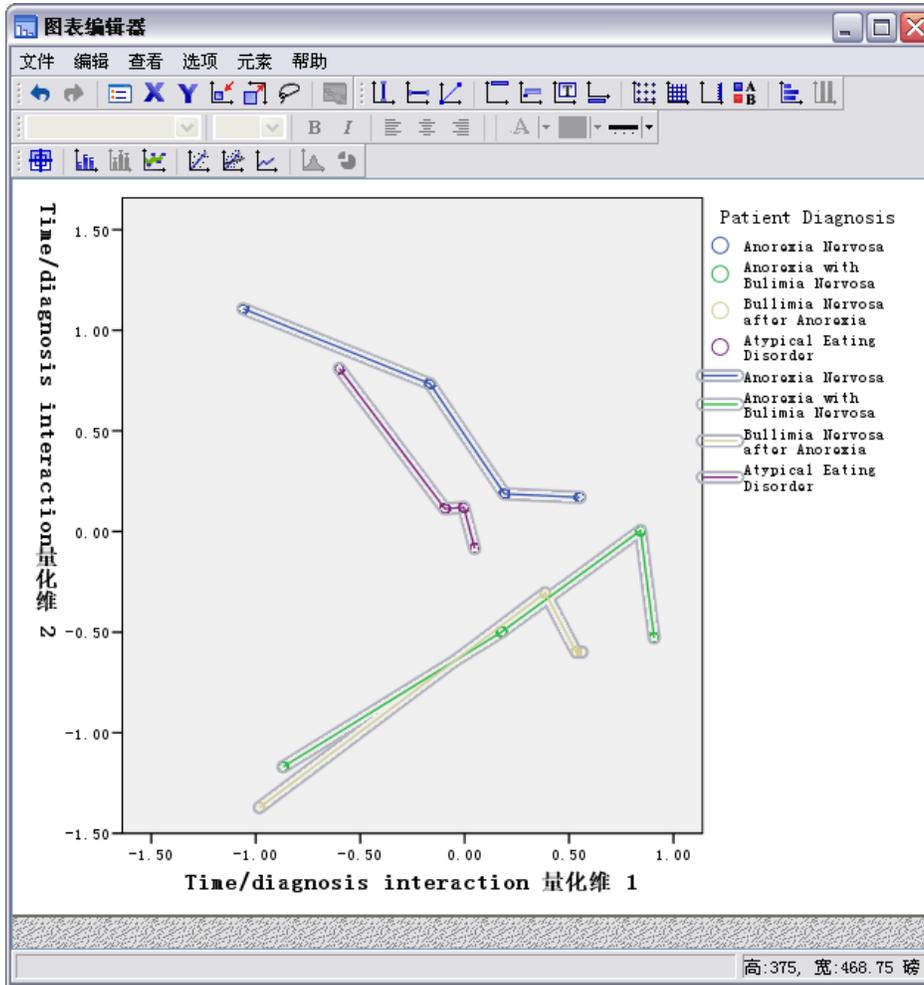
- 选择散点图/点图库并选择“分组的散点图”。

图片 10-41  
图表构建程序



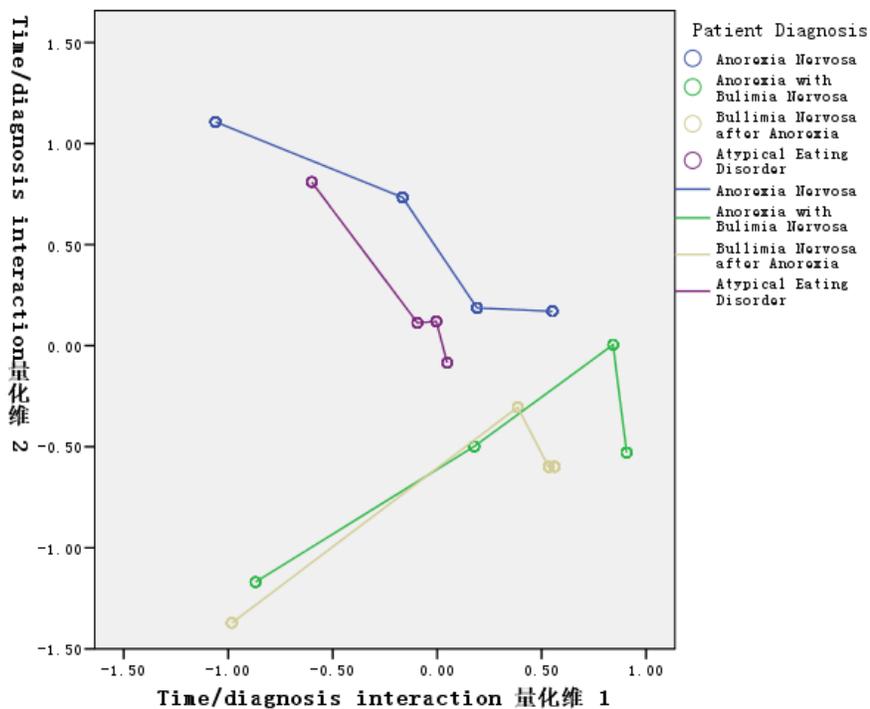
- ▶ 选择 Time/diagnosis interaction 量化维 2 作为 y 轴变量，并选择 Time/diagnosis interaction 量化维 1 作为 x 轴变量。
- ▶ 选择按 Patient Diagnosis 设置颜色。
- ▶ 单击确定。

图片 10-42  
疾病过程结构



- ▶ 然后将各点连接起来，双击该图，然后单击图表编辑器中的添加内插线工具。
- ▶ 关闭图表编辑器。

图片 10-43  
疾病过程结构



通过连接每个随时间变化的诊断类别的类别点，则可通过模式直接看出第一个维度与时间相关，而第二个维度与诊断相关，就像前面通过对象得分图确定的一样。

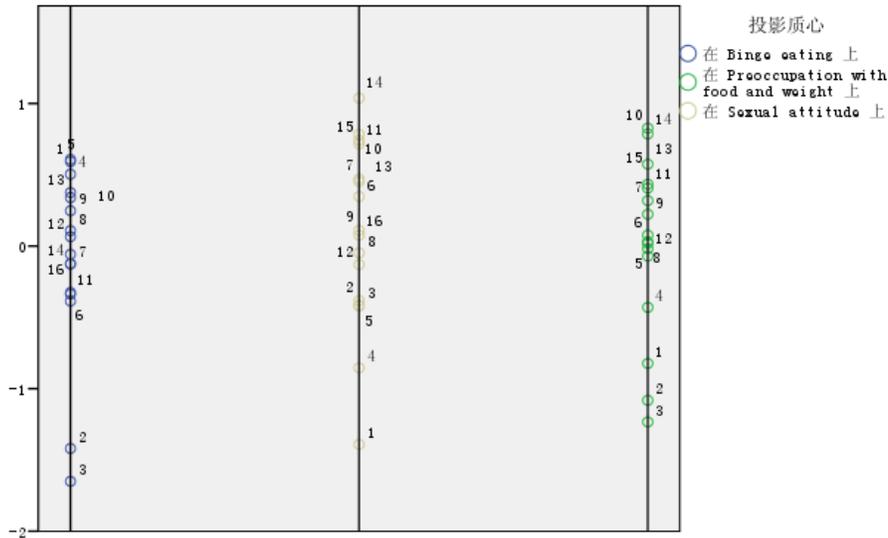
不过，该图进一步表明病症随时间而变得越来越相似。此外，在所有组中，时间点 1 和 2 之间的渐近最大；厌食症患者组显示时间点 2 和 3 之间有一些渐近，但其他组则几乎没有渐近。

### 选定变量的差异发展

每个症状束中有一个变量由选择为该束“代表”的成份载荷进行标识。将从暴食症状束中选择暴食变量；从厌食/心理症状束中选择性态度变量；从第三个症状束中选择身体知觉变量。

为了检测可能的不同病程，在下图中计算并绘制了 Time/diagnosis interaction 在变量 Binge eating、Sexual attitude 和 Preoccupation with food and weight 上的投影。

图片 10-44  
时间/诊断交叉组合在变量暴食、性态度和对食物和体重的注重程度上的投影质心

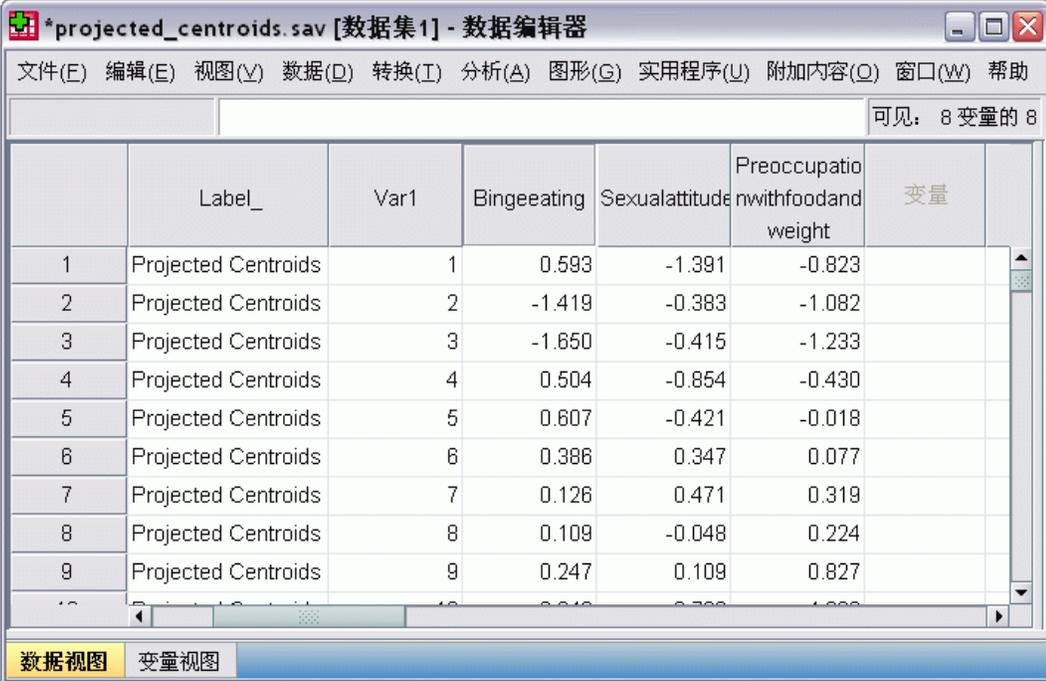


该图显示了在第一个时间点，暴食变量将暴食症患者（2 和 3）与其他患者区分开；性态度变量将厌食症患者和非典型患者（1 和 4）与其他患者区分开；身体知觉变量实际上并没有区分患者。在许多应用中，该图足以描述症状和诊断之间的关系，但由于多时间点的复杂性，该图比较混乱。

为了查看这些投影随时间的变化，需要绘制投影质心表的内容。这一点可以通过 OMS 请求将此信息保存到 projected\_centroids.sav 中来实现。

图片 10-45

Projected\_centroids.sav



	Label_	Var1	Bingeating	Sexualattitude	Preoccupatio with food and weight	变量
1	Projected Centroids	1	0.593	-1.391	-0.823	
2	Projected Centroids	2	-1.419	-0.383	-1.082	
3	Projected Centroids	3	-1.650	-0.415	-1.233	
4	Projected Centroids	4	0.504	-0.854	-0.430	
5	Projected Centroids	5	0.607	-0.421	-0.018	
6	Projected Centroids	6	0.386	0.347	0.077	
7	Projected Centroids	7	0.126	0.471	0.319	
8	Projected Centroids	8	0.109	-0.048	0.224	
9	Projected Centroids	9	0.247	0.109	0.827	
10	Projected Centroids	10	0.018	0.788	1.088	

变量 Bingeating、Sexualattitude 和 Preoccupationwithfoodandweight 包含在每个相关症状上投影得到的质心值。个案编号（1 至 16）与时间/诊断交叉组合相对应。将需要计算将“时间”和“诊断”值区分开的新变量。

- ▶ 从菜单中选择：  
转换 > 计算变量...

图片 10-46  
“计算变量”对话框



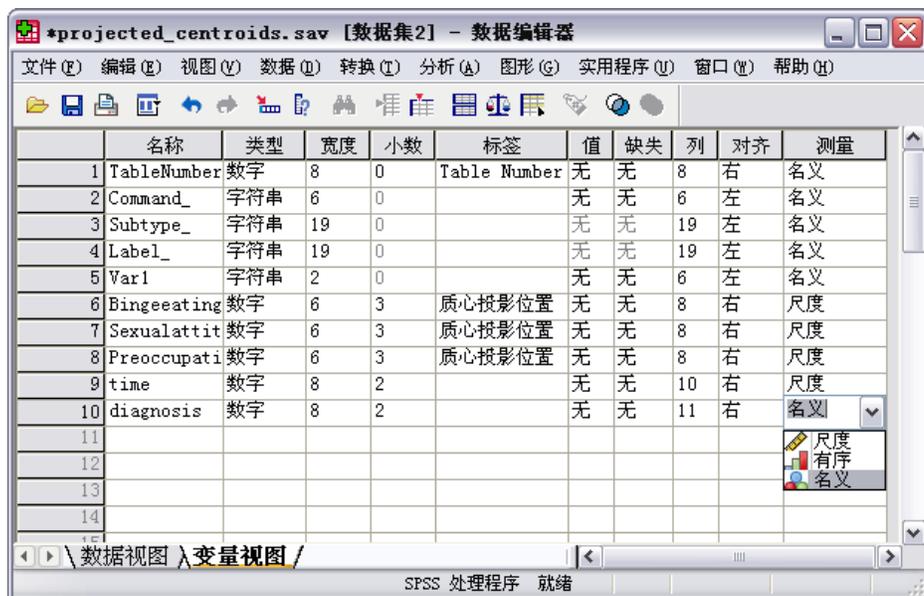
- ▶ 键入 time 作为目标变量。
- ▶ 键入  $\text{trunc}((\$casenum-1)/4) + 1$  作为数值表达式。
- ▶ 单击确定。

图片 10-47  
“计算变量”对话框



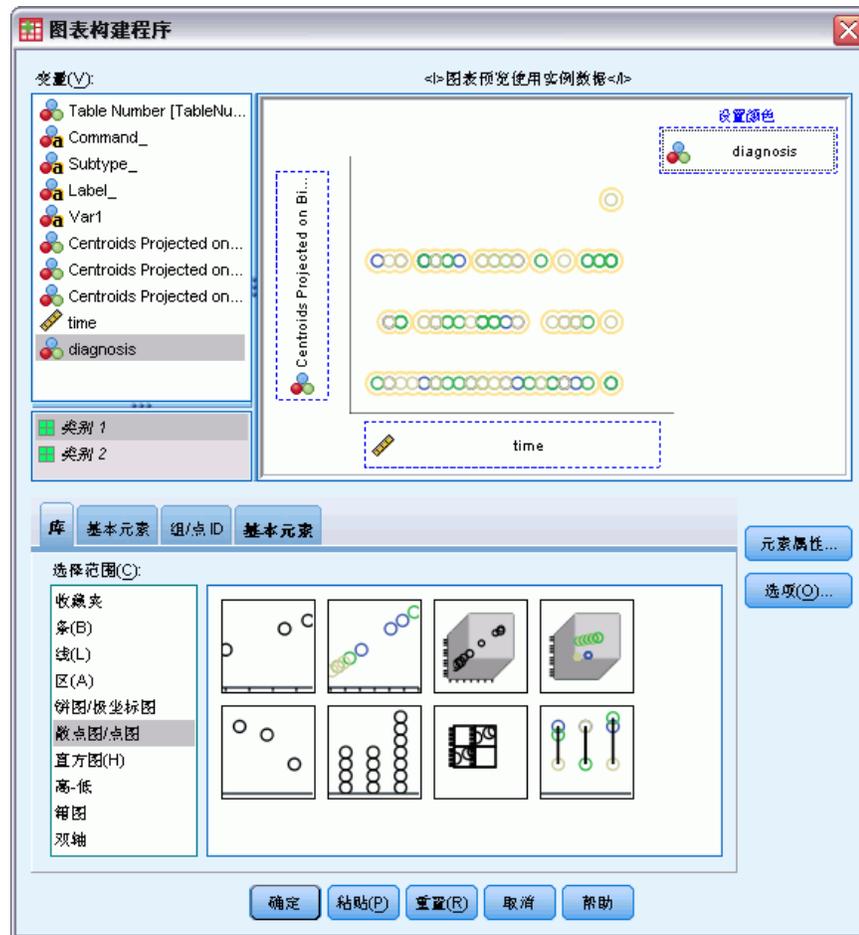
- ▶ 调用“计算变量”对话框。
- ▶ 键入 diagnosis 作为目标变量。
- ▶ 键入  $\text{mod}(\$casenum-1, 4) + 1$  作为数值表达式。
- ▶ 单击确定。

图片 10-48  
Projected\_centroids.sav



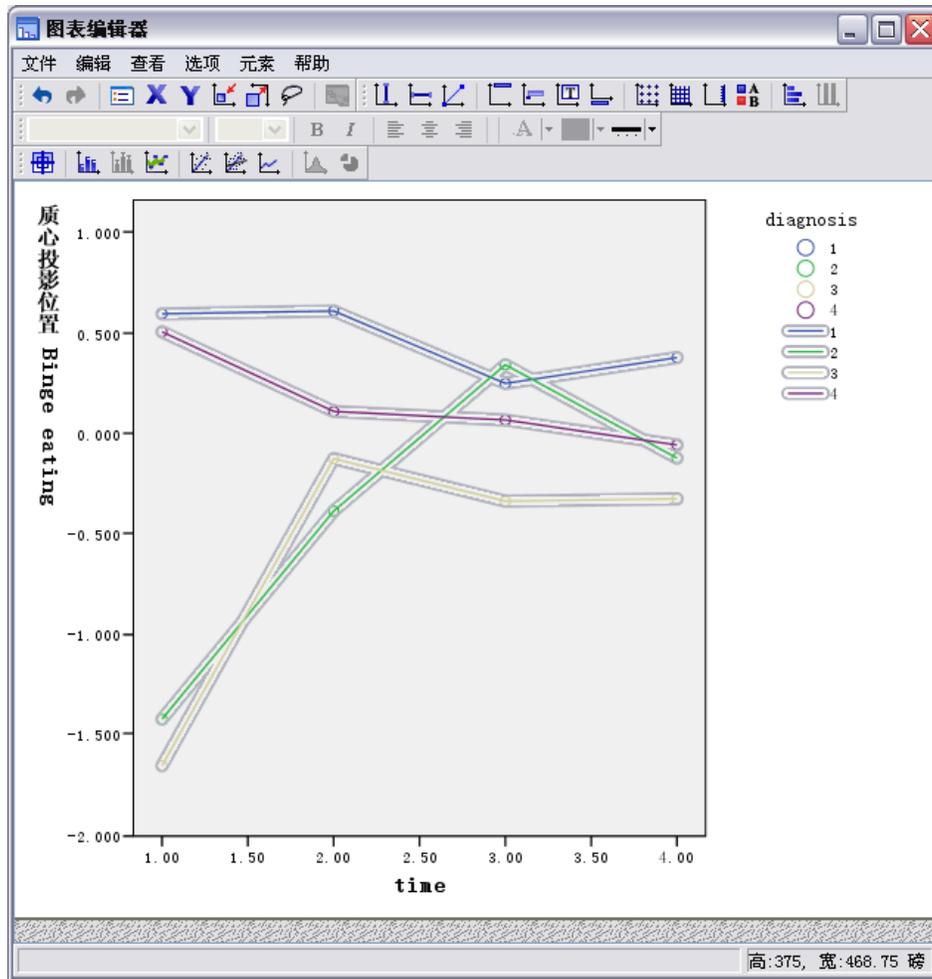
在变量视图中，将 diagnosis 的测量从“尺度”更改为“名义”。

图片 10-49  
图表构建程序



- ▶ 最后，要查看诊断时间在暴食变量上的投影质心随时间的变化，请调用图表生成器，然后单击重置清除以前的选择。
- ▶ 选择散点图/点图库并选择“分组的散点图”。
- ▶ 选择 Centroids Projected on Binge eating 作为 y 轴变量，并选择 time 作为 x 轴变量。
- ▶ 选择按 diagnosis 设置颜色。
- ▶ 单击确定。

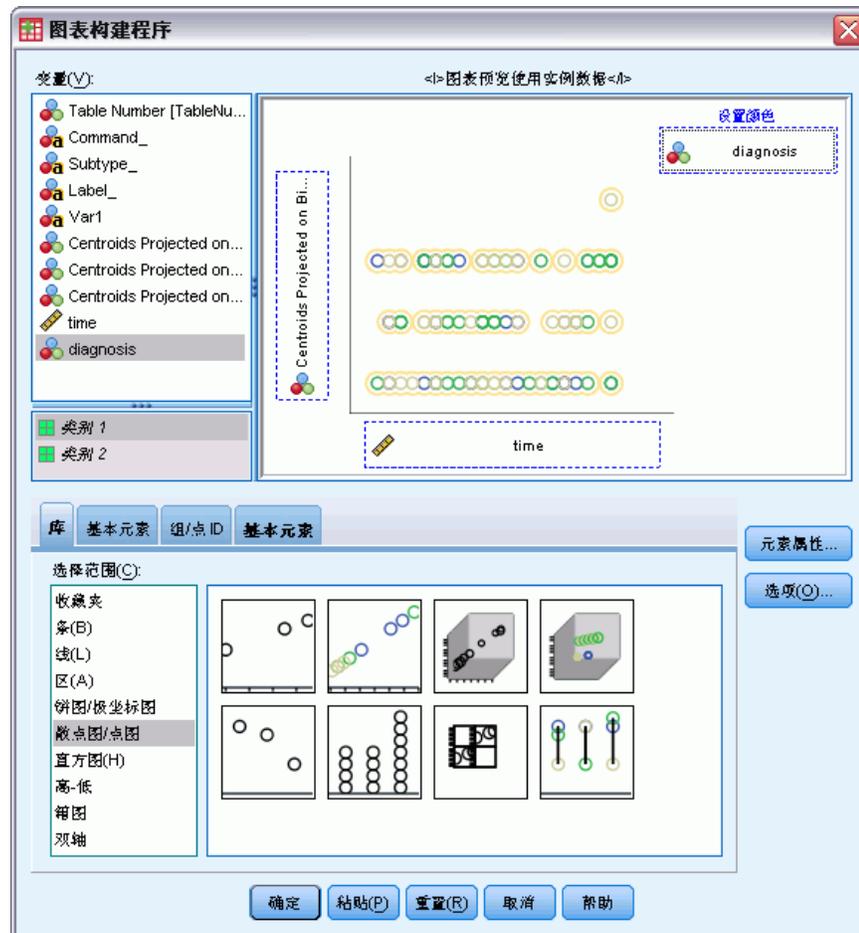
图片 10-50  
诊断时间在暴食变量上的投影质心随时间的变化



- ▶ 然后将各点连接起来，双击该图，然后单击图表编辑器中的添加内插线工具。
- ▶ 关闭图表编辑器。

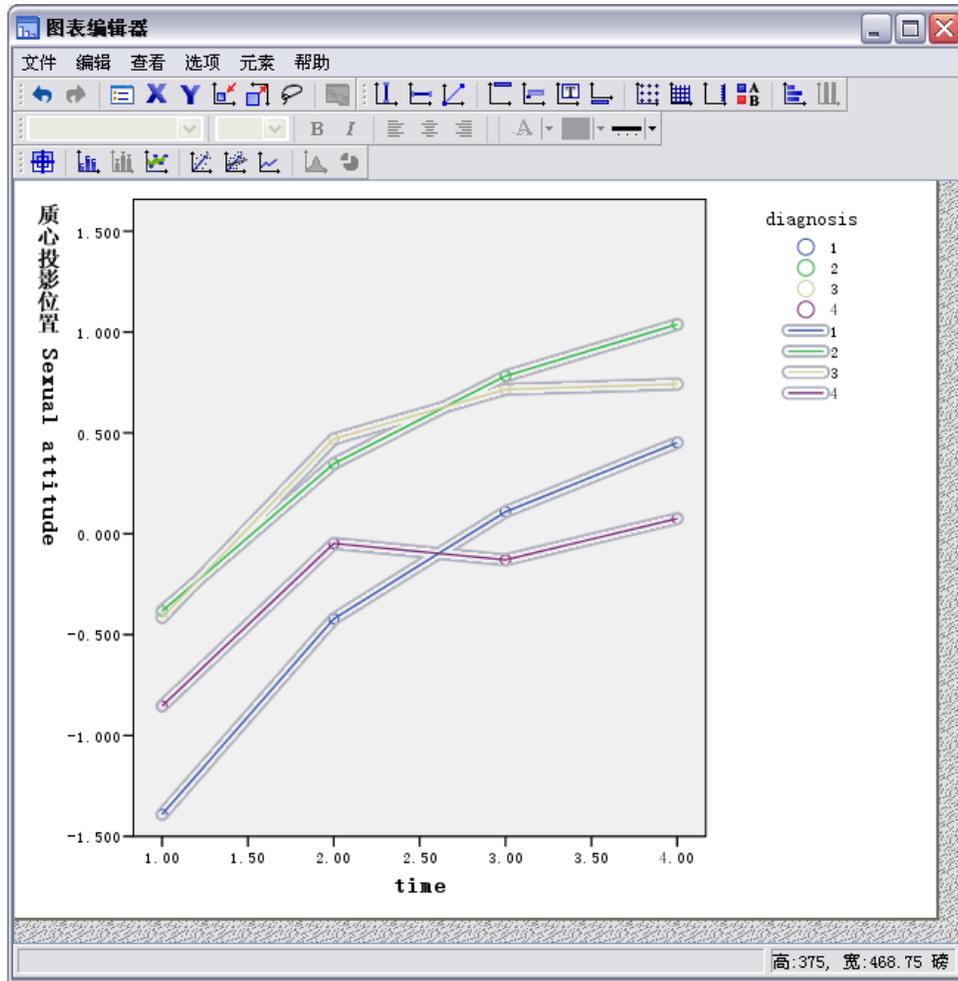
就暴食而言，可能明显地看到厌食组与暴食组具有不同的起始值。此差异会随时间缩小，因为厌食组几乎没有变化，而暴食组却显现出了渐近。

图片 10-51  
图表构建程序



- ▶ 调用图表生成器。
- ▶ 取消选择 Centroids Projected on Binge eating 作为 y 轴变量，然后选择 Centroids Projected on Sexual attitude 作为 y 轴变量。
- ▶ 单击确定。

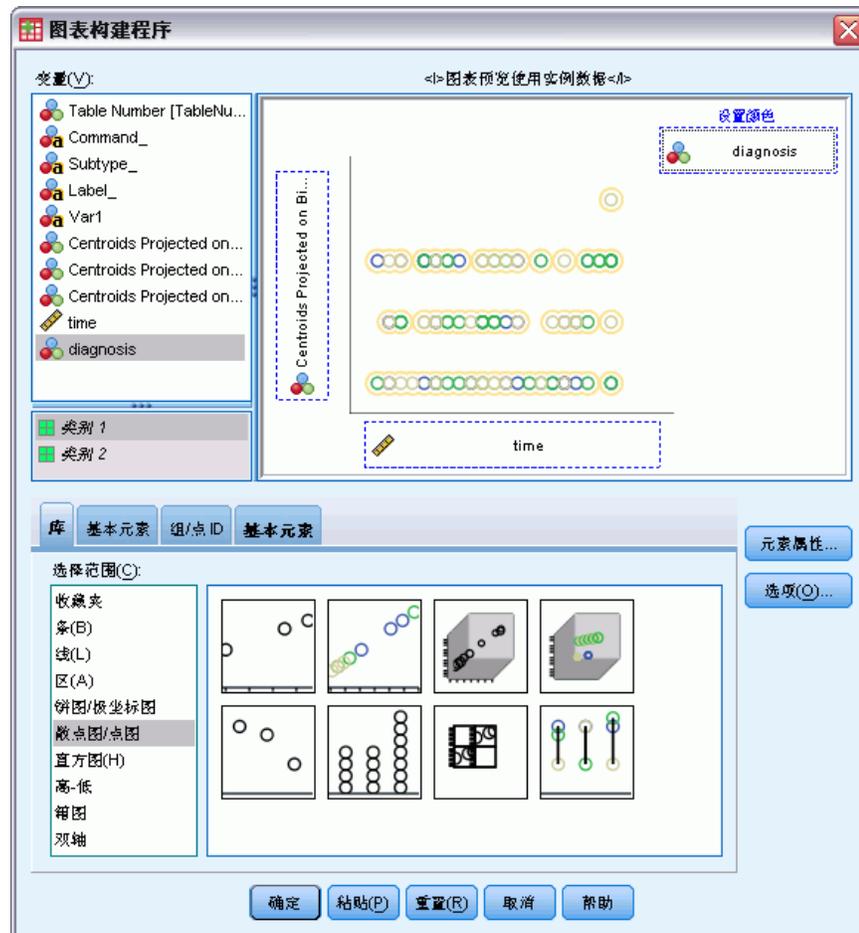
图片 10-52  
诊断时间在性态度变量上的投影质心随时间的变化



- ▶ 然后将各点连接起来：双击该图，然后单击图表编辑器中的添加内插线工具。
- ▶ 关闭图表编辑器。

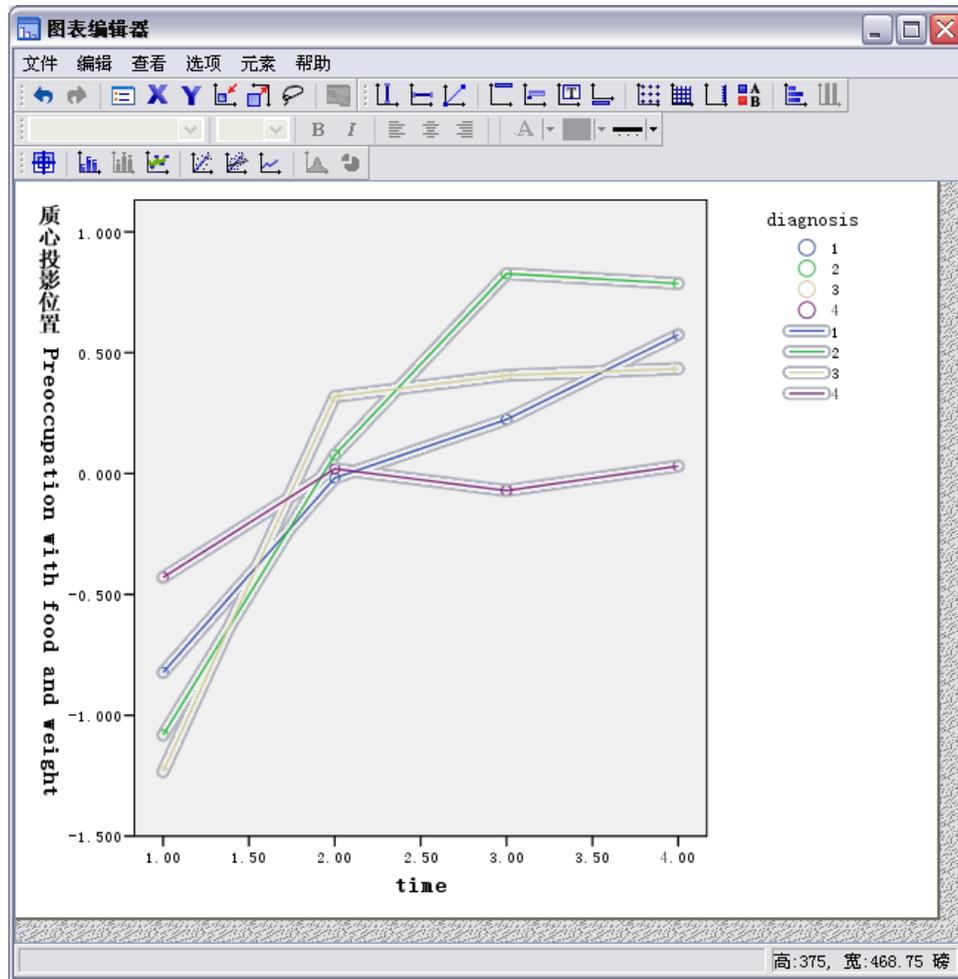
就性态度而言，四个轨迹线与时间或多或少是平行的，并且所有组都显现出渐近。但暴食组比厌食组具有更高（更好）的得分。

图片 10-53  
图表构建程序



- ▶ 调用图表生成器。
- ▶ 取消选择 Centroids Projected on Sexual attitude 作为 y 轴变量，然后选择 Centroids Projected on Preoccupation with food and weight 作为 y 轴变量。
- ▶ 单击确定。

图片 10-54  
诊断时间在身体知觉变量上的投影质心随时间的变化



- ▶ 然后将各点连接起来：双击该图，然后单击图表编辑器中的添加内插线工具。
- ▶ 关闭图表编辑器。

身份知觉是一个表示核心症状（即四个不同的组共有的症状）的变量。除非典型进食障碍症患者之外，厌食组和两个暴食组在开头和结尾时具有非常类似的级别。

## 推荐参考

有关分类主成份分析的更多信息，请参见以下内容：

De Haas, M., J. A. Algera, H. F. J. M. Van Tuijl, 和 J. J. Meulman. 2000. Macro and micro goal setting: In search of coherence. *Applied Psychology*, 49, .

De Leeuw, J. 1982. Nonlinear principal components analysis. 输入: COMPSTAT Proceedings in Computational Statistics, Vienna: Physica Verlag.

- Eckart, C., 和 G. Young. 1936. The approximation of one matrix by another one of lower rank. *Psychometrika*, 1, .
- Gabriel(G), K. R. 1971. The biplot graphic display of matrices with application to principal components analysis. *Biometrika*, 58, .
- Gifi, A. 1985. PRINCALS. Research Report UG-85-02. Leiden: Department of Data Theory, University of Leiden.
- Gower, J. C., 和 J. J. Meulman. 1993. The treatment of categorical information in physical anthropology. *International Journal of Anthropology*, 8, .
- Heiser, W. J., 和 J. J. Meulman. 1994. Homogeneity analysis: Exploring the distribution of variables and their nonlinear relationships. 输入: *Correspondence Analysis in the Social Sciences: Recent Developments and Applications*, M. Greenacre, 和 J. Blasius, eds. New York: Academic Press.
- Kruskal, J. B. 1978. Factor analysis and principal components analysis: Bilinear methods. 输入: *International Encyclopedia of Statistics*, W. H. Kruskal, 和 J. M. Tanur, eds. New York: The Free Press.
- Kruskal, J. B., 和 R. N. Shepard. 1974. A nonmetric variety of linear factor analysis. *Psychometrika*, 39, .
- Meulman, J. J. 1993. Principal coordinates analysis with optimal transformations of the variables: Minimizing the sum of squares of the smallest eigenvalues. *British Journal of Mathematical and Statistical Psychology*, 46, .
- Meulman, J. J., 和 P. Verboon. 1993. Points of view analysis revisited: Fitting multidimensional structures to optimal distance components with cluster restrictions on the variables. *Psychometrika*, 58, .
- Meulman, J. J., A. J. Van der Kooij, 和 A. Babinec. 2000. New features of categorical principal components analysis for complicated data sets, including data mining. 输入: *Classification, Automation and New Media*, W. Gaul, 和 G. Ritter, eds. Berlin: Springer-Verlag.
- Meulman, J. J., A. J. Van der Kooij, 和 W. J. Heiser. 2004. Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. 输入: *Handbook of Quantitative Methodology for the Social Sciences*, D. Kaplan, ed. Thousand Oaks, Calif.: Sage Publications, Inc..
- Theunissen, N. C. M., J. J. Meulman, A. L. Den Ouden, H. M. Koopman, G. H. Verrips, S. P. Verloove-Vanhorick, 和 J. M. Wit. 2003. Changes can be studied when the measurement instrument is different at different time points. *Health Services and Outcomes Research Methodology*, 4, .
- Tucker, L. R. 1960. Intra-individual and inter-individual multidimensionality. 输入: *Psychological Scaling: Theory & Applications*, H. Gulliksen, 和 S. Messick, eds. New York: John Wiley and Sons.
- Vlek, C., 和 P. J. Stallen. 1981. Judging risks and benefits in the small and in the large. *Organizational Behavior and Human Performance*, 28, .

Wagenaar, W. A. 1988. Paradoxes of gambling behaviour. London: Lawrence Erlbaum Associates, Inc.

Young, F. W., Y. Takane, 和 J. De Leeuw. 1978. The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika*, 43, .

Zeijl, E., Y. te Poel, M. du Bois-Reymond, J. Ravesloot, 和 J. J. Meulman. 2000. The role of parents and peers in the leisure activities of young adolescents. *Journal of Leisure Research*, 32, .

# 非线性典型相关性分析

非线性典型相关性分析的目的是确定两个或更多变量集彼此之间的相似性。同线性典型相关性分析一样，其目的也是在低维空间中尽可能多地解释变量集间关系中的方差。然而，与线性典型相关性分析不同的是，非线性典型相关性分析不假定度量的区间度，也不假定关系为线性。另一个重要的区别是非线性典型相关性分析通过同时将每个集中变量的线性组合与未知集一对象得分相比较，建立集之间的相似性。

## 示例：调查结果的分析

本章中的示例取自一个调查 (Verdegaal, 1985)。该调查记录了 15 个主体对 8 个变量的响应。下表显示了数据集中的变量、变量标签和值标签（类别）。

表 11-1  
调查数据

变量名	变量标签	值标签
age	Age in years	20 - 25, 26 - 30, 31 - 35, 36 - 40, 41 - 45, 46 - 50, 51 - 55, 56 - 60, 61 - 65, 66 - 70
marital	Marital status	Single, Married, Other
pet	Pets owned	No, Cat(s), Dog(s), Other than cat or dog, Various domestic animals
news	Newspaper read most often	None, Telegraaf, Volkskrant, NRC, Other
music	Music preferred	Classical, New wave, Popular, Variety, Don' t like music
live	Neighborhood preference	Town, Village, Countryside
math	Math test score	0 - 5, 6 - 10, 11 - 15
语言	Language test score	0 - 5, 6 - 10, 11 - 15, 16 - 20

您可以在 `verd1985.sav` 中找到此数据集。有关详细信息，请参阅附录 A 中的样本文件中的 [IBM SPSS Categories 21](#)。感兴趣的变量是前六个变量；这六个变量被分割到三个数据集中。数据集 1 包含 年龄 和 婚姻；数据集 2 包含 宠物 和 新闻；数据集 3 包含 音乐 和 居住。宠物 被尺度化为多名义；年龄 被尺度化为有序；所有其他变量都被尺度化为单名义。该分析要求一个随机初始配置。默认情况下，初始配置是数值。但是，当某些变量被视为无法排列顺序的单名义时，最好选择随机初始配置。本项研究中的大多数变量都是这种情况。

## 检查数据

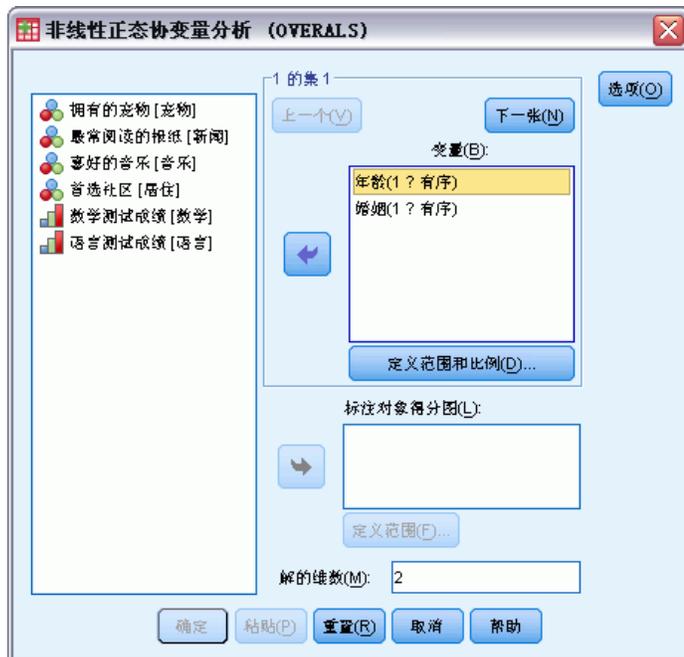
- ▶ 若要获取该数据集的非线性典型相关性分析，请从菜单中选择：  
分析 > 降维 > 最优尺度...

图片 11-1  
“最优尺度”对话框



- ▶ 在“最佳度量水平”组中，选择有些变量不是无序多分类。
- ▶ 在“变量集的数目”组中，选择多个集合。
- ▶ 单击定义。

图片 11-2  
“非线性正态协变量分析”对话框



- ▶ 选择 年龄 和 婚姻状况 作为第一个集的变量。
- ▶ 选择 年龄 并单击定义范围和比例。

图片 11-3  
“定义范围和尺度”对话框



- ▶ 键入 10 作为此变量的最大值。
- ▶ 单击继续。
- ▶ 在“非线性正态协变量分析”对话框中，选择 婚姻 并单击定义范围和比例。

图片 11-4  
“定义范围和尺度”对话框



- ▶ 键入 3 作为此变量的最大值。
- ▶ 选择单标定作为度量标度。
- ▶ 单击继续。
- ▶ 在“非线性正态协变量分析”对话框中，单击下一步定义下一个变量集。

图片 11-5  
“非线性正态协变量分析”对话框



- ▶ 选择 宠物 和 新闻 作为第二个集的变量。
- ▶ 选择 宠物 并单击定义范围和比例。

图片 11-6  
“定义范围和尺度”对话框



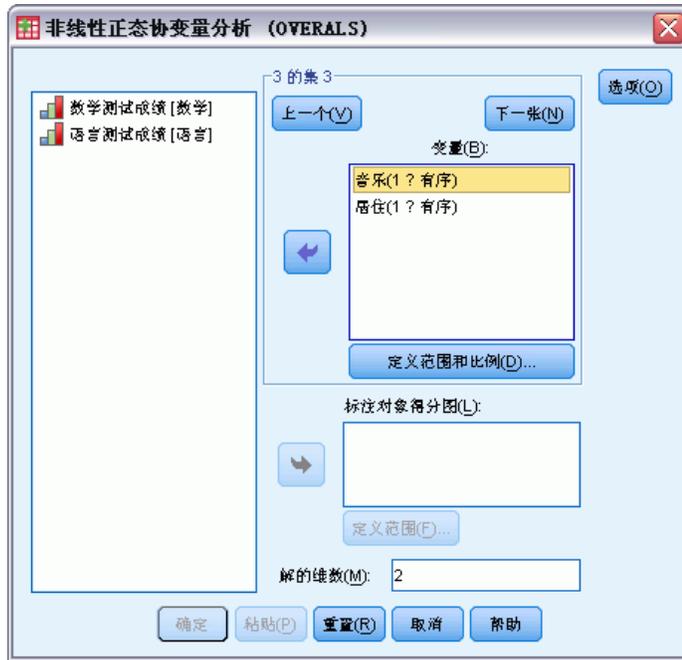
- ▶ 键入 5 作为此变量的最大值。
- ▶ 选择多标定作为度量标度。
- ▶ 单击继续。
- ▶ 在“非线性正态协变量分析”对话框中，选择 新闻 并单击定义范围和比例。

图片 11-7  
“定义范围和尺度”对话框



- ▶ 键入 5 作为此变量的最大值。
- ▶ 选择单标定作为度量标度。
- ▶ 单击继续。
- ▶ 在“非线性正态协变量分析”对话框中，单击下一步定义最后一个变量集。

图片 11-8  
“非线性正态协变量分析”对话框



- ▶ 选择 音乐 和 居住 作为第三个集的变量。
- ▶ 选择 音乐 并单击定义范围和比例。

图片 11-9  
“定义范围和尺度”对话框



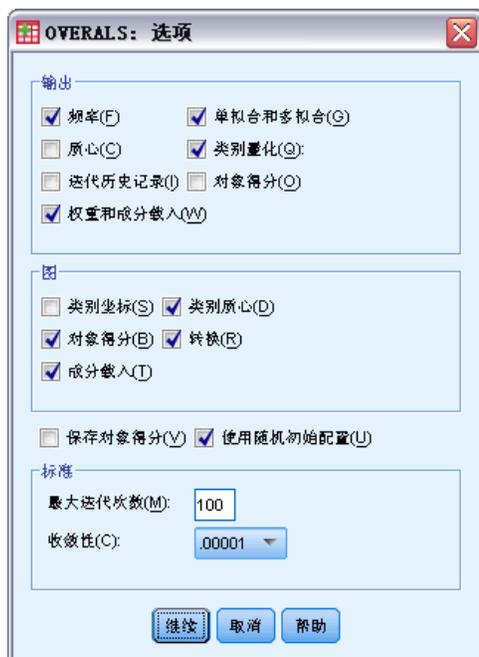
- ▶ 键入 5 作为此变量的最大值。
- ▶ 选择单标定作为度量标度。
- ▶ 单击继续。
- ▶ 在“非线性正态协变量分析”对话框中，选择 居住 并单击定义范围和比例。

图片 11-10  
“定义范围和尺度”对话框



- ▶ 键入 3 作为此变量的最大值。
- ▶ 选择单标定作为度量标度。
- ▶ 单击继续。
- ▶ 在“非线性正态协变量分析”对话框中，单击选项。

图片 11-11  
“选项”对话框



- ▶ 在“显示”组中，取消对质心的选择并选择权重和成份载荷。
- ▶ 在“图”组中，选择类别质心和转换。
- ▶ 选择使用随机初始配置。
- ▶ 单击继续。
- ▶ 在“非线性正态协变量分析”对话框中单击确定。

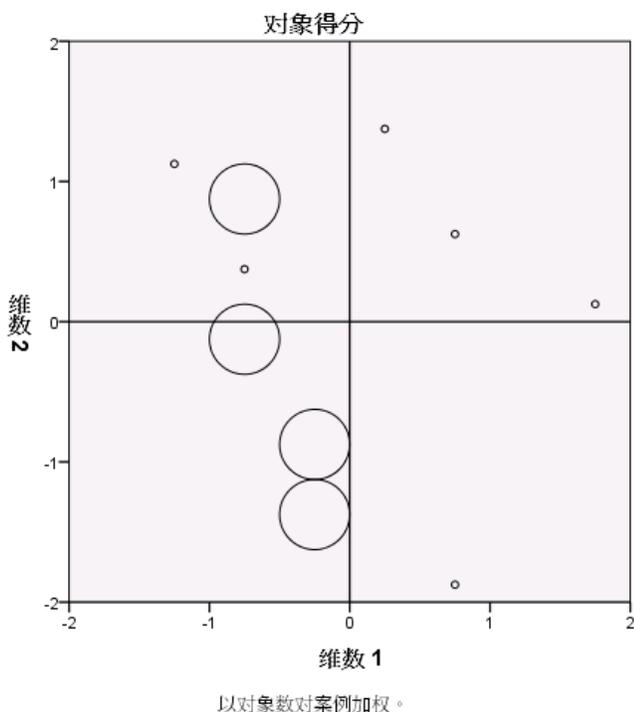
在带有最佳度量水平的变量的列表之后，采用最优尺度的分类典型相关性分析会为分析中每个变量按类别显示对象频率的表。由于几乎为空的类别更有可能在解中占优势，因此如果有缺失的数据，这些表尤为重要。在本例中没有缺失的数据。

第二个初步检查是通过查看对象得分图来寻找离群值。离群值的定量与其他对象如此不同，以致离群值出现在了图的边界，因此它们在一个或多个维度中占优势。

如果找到离群值，可以用两种方法之一来进行处理。可以简单地将其从数据中删除，然后再次运行非线性典型相关性分析。另一种方法，可以尝试通过拼并（合并）某些类别来重新编码离群对象的极端响应。

如对象得分图中所示，本调查数据没有离群值。

图片 11-12  
对象得分



## 解释集之间的相似性

非线性典型相关性分析中有多种度量集之间的关联性的方法（每种方法都在单独的表或表集中详细说明）。

### 分析摘要

拟合和损失值可以告诉您在考虑集之间的关联性的情况下，非线性典型相关性分析解和最优量化的数据之间的拟合度如何。分析摘要表显示了调查示例的拟合值、损失值和特征值。

图片 11-13  
分析摘要

		维		和
		1	2	
损失	集合 1	.240	.183	.423
	集合 2	.184	.408	.593
	集合 3	.171	.205	.376
	均值	.199	.265	.464
特征值		.801	.735	
拟合				1.536

跨维度和集划分损失。对于每个维度和集，损失表示集中变量的加权组合无法解释的对象得分变差的比例。平均损失标记为均值。在本例中，集上的平均损失是 0.464。请注意：第二维度的损失比第一维度的大。

每个维度的特征值等于 1 减去该维度的平均损失；该特征值表示每个维度所显示的关系程度。特征值构成总拟合。对于 Verdegaal 的数据，第一维度解释了  $0.801 / 1.536 = 52\%$  的实际拟合。

最大拟合值等于维数；如果达到了最大值，则表示关系是完全的。集上的平均损失值和维度可以告诉您最大拟合和实际拟合之间的差异。拟合加平均损失等于维数。很少会出现完全相似的情况，并且完全相似通常会利用数据中的普通特征。

另一个常用的包含两个变量集的统计量是典型相关性。由于典型相关性与特征值相关，它又不会提供附加信息，因此非线性典型相关性分析输出中不包含典型相关性。对于两个变量集，可以通过以下公式获取每维典型相关性：

$$\rho_d = 2 \times E_d - 1$$

其中 d 是维数，而 E 是特征值。

可以用以下公式一般化超过两个集的典型相关性：

$$\rho_d = ((K \times E_d) - 1) / (K - 1)$$

其中 d 是维数；K 是集的数量；E 是特征值。对于我们的示例来讲，

$$\rho_1 = ((3 \times 0.801) - 1) / 2 = 0.702$$

和

$$\rho_2 = ((3 \times 0.735) - 1) / 2 = 0.603$$

## 权重和成份载荷

相关性的另一个度量是来自每个集的线性组合和对象得分之间的多相关性。如果集中没有多名义的变量，该度量可以通过以下方法计算：用权重乘以集中每个变量的成份载荷、加总这些乘积并对和取平方根。

图片 11-14  
权重

集合	维	
	1	2
1 年龄	.680	.789
婚姻状况	.296	-1.016
2 最常阅读的报纸	-.845	-.361
3 喜爱的音乐	.631	-.749
首选社区	-.484	-.780

图片 11-15  
成分载入

集合	维	
	1	2
1 年龄 <sup>a, b</sup>	.834	.259
婚姻状况 <sup>c, b</sup>	.651	-.604
2 拥有的宠物 <sup>d, e</sup> 维 1	.397	-.431
2 2	-.277	.680
最常阅读的报纸 <sup>c, b</sup>	-.867	-.391
3 喜爱的音乐 <sup>c, b</sup>	.786	-.500
首选社区 <sup>c, b</sup>	-.687	-.540

- a. 最佳刻度水平: 序数
- b. 对象空间中单一量化变量的投影
- c. 最佳刻度水平: 单定类
- d. 最佳刻度水平: 多定类
- e. 对象空间中多个量化变量的投影

这些图给出了本例中变量的权重和成份载荷。如下所示，多相关性 (R) 是指最优尺度化变量 (年龄 和 婚姻状况) 的首个加权和与对象得分的第一维度之间的多相关性：

$$\begin{aligned}
 R &= \sqrt{(0.701 \times 0.841 + (-0.273 \times -0.631))} \\
 &= \sqrt{(0.5895 + 0.1723)} \\
 &= 0.873
 \end{aligned}$$

对于每个维度， $1 - \text{损失} = R^2$ 。例如，从分析摘要表中， $1 - 0.238 = 0.762$ ，即，0.873 的平方 (加上特定误差的四舍五入结果)。因此，较小的损失值表示最优尺度化变量的加权和与维度之间具有较强的多相关性。对于多名义变量，权重不是唯一的。对于多名义变量，请使用  $1 - \text{每集损失}$ 。

## 划分拟合和损失

非线性典型相关性分析通过多种方法划分每集的损失。拟合表提供了由调查示例的非线性典型相关性分析生成的多拟合、单拟合和单一损失表。请注意：多拟合减去单拟合等于单一损失。

图片 11-16  
划分拟合和损失

集合	多拟合度			单一拟合度			单一损失		
	维		和	维		和	维		和
	1	2		1	2		1	2	
1 年龄 <sup>a</sup>	.494	.676	1.170	.462	.622	1.085	.032	.054	.065
婚姻状况 <sup>b</sup>	.089	1.033	1.122	.088	1.033	1.120	.001	.000	.001
2 拥有的宠物 <sup>c</sup>	.402	.439	.841						
最常阅读的报纸 <sup>b</sup>	.724	.187	.911	.714	.130	.844	.010	.057	.067
3 喜爱的音乐 <sup>b</sup>	.421	.577	.998	.398	.561	.960	.022	.016	.039
首选社区 <sup>b</sup>	.234	.609	.843	.234	.608	.843	.000	.000	.000

a. 最佳刻度水平: 序数

b. 最佳刻度水平: 单定类

c. 最佳刻度水平: 多定类

单一损失指将变量约束为一个定量集（即，单名义、有序或名义）所得到的损失。如果单一损失较大，则最好将这些变量视为多名义。但是在本例中，单拟合和多拟合几乎相等，这意味着多个坐标在由权重给定的方向上几乎位于一条直线上。

多拟合等于每个变量的多个类别坐标的方差。这些度量类似于在齐性分析中找到的区分测量。可以检查多拟合表，以查看哪些变量区分最好。例如，检查 婚姻状况 和 最常阅读的报纸 的多拟合表。跨两个维度计算拟合值的和：婚姻状况 的拟合值是 1.122，而 最常阅读的报纸 的拟合值是 0.911。这个信息告诉我们一个人的婚姻状况比他最常阅读的报纸提供了更大的区分效能。

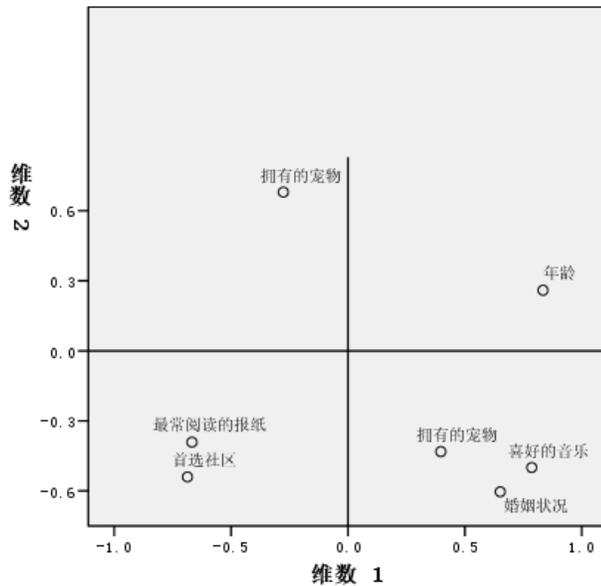
单拟合与每个变量的平方权重相对应，并等于单个类别坐标的方差。因此，权重等于单个类别坐标的标准差。通过检查跨维度分解单拟合的方式，可以看到变量 最常阅读的报纸 主要在第一维度上区分；还可以看到变量 婚姻状况 在第二维度上几乎完全区分。换言之，最常阅读的报纸 的类别在第一维度中比第二维度分散，而 婚姻状况 的情况与此模式相反。相比之下，年龄 在第一和第二维度中都有区分；因此，该变量的类别沿两个维度的分布相同。

## 成份载荷

下图显示调查数据的成份载荷图。当没有缺失数据时，成份载荷等同于已量化的变量和对象得分之间的 Pearson 相关性。

从原点到每个变量点的距离近似于该变量的重要性。图上不会显示典型变量，但是可以用穿越原点的水平和垂直线表示典型变量。

图片 11-17  
成分载入



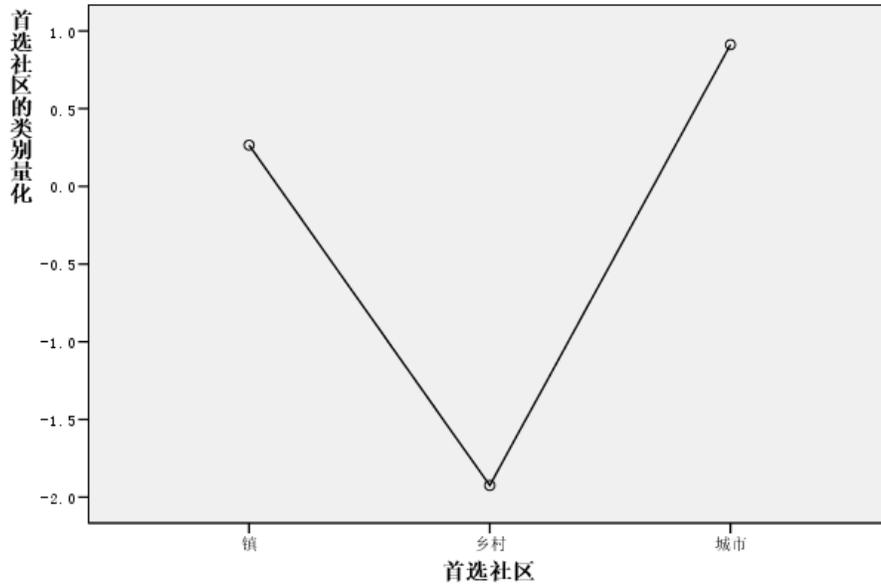
变量之间的关系是显而易见的。有两个不与水平和垂直轴重合的方向。一个方向由 年龄、最常阅读的报纸 和 首选社区 确定。另一个方向由变量 婚姻状况、喜好的音乐 和 拥有的宠物 定义。拥有的宠物 变量是多名义变量，因此图上会为该变量显示两个点。将每个定量都视为单个变量。

## 转换图

每个变量的不同尺度级别对定量施加了约束。变换图说明了从选定的最佳度量水平得出的量化和原始类别之间的关系。

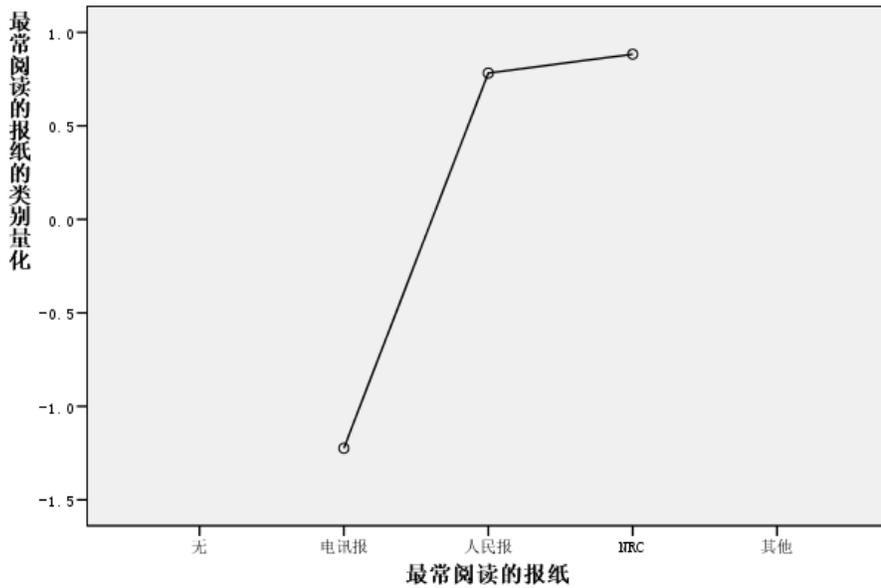
首选社区的变换图（被视为标定）显示了一个 U 型的模式，在该模式中，中间的类别取得最低的定量，而两端的类别取得的值彼此相似。这种模式表示初始变量和转换后的变量之间具有二次关系。不建议对 首选社区 使用其他最佳度量水平。

图片 11-18  
 首选社区的变换图（名义）

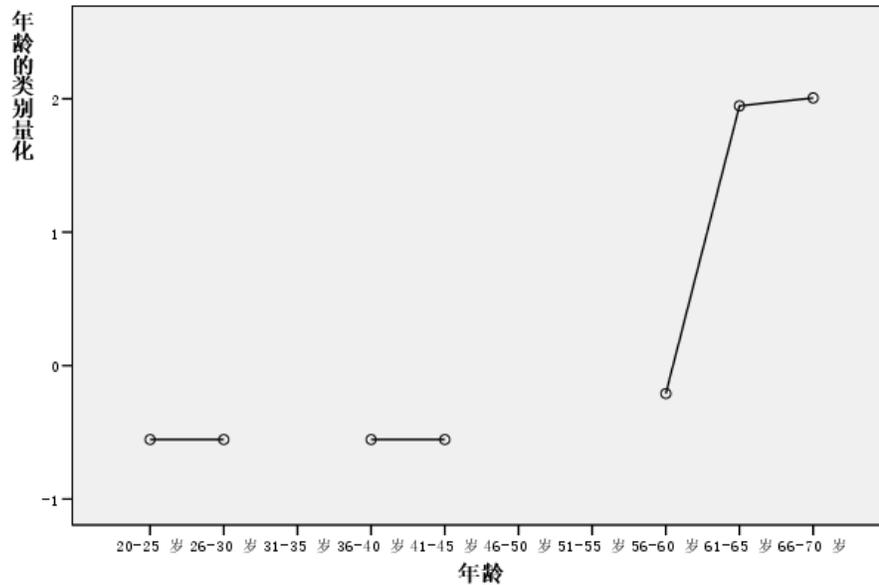


相比之下，最常阅读的报纸 的定量与具有观察个案的三个类别的递增趋势相对应。第一个类别取得的定量最小；第二个类别取得的值稍大；第三个类别取得的值最大。尽管变量被尺度化为名义变量，但是在定量中可以获得类别顺序。

图片 11-19  
 最常阅读的报纸 的变换图（名义）



图片 11-20  
年龄 的变换图（有序）



年龄 的变换图显示一个 S 型曲线。四个年龄最小的观察类别都取得了相同的负的定量，而两个年龄最大的类别取得了相似的正值。因此，可以尝试将所有年龄较小的类别拼并为一个公共类别（即，50 以下），并将两个年龄最大的类别拼并为一个类别。但是，年龄较小的组的定量的精确等同性表明：将定量的顺序约束为初始类别的顺序可能并不合理。因为 26 - 30、36 - 40 和 41 - 45 组的定量不能比 20 - 25 组的定量低，所以将这些值设为与边界值相等。允许这些值小于最年轻的组的定量（即，将年龄视为名义变量）可以提高拟合度。所以，尽管可以将年龄视为有序变量，但是在本个案中将其视为有序变量并不妥当。另外，将年龄视为数值，因而保持类别之间的距离会明显地降低拟合度。

## 单一类别与多类别坐标

对于每个被视为单名义、有序或数值的变量，确定定量、单一类别坐标和多类别坐标。显示 年龄 的这些统计量。

图片 11-21  
年龄 的坐标

	边际频率	量化	单一类别坐标		多类别坐标	
			维		维	
			1	2	1	2
20-25 岁	3	-.554	-.377	-.437	-.192	-.139
26-30 岁	5	-.554	-.377	-.437	-.404	-.623
31-35 岁	0	.000				
36-40 岁	1	-.554	-.377	-.437	-.318	-.733
41-45 岁	1	-.554	-.377	-.437	-.356	-.534
46-50 岁	0	.000				
51-55 岁	0	.000				
56-60 岁	2	-.209	-.142	-.165	-.435	.087
61-65 岁	1	1.947	1.324	1.536	1.710	1.204
66-70 岁	2	2.006	1.364	1.583	1.215	1.711
缺失	0					

每个没有记录个案的类别都取得定量 0。对于年龄，这包含类别 31 - 35、46 - 50 和 51 - 55。这些类别没有约束为与其他类别一同进行排序，并且这些类别不影响任何计算。

对于多名义变量，每个类别都会在各个维度上取得不同的定量。对于所有其他变换类型，不论解的维数是多少，一个类别只能有一个定量。单一类别坐标的每个集都表示对象空间中类别在线上的位置。给定类别的坐标等于定量和变量维度权重的乘积。例如，在年龄的表中，类别 56-60 的单一类别坐标 (-0.142, -0.165) 等于定量 (-0.209) 乘以维度权重 (0.680, 0.789)。

视为单名义、有序或数值的变量的多类别坐标表示应用有序或线性约束之前对象空间中的类别的坐标。这些值是损失的非约束最小化值。对于多名义变量，这些坐标表示类别的定量。

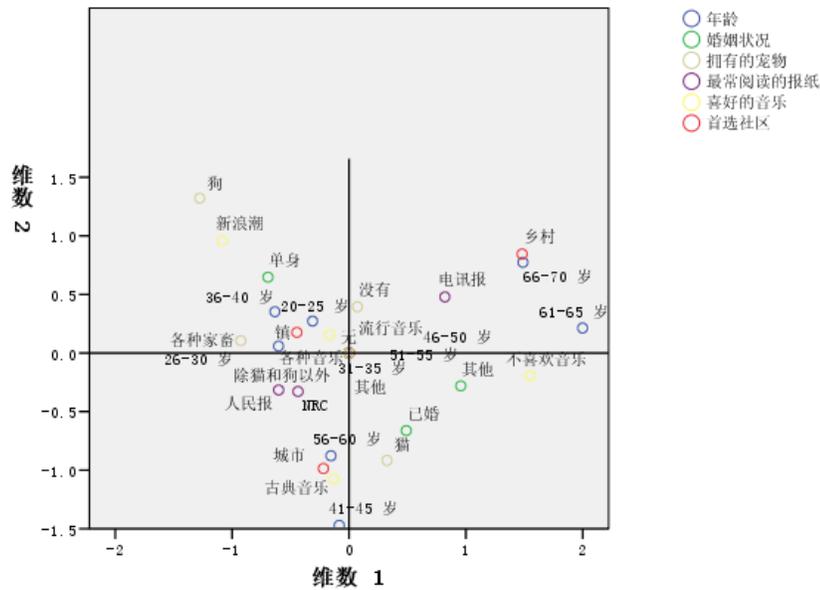
通过比较单一类别坐标和多类别坐标，可以揭示对类别和其定量的关系施加约束的效果。在第一维度上，年龄的多类别坐标减小到类别 2，并相对保持在同一级别上，直到类别 9。在类别 9 的点，坐标急剧上升。经过验证，第二维度上的模式与此相似。在单一类别坐标中会移去这些关系，并应用有序约束。现在在两个维度中，坐标都不降低。两个坐标集的不同结构表明标定处理可能会更加合适。

## 质心和投影质心

由变量标记的质心图的解释方式应与齐性分析中的类别量化图或非线性主要分量分析中的多类别坐标相同。就其本身而言，这样的图显示变量区分对象组的情况（质心位于对象的重心）。

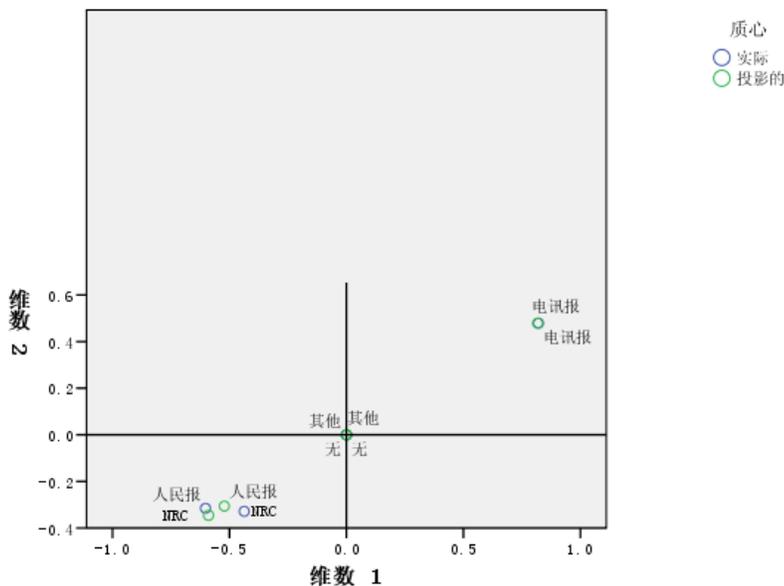
请注意：年龄的类别区分得不是很清晰。较小年龄的类别组成一组，位于图的左边。如之前所建议的，有序可能是一个过于严格的尺度级别，不能施加于年龄。

图片 11-22  
由变量标记的质心



在请求质心图时，还会生成由值标签标记的每个变量的单个质心和投影质心图。投影质心位于对象空间中的一条线上。

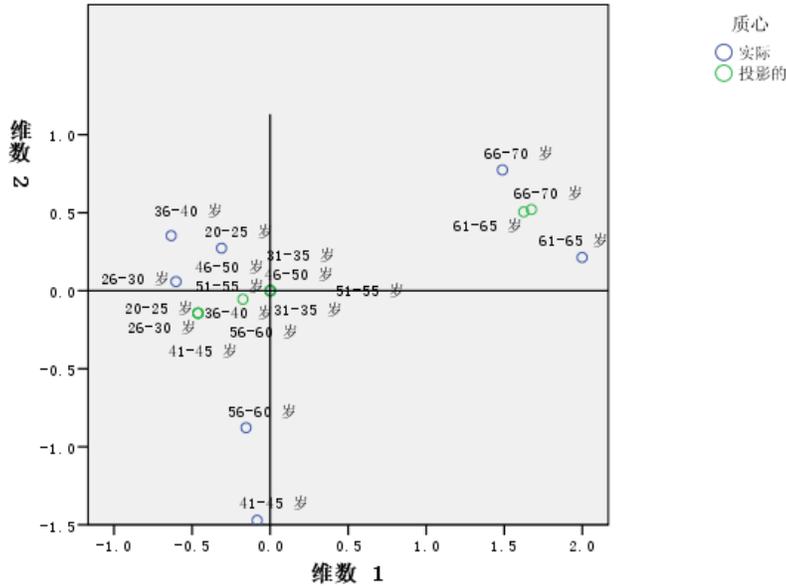
图片 11-23  
最常阅读的报纸 的质心和投影质心



实际的质心被投影到由成份载荷定义的矢量上。将这些矢量添加到质心图是为了帮助区分投影质心和实际质心。延伸两个垂直参考线，使其穿过原点，这样会形成四个象限，投影质心便落到其中的一个象限。从投影质心的位置可以获取对单名义、有序或

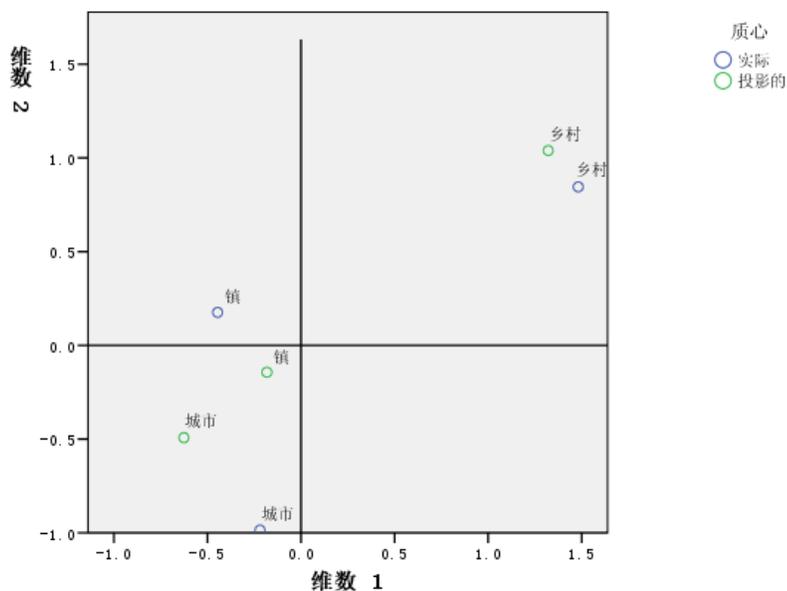
数值变量的方向的解释。例如，变量 最常阅读的报纸 被指定为单名义。投影质心显示 Volkskrant 和 NRC 与 Telegraaf 之间成鲜明对比。

图片 11-24  
年龄的质心和投影质心



从投影质心来看，年龄的问题较为明显。将年龄视为有序意味着必须保持年龄组的排列顺序。若要满足此约束，45 岁以下的的所有年龄组都会投影到同一点上。沿着由年龄、最常阅读的报纸和首选社区定义的方向，较年轻的年龄组没有区分。这个发现意味着将变量视为了名义变量。

图片 11-25  
首选社区的质心和投影质心



若要理解变量之间的关系，需先明确对于质心图中的类别的聚类都有哪些特定的类别（值）。通过查看图的右上角和左下角，可以描述 年龄、最常阅读的报纸 和 首选社区 之间的关系。在右上角，年龄组是年纪较长的响应者；他们阅读 Telegraaf 并喜欢居住在村庄。检查每个图的左下角，可以看到较为年轻的响应者到中年的响应者阅读 Volkskrant 或 NRC，并且希望居住在乡村或城镇。但是，要区分较年轻的组是非常不容易的。

对于其他方向（喜好的音乐、婚姻状况 和 拥有的宠物），通过将重点集中在质心图的左上角和右下角，可以进行同样的解释。在左上角，可以发现单身人士倾向于养狗，并喜欢新潮音乐。已婚人士和其他婚姻类别的人士则养猫；前者更喜欢古典音乐，而后者不喜欢音乐。

## 备用分析

分析结果表明将 年龄 视为有序不太恰当。尽管 年龄 用有序级别来度量，但它与其他变量之间的关系不是单调的。若要探究将最佳度量水平更改为单名义的效应，可以重新运行分析。

### 若要运行分析，请执行以下操作：

- ▶ 调用“非线性正态协变量分析”对话框并导航到第一个集。
- ▶ 选择 年龄 并单击定义范围和比例。
- ▶ 在”定义全距和刻度”对话框中，选择单标定作为度量标度。
- ▶ 单击继续。

- 在“非线性正态协变量分析”对话框中单击**确定**。

二维解的特征值分别是 0.806 和 0.757，而总拟合度为 1.564。

图片 11-26  
二维解的特征值

		维		和
		1	2	
损失	集合 1	.249	.115	.363
	集合 2	.176	.408	.584
	集合 3	.157	.205	.363
	均值	.194	.243	.436
特征值		.806	.757	
拟合				1.564

多拟合和单拟合表显示 年龄 仍然是高度区分的变量，如同多拟合值的和所表明的一样。但是，与之前的结果相对照的是，对单拟合所进行的检查揭示沿第二维度的区分几乎是完全的。

图片 11-27  
划分拟合和损失

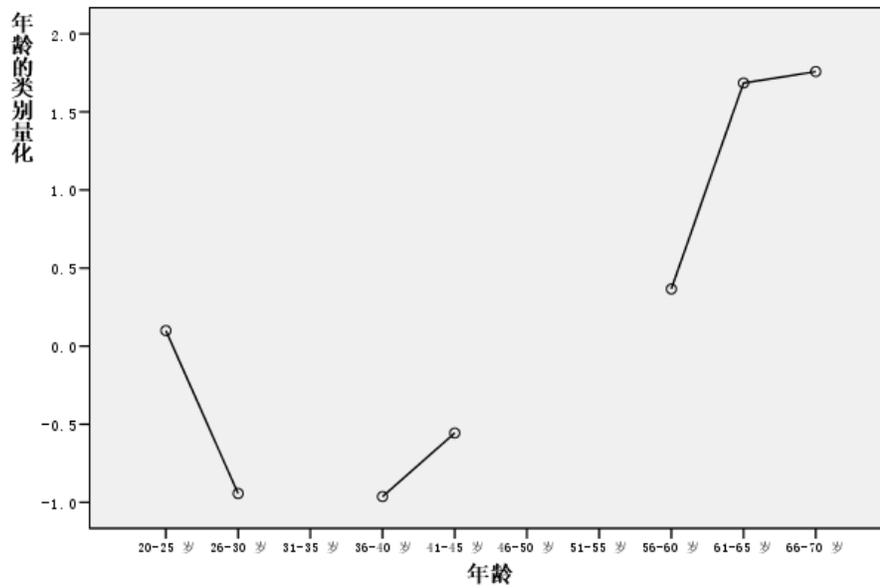
集合	多拟合度			单一拟合度			单一损耗		
	维		和	维		和	维		和
	1	2		1	2		1	2	
1 年龄 <sup>a</sup>	.246	1.197	1.443	.195	1.188	1.384	.051	.008	.059
婚姻状况 <sup>a</sup>	.273	1.136	1.409	.272	1.135	1.407	.001	.000	.002
2 拥有的宠物 <sup>b</sup>	.530	.392	.921						
最常阅读的报纸 <sup>a</sup>	.639	.185	.824	.631	.149	.780	.008	.036	.044
3 喜好的音乐 <sup>a</sup>	.604	.438	1.041	.603	.437	1.040	.000	.001	.001
首选社区 <sup>a</sup>	.075	.822	.897	.075	.822	.897	.000	.000	.000

a. 最佳刻度水平：单定类

b. 最佳刻度水平：多定类

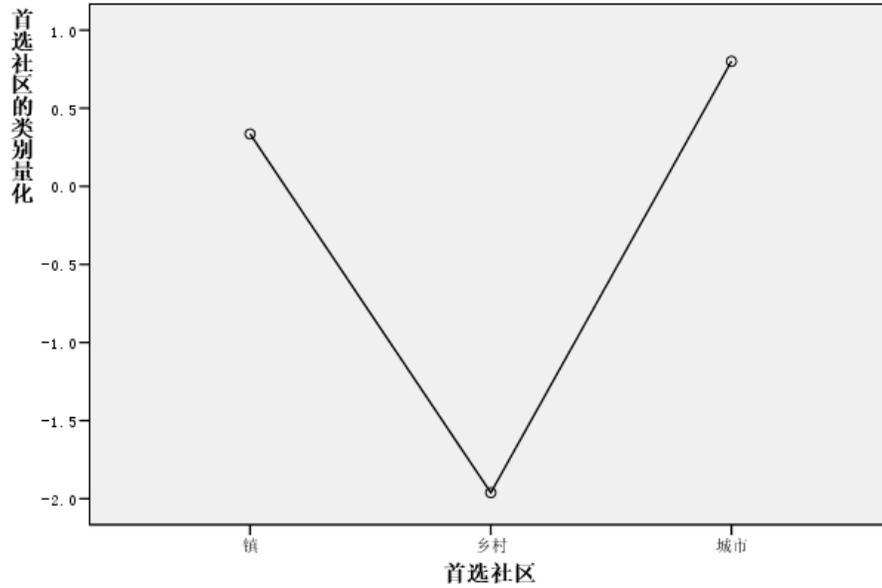
转到 年龄 变换图。名义变量的定量是非约束的，因此当 年龄 被视为有序时，曾显示的非减少趋势将不再出现。在 40 岁的年龄之前会有一个递减的趋势，之后便会出现增加的趋势；这与 U 型关系相对应。两个年龄较长的类别仍然取得相似的得分；后续的分析会涉及到组合这些类别。

图片 11-28  
年龄 的变换图（名义）



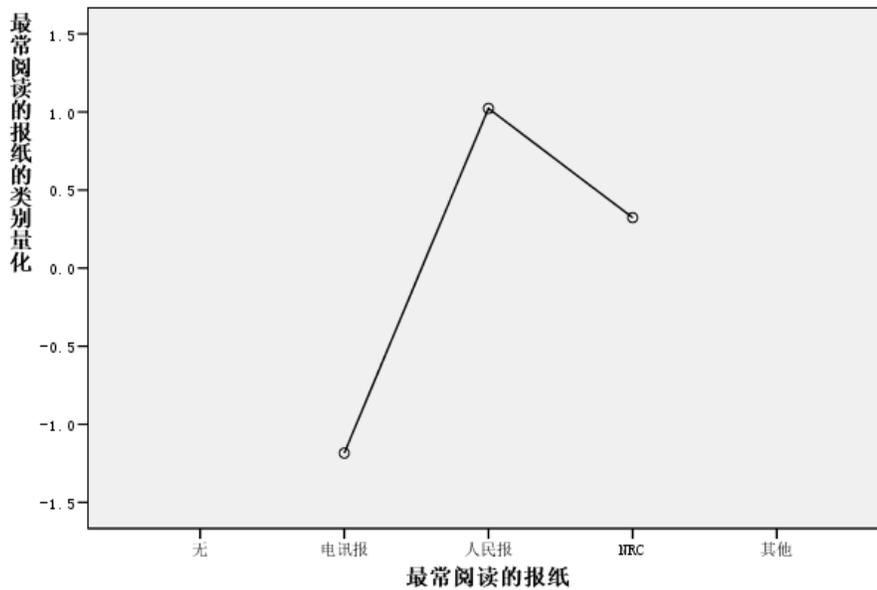
这里显示 首选社区 的变换图。将 年龄 视为标定不会对 首选社区 的定量有任何显著的影响。中间类别取得的定量最小，而两端的类别则取得较大的正值。

图片 11-29  
首选社区的变换图（名义年龄）



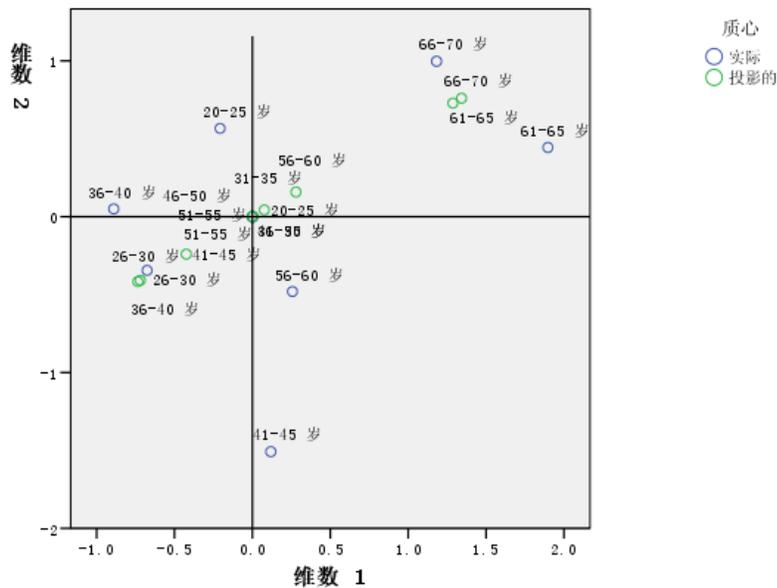
在 最常阅读的报纸 的变换图中发现了一个变化。之前，定量表现为增长的趋势，这可能意味着将该变量视为了有序。但是，将 年龄 视为标定则将这一趋势从新闻定量中去除。

图片 11-30  
最常阅读的报纸 的变换图（名义年龄）



该图是 年龄 的质心图。请注意：类别并不沿着连接投影质心的线按时间顺序分布。20 - 25 岁的组位于中间而不是末端。在之前所显示的相应有序上，类别的分布已得到了较大的改进。

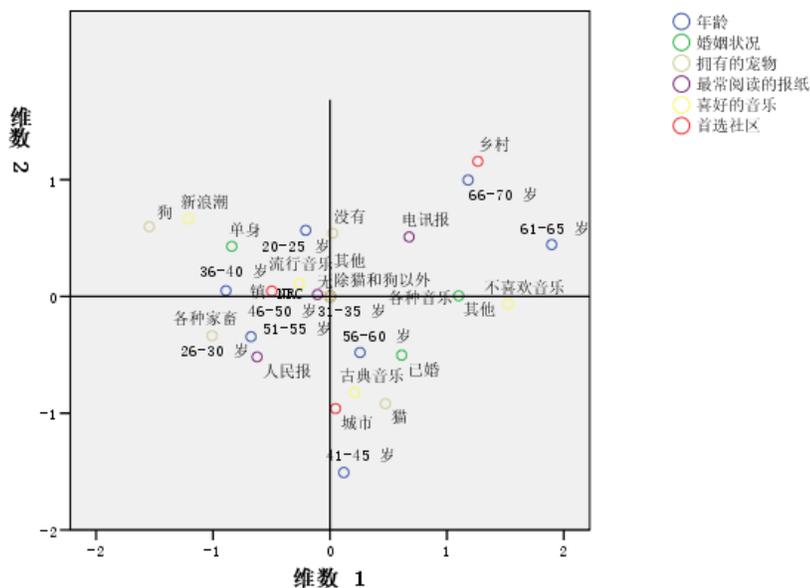
图片 11-31  
年龄 的质心和投影质心（名义）



现在，可以从质心图得到较年轻的组的解释。人民报和 NRC 类别之间的距离比在先前的分析中更远，因此可以单独解释每个类别。26 和 45 岁之间的年龄组阅读人民报并更喜欢乡村生活。20 - 25 和 56 - 60 年龄组阅读 NRC；前者更喜欢居住在城镇，而后者更喜欢乡村生活。年龄最大的组阅读电讯报并更喜欢乡村生活。

对其他方向（喜爱的音乐、婚姻状况 和 拥有的宠物）的解释基本上与先前的分析相比没有变化。唯一明显的区别在于婚姻状态为 Other 的人士或者养猫或者没有宠物。

图片 11-32  
由变量标记的质心（名义年龄）



## 一般建议

在检查了初始结果之后，您可能希望通过更改非线性典型相关性分析的某些规范来改进分析。以下是有关使分析结构化的一些提示：

- 创建尽可能多的集。将要预测的重要变量独自放入单独的集中。
- 将您认为是预测变量的变量一起放入单独的集中。如果有许多预测变量，则尝试将它们划分成若干个集。
- 将每个多名义变量都独自放入单独的集中。
- 如果变量之间高度相关，并且您不想让这种关系在解中占优势，则将这些变量一起放入同一集中。

## 推荐参考

有关非线性典型相关性分析的详细信息，请参见以下内容：

Carroll, J. D. 1968. Generalization of canonical correlation analysis to three or more sets of variables. 输入：Proceedings of the 76th Annual Convention of the American Psychological Association, 3, Washington, D.C.: American Psychological Association.

- De Leeuw, J. 1984. Canonical analysis of categorical data, 2nd ed. Leiden: DSWO Press.
- Horst, P. 1961. Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology*, 17, .
- Horst, P. 1961. Relations among m sets of measures. *Psychometrika*, 26, .
- Kettenring, J. R. 1971. Canonical analysis of several sets of variables. *Biometrika*, 58, .
- Van der Burg, E. 1988. Nonlinear canonical correlation and some related techniques. Leiden: DSWO Press.
- Van der Burg, E., 和 J. De Leeuw. 1983. Nonlinear canonical correlation. *British Journal of Mathematical and Statistical Psychology*, 36, .
- Van der Burg, E., J. De Leeuw, 和 R. Verdegaal. 1988. Homogeneity analysis with k sets of variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 53, .
- Verboon, P., 和 I. A. Van der Lans. 1994. Robust canonical discriminant analysis. *Psychometrika*, 59, .

# 对应分析

单元格包含行和列之间的对应的某种度量的任何双向表都是**对应表**。对应的测量可以是行和列变量之间的相似性、仿射性、混乱、关联或交互的任何指标。交叉制表是对应表的一个非常常见的类型，在交叉制表中单元格包含频率计数。

通过交叉表过程可以很容易地得到交叉制表。不过，交叉制表并不总是可以清晰地描述两个变量之间的关系性质。如果感兴趣的变量是名义变量（没有固有的阶或秩）并包含很多类别，则情况尤为如此。交叉制表可能会告诉您观察到的单元格频率与职业和早餐食品的  $10 \times 9$  交叉表中的期望值有显著差异，但是可能较难了解哪些职业群体具有相似的口味或都有哪些口味。

对应分析使您可以在多维尺度空间中用图形方式检查两个名义变量之间的关系。它计算行和列得分并根据得分来生成图。彼此相似的类别在图中相互之间的距离会比较接近。这样，可以很容易地分辨一个变量的哪些类别彼此相似或两个变量的哪些类别是相关的。对应分析过程还使您可以将补充点拟合到由活动点定义的空间。

如果根据类别的得分所排列的类别顺序不理想或与直觉相悖，则可以通过将某些类别的得分约束为相等来施加顺序限制。例如，假设您期望变量 smoking behavior（有 none、light、medium 和 heavy 四个类别）具有与该顺序排列一致的得分。但是，如果分析将类别的顺序排列为 none、light、heavy 和 medium，则将 heavy 和 medium 的得分约束为相等会使类别的顺序仍以得分排列。

依据距离对对应分析进行解释要依赖于所使用的标准化方法。对应分析过程既可用于分析一个变量的不同类别之间的差异，又可用于分析两个变量之间的差异。在默认的标准化方法下，它分析行和列变量之间的差异。

对应分析算法可以进行很多种分析。标准的对应分析是将行和列居中并使用卡方距离。不过，使用与欧氏距离组合的其他居中选项允许在低维空间中使用其他的矩阵表示法。

下面将给出三个示例。第一个使用相对较小的对应表并说明对应分析中固有的概念。第二个示例描述一个实际的营销应用。最后一个示例在多维尺度方法中使用距离表。

## 标准化

标准化用于在行得分和列得分上分布惯量。对应分析的解的某些方面（例如，奇异值、每维惯量和贡献）在不同的标准化方法下不会改变。行和列得分以及它们的方差会受到影响。对应分析有多种分布惯量的方法。三个最常用的方法包括只在行得分上分布惯量、只在列得分上分布惯量或对称地在行得分和列得分上分布惯量。

**主要行。**在主要行标准化中，行点之间的欧氏距离与对应表的行之间的卡方距离近似。行得分是列得分的加权平均值。列得分被标准化为与 1 的质心的距离的加权平方和。由于该方法最大化行类别之间的距离，因此如果您主要关心的是查看行变量的不同类别之间的差异，则应使用主要行标准化。

**主要列。**另一方面，您可能要估算对应表的列之间的卡方距离。这种情况下，列得分应是行得分的加权平均。行得分被标准化为与 1 的质心的距离的加权平方和。该方法最大化列类别之间的距离；如果您主要关心的是列变量的不同类别之间的差异，则应使用该方法。

**对称。**还可以对称地处理行和列。该标准化方法在行和列得分上平均地分布惯量。请注意，在这种情况下，行点之间的距离和列点之间的距离都不是卡方距离的近似值。如果您主要关心的是两个变量之间的差异或相似性，则请使用该方法。通常，这是生成双标图的首选方法。

**主体。**第四个选项称为主标准化，在主标准化中惯量在解中分布两次——一次在行得分上，一次在列得分上。如果您分别关注行点之间的距离和列点之间的距离，而不是行点和列点的相互关系，则应使用该方法。双标图不适用于本标准化选项，因此如果您已经指定了主标准化方法，则双标图不可用。

## 示例：咖啡品牌的认知度

上个示例中用到一个假设数据的小型表。而实际应用中用到的表通常要大得多。在本例中，将使用与六种冰咖啡的认知品牌形象相关的数据（Kennedy, Riquier, 和 Sharp, 1996）。此数据集可在 coffee.sav 中找到。[有关详细信息，请参阅附录 A 中的样本文件中的 IBM SPSS Categories 21。](#)

对于 23 种冰咖啡特征属性中的每种属性，人们选择了由该属性所描述的所有品牌。为保密起见，六种品牌用 AA、BB、CC、DD、EE 和 FF 来表示。

表 12-1  
冰咖啡属性

特征属性	标签	特征属性	标签
good hangover cure	cure	fattening brand	fattening
low fat/calorie brand	low fat	appeals to men	men
brand for children	children	South Australian brand	South Australian
working class brand	working	traditional/old fashioned brand	traditional
rich/sweet brand	sweet	premium quality brand	premium
unpopular brand	unpopular	healthy brand	healthy
brand for fat/ugly people	ugly	high caffeine brand	caffeine
very fresh	fresh	new brand	new
brand for yuppies	yuppies	brand for attractive people	attractive
nutritious brand	nutritious	tough brand	tough
brand for women	women	popular brand	popular
minor brand	minor		

开始时，您把重点集中在属性之间的相关性和品牌之间的相关性上。使用主标准化在行上分布一次总惯量并在列上分布一次总惯量。尽管这样会妨碍双标图解释，但是可以检查每个变量的各类别之间的距离。

## 运行分析

- ▶ 数据设置要求用变量 频率 对个案进行加权。为此，请从菜单中选择：  
数据 > 加权个案...

图片 12-1  
“加权个案”对话框



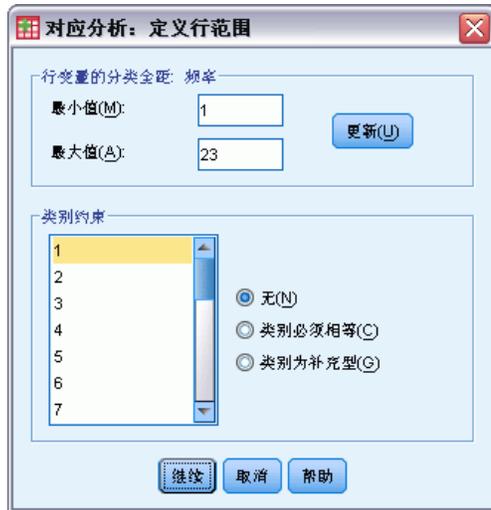
- ▶ 个案的频率变量：频率。
- ▶ 单击确定。
- ▶ 若要使用主标准化在五个维度中获取初始解，请从菜单中选择：  
分析 > 降维 > 对应分析...

图片 12-2  
“对应分析”对话框



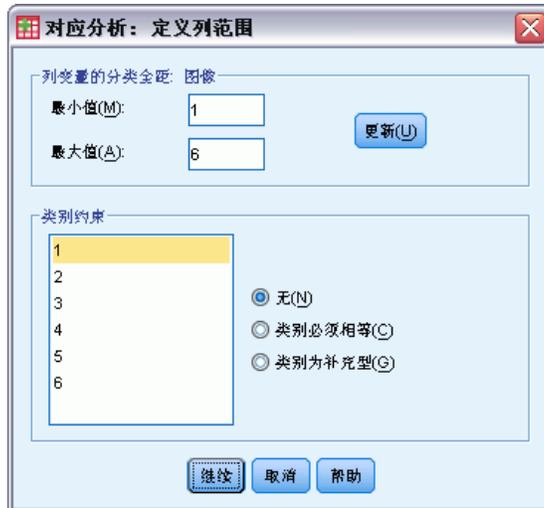
- ▶ 选择 图像 作为行变量。
- ▶ 单击定义范围。

图片 12-3  
“定义行范围”对话框



- ▶ 键入 1 作为最小值。
- ▶ 键入 23 作为最大值。
- ▶ 单击更新。
- ▶ 单击继续。
- ▶ 选择 品牌 作为列变量。
- ▶ 在“对应分析”对话框中单击定义范围。

图片 12-4  
“定义列范围”对话框



- ▶ 键入 1 作为最小值。
- ▶ 键入 6 作为最大值。

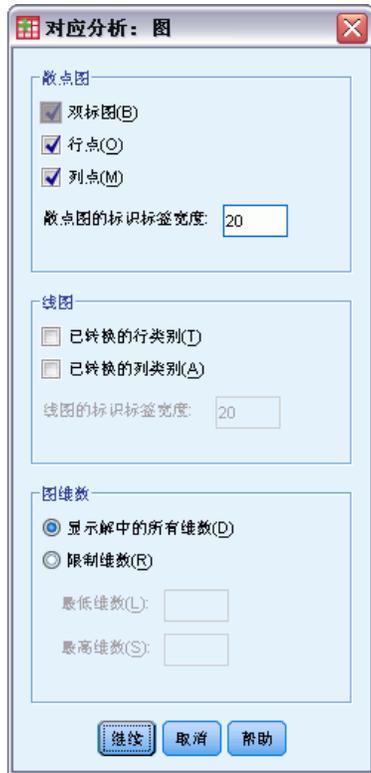
- ▶ 单击更新。
- ▶ 单击继续。
- ▶ 在“对应分析”对话框中单击模型。

图片 12-5  
“模型”对话框



- ▶ 选择主要作为正态化方法。
- ▶ 单击继续。
- ▶ 在“对应分析”对话框中单击图。

图片 12-6  
“图”对话框



- ▶ 在“散点图”组中选择行点和列点。
- ▶ 单击继续。
- ▶ 在“对应分析”对话框中单击确定。

## 维数

每维惯量显示总惯量沿每个维度的分解。两个维度占总惯量的 83%。添加第三维度只贡献了 8.6% 比例的惯量。于是您决定使用二维表示形式。

图片 12-7  
每维惯量

维	奇异值	惯量	卡方	显著性	惯量比例		置信奇异值	
					考虑情况	累积	标准差	相关性
1	.711	.506			.629	.629	.009	.132
2	.399	.159			.198	.827	.014	
3	.263	.069			.086	.913		
4	.234	.055			.068	.982		
5	.121	.015			.018	1.000		
总计		.804	3746.968	.000 <sup>a</sup>	1.000	1.000		

a. 110 自由度

## 贡献率

行点概览显示了行点对维度惯量的贡献和维度对行点惯量的贡献。如果所有点对惯量的贡献都相等，则贡献将是 0.043。健康的 和 低脂肪 都对第一维度的惯量贡献了相当大的比例。男人 和 强硬的 对第二维度的惯量的贡献量最大。丑的 和 新鲜的 对两个维度的贡献都很小。

图片 12-8  
属性贡献

图像	质量	维中的得分		惯量	贡献				
		1	2		点对维惯量		维对点惯量		总计
					1	2	1	2	
使人发胖的	.080	-.514	-.265	.033	.042	.035	.652	.173	.825
男人	.051	-.852	.825	.072	.073	.219	.512	.480	.992
南澳大利亚	.057	-.303	-.350	.046	.010	.044	.114	.152	.266
传统的	.040	-.703	-.532	.043	.039	.071	.454	.260	.715
奖金	.042	-.444	-.582	.028	.016	.090	.296	.509	.805
健康的	.053	1.200	.174	.081	.152	.010	.953	.020	.973
咖啡因	.047	-.452	.124	.014	.019	.005	.702	.053	.755
新的	.047	.960	.147	.046	.086	.006	.693	.021	.914
迷人的	.041	.657	-.056	.019	.035	.001	.911	.007	.918
强硬的	.039	-.850	1.002	.070	.056	.246	.404	.560	.964
流行的	.060	-.697	-.042	.036	.058	.001	.771	.003	.774
治疗	.026	-.389	.266	.009	.008	.011	.446	.209	.655
低脂肪	.052	1.305	.196	.094	.175	.013	.941	.021	.962
孩子	.024	-.352	-.513	.017	.006	.041	.179	.380	.559
工作的	.045	-.785	.477	.040	.055	.064	.693	.255	.948
甜的	.038	-.519	-.683	.046	.020	.112	.212	.368	.580
不受欢迎的	.024	.489	.186	.010	.011	.005	.585	.085	.670
丑的	.030	.006	-.109	.003	.000	.002	.000	.131	.131
新鲜的	.036	-.096	-.100	.002	.001	.002	.196	.214	.410
雅皮士	.034	.380	-.301	.012	.010	.019	.392	.246	.637
有营养的	.040	.722	.055	.022	.041	.001	.946	.006	.951
女人	.054	.758	-.063	.032	.062	.001	.965	.007	.972
未成年的	.040	.579	.063	.023	.027	.001	.593	.007	.600
活动总计	1.000			.804	1.000	1.000			

两个维度对大多数行点的惯量贡献都很大。第一维度对 健康的、新的、迷人的、低脂肪、有营养的 和 女人 的大量贡献表明在一个维度中可以很好地描述这些点。因此，更高的维度对这些点的惯量贡献很小，这些惯量与水平轴会非常接近。第二维度对 男人、奖金 和 强硬的 贡献最大。两个维度对 南澳大利亚 和 丑的 的惯量的贡献都很小，因此未能较好地表示这些点。

列点概览显示了涉及列点的贡献。品牌 CC 和 DD 对第一维度贡献最大，而 EE 和 FF 解释了第二维度的大量惯量。AA 和 BB 对两个维度的贡献都很小。

图片 12-9  
品牌贡献

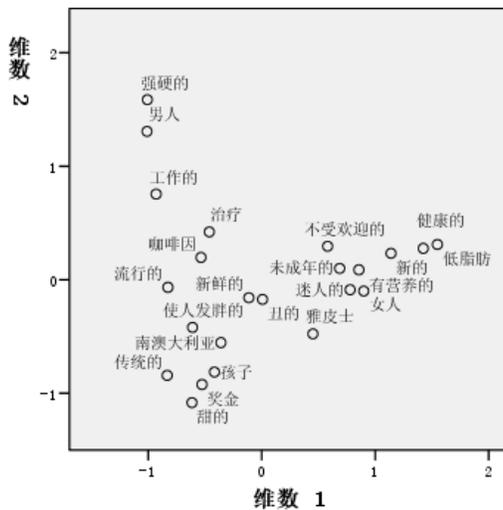
品牌	质量	维中的得分		惯量	贡献				
		1	2		点对维惯量		维对点惯量		总计
					1	2	1	2	
AA	.217	-.659	.046	.127	.187	.003	.744	.004	.748
BB	.131	-.284	-.404	.078	.021	.134	.135	.272	.407
CC	.185	.996	.076	.193	.362	.007	.951	.006	.957
DD	.162	.915	.101	.146	.267	.010	.928	.011	.939
EE	.152	-.651	.706	.153	.127	.477	.420	.494	.914
FF	.153	-.343	-.618	.107	.036	.369	.169	.550	.718
活动总计	1.000			.804	1.000	1.000			

在两个维度中，除 BB 之外所有品牌都得到了很好的表示。CC 和 DD 在一个维度中得到了很好的表示。第二维度对 EE 和 FF 的贡献最大。请注意，AA 在第一维度中得到了很好的表示，但对该维度并没有很大的贡献。

## 图

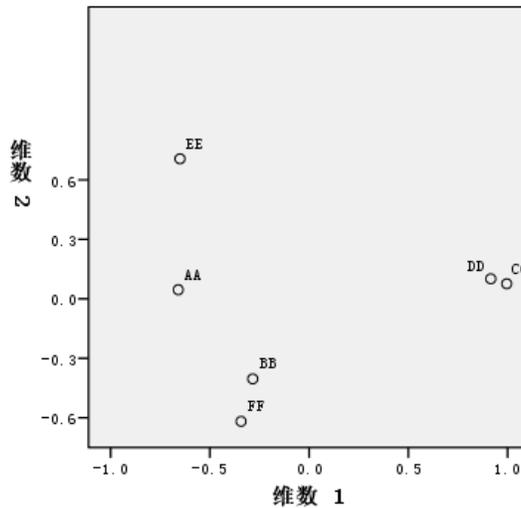
行得分图显示 新鲜的 和 丑的 距离原点都很近，这表明它们与平均行轮廓表的差异很小。出现三个一般分类。强硬的、男人 和 工作的 位于图的左上角，它们彼此之间都很相似。左下角包含 甜的、使人发胖的、孩子 和 奖金。与之相反，健康的、低脂肪、有营养的 和 新的 聚集在图的右侧。

图片 12-10  
特征属性图（主标准化）



请注意，在列点图中所有品牌都与原点距离很远，因此没有与整体质心相似的品牌。品牌 CC 和 DD 在右侧组成一组，而品牌 BB 和 FF 则聚集在图的下半部分。品牌 AA 和 EE 与任何其他品牌都不相似。

图片 12-11  
品牌的图（主标准化）



## 对称标准化

品牌与特征属性的相关性如何？主标准化不能确定这些关系。若要将重点集中在变量之间的相关性，请使用对称标准化。对称标准化不会两次分布惯量（如在主标准化中），而是在行和列上平均地划分惯量。不能解释单个变量的类别之间的距离，但是不同变量的类别之间的距离是有意义的。

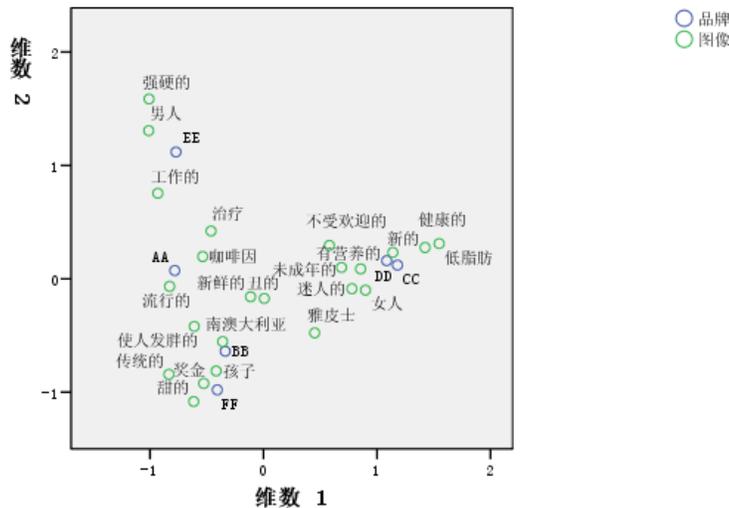
图片 12-12  
“模型”对话框



- ▶ 若要用对称标准化生成以下解，请调用“对应分析”对话框，并单击模型。
- ▶ 选择对称作为正态化方法。
- ▶ 单击继续。
- ▶ 在“对应分析”对话框中单击确定。

在生成的双标图的左上角，品牌 EE 是唯一具备 强硬的、工作的和男人属性的品牌。品牌 AA 是最受欢迎的品牌，也被认为是咖啡因含量最高的品牌。甜的和使人发胖的品牌包括 BB 和 FF。品牌 CC 和 DD 尽管被认为是新的和健康的品牌，但也是不受欢迎的品牌。

图片 12-13  
品牌和属性的双标图（对称标准化）



为进行进一步的解释，可以画一条穿越原点以及 男人 和 雅皮士 两个特征属性的线，并将品牌投影到这条线上。这两个属性是彼此相对的，这表明 男人 的品牌关联方式与 雅皮士 的方式相反。即，男人与品牌 EE 的关联最频繁，与品牌 CC 的关联最少，而 雅皮士 与品牌 CC 的关联最频繁，与 EE 的关联最少。

## 推荐参考

有关对应分析的更多信息，请参见以下内容：

Fisher, R. A. 1938. Statistical methods for research workers. Edinburgh: Oliver and Boyd.

Fisher, R. A. 1940. The precision of discriminant functions. Annals of Eugenics, 10, .

Gilula, Z., 和 S. J. Haberman. 1988. The analysis of multivariate contingency tables by restricted canonical and restricted association models. Journal of the American Statistical Association, 83, .

## 多重对应分析

多重对应分析又称齐性分析，其目的是在尽可能地使类别相互分离的意义上寻找最佳定量。这意味着相同类别中的对象在图中的位置接近，而不同类别中的对象在图中的位置尽可能地远离。术语**齐性**还指当变量为均一变量（即，变量将对象划分为具有相同或相似类别的聚类）时，分析会最为成功这一事实。

### 示例：五金件的特征

要探索多重对应分析的工作原理，将会用到 Hartigan (Hartigan, 1975) 提供的数据，这些数据可以在 screws.sav 中找到。[有关详细信息，请参阅附录 A 中的样本文件中的 IBM SPSS Categories 21。](#) 该数据集包含关于螺钉、螺栓、螺母和图钉的特性的信息。下表显示变量和变量标签，以及为 Hartigan 五金件数据集中每个变量的类型指定的值标签。

表 13-1  
Hartigan 五金件数据集

变量名	变量标签	值标签
thread	螺纹	Yes_Thread, No_Thread
head	头部形状	Flat、Cup、Cone、Round、Cylinder
indhead	头部凹槽	None、Star、Slit
bottom	底部形状	sharp、flat
length	长度（以半英寸计）	1/2_in、1_in、1_1/2_in、2_in、2_1/2_in
brass	铜制	Yes_Br、Not_Br
object	对象	tack、nail1、nail2、nail3、nail4、nail5、nail6、nail7、nail8、screw1、screw2、screw3、screw4、screw5、bolt1、bolt2、bolt3、bolt4、bolt5、bolt6、tack1、tack2、nailb、screwb

### 运行分析

- ▶ 若要获取“多重对应分析”，请从菜单中选择：  
分析 > 降维 > 最优尺度...

图片 13-1  
“最优尺度”对话框



- ▶ 确保已选择所有变量都是多名义和一个集合，然后单击定义。

图片 13-2  
“多重对应分析”对话框



- ▶ 选择 螺纹 到 长度（以半英寸计） 作为分析变量。
- ▶ 选择 对象 作为标记变量。
- ▶ 单击“图”组中的对象。

图片 13-3  
“对象图”对话框



- ▶ 选择按照变量标记对象。
- ▶ 选择 螺纹 到 对象 作为标记变量。
- ▶ 单击继续，然后在“多重对应分析”对话框的“图”组中单击变量。

图片 13-4  
“变量图”对话框



- ▶ 选择生成 螺纹 到 长度 的联合类别图。
- ▶ 单击继续。
- ▶ 在“多重对应分析”对话框中单击确定。

## 模型摘要

齐性分析可以计算多个维度的解。最大维数等于以下两者中的较小者：类别的数量减去无缺失数据的变量数，和观察值的数量减去 1。不过，应该很少会用到最大维数。维数越少就越容易解释；而到达特定的维数之后，归因于附加关联的量可以被忽略。在齐性分析中，颇为常见的解是一维、二维或三维解。

图片 13-5  
模型汇总

维	Cronbach's Alpha	方差考虑情况		
		总计(特征值)	惯量	方差的 %
1	.878	3.727	.621	62.123
2	.657	2.209	.368	36.809
总计		5.936	.989	
均值	.796 <sup>a</sup>	2.968	.495	49.466

a. 总 Cronbach's Alpha 基于平均特征值。

该解将解释数据中几乎所有的方差，其中第一维度解释 62.1%，第二维度解释 36.8%。

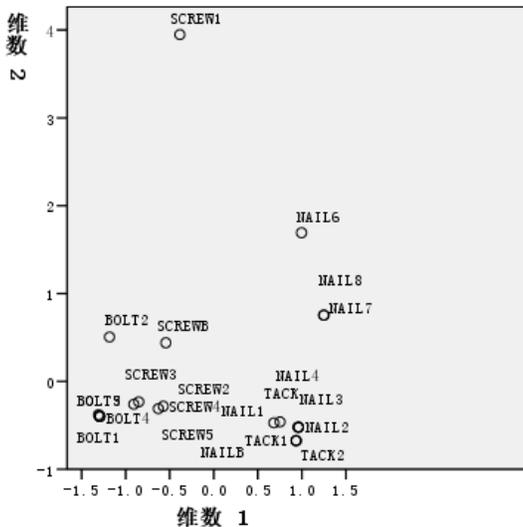
这两个维度共同以距离提供了解释。如果变量可以良好地区分，则对象会接近其所从属的类别。理想的情况下，同一类别中的对象彼此之间会较接近（即，它们应具有相似的得分），并且如果不同变量的类别属于同一对象，则它们也会较接近（即，如果两个对象对于一个变量具有相似的得分，则它们对于解中的其他变量得分也会较接近）。

## 对象得分

检查了模型摘要后，应查看对象得分。可以指定一个或多个变量以标记对象得分图。每个标记变量都会生成一个单独的图，该图以该变量的值作为标记。我们' 将查看这个由变量对象标记的对象得分图。它只是一个个案标识变量，并且不会在任何计算过程中使用。

对象到原点的距离反映了与“平均”响应模式之间的变差。平均响应模式与每个变量的最频繁类别相对应。具有许多与最频繁类别相对应的特征的对象位于距离原点较近的位置。相比之下，具有独特特征的对象位于距离原点较远的位置。

图片 13-6  
由对象标记的对象得分图



检查图，可以看到第一维度（水平轴）将螺钉和螺栓（两者有螺纹）与钉子和图钉（两者没有螺纹）区分开。由于螺钉和螺栓位于水平轴的一端，而钉子和图钉位于另一端，因此可以容易地从图中看出这一点。从更小的范围来讲，第一维度还将螺栓（它的底部是平的）与所有其他五金件（它们的底部是尖的）区分开。

第二维度（垂直轴）似乎是将 SCREW1 和 NAIL6 与所有其他五金件区分开。SCREW1 和 NAIL6 的共同之处是它们的 length 变量的值它们是本数据中最长的对象。另外，SCREW1 与原点的距离比其他对象要远的多，这表示从整体上看，这个对象的许多特征都不与其他对象共享。

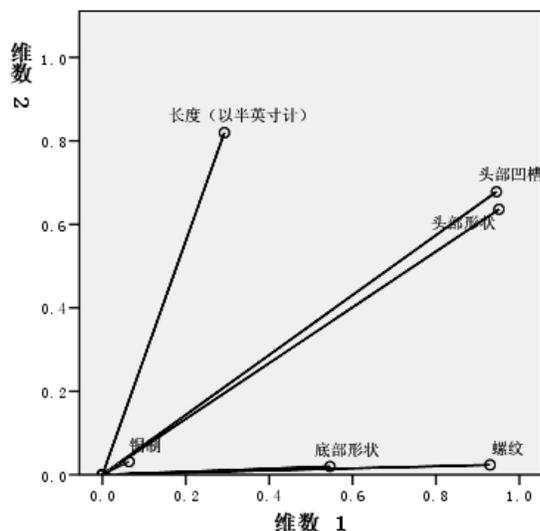
对象得分图在辨认离群值方面特别有用。可将 SCREW1 视为离群值。随后我们将考察如果丢弃该对象会发生什么状况。

## 区分测量

在检查其他对象得分图之前，让我们来查看区分测量是否与我们到目前为止所讲述的内容相符。对于每个变量，会为每个维度计算可视为平方成份载荷的区分测量。该测量也是已量化的变量在该维度中的方差。它的最大值是 1。如果对象得分落到互斥的组，并且同一类别中的所有对象得分都相同，则区分测量达到该最大值。（注意：如果有缺失数据，则该测量的值可能会大于 1。）取值大的区分测量与变量的不同类别之间分布的距离远相对应，因此，它表示变量的不同类别之间在该维度上是高度区分的。

任何维度中区分测量的平均值都等于归因于该维度的方差的百分比。因此，维度根据平均区分进行排序。第一维度的平均区分最大；第二维度的平均区分次之；对于解中的所有维度均依此类推。

图片 13-7  
区分测量的图



如同对象得分图上所表示的一样，区分测量图显示第一维度与变量 螺纹 和 底部形状 相关。这些变量在第一维度中的区分测量值较大，而在第二维度中的区分测量值较小。因此，对于这两个变量，类别均只在第一维度的分布彼此远离。长度（以半英寸计）在第二维度上的区分测量值较大，而在第一维度上则较小。因此，长度 最靠近第二维度，这与从对象得分图观察的结果（即，第二维度似乎是将最长的对象与其他对象区分

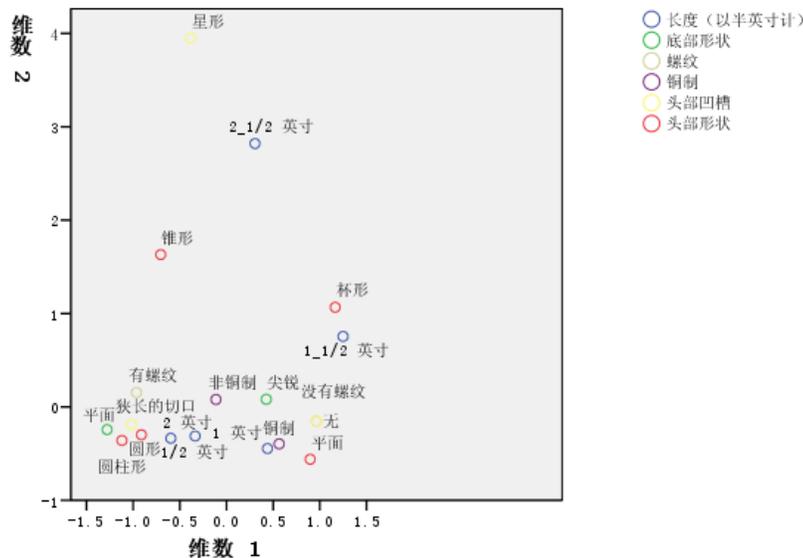
开)相一致。头部凹槽 和 头部形状 在两个维度上都有相对较大的值, 这表明它们在第一和第二维度上都处于区分状态。变量 铜制 与原点距离很近, 它在前两个维度上都没有区分。这是讲得通的, 因为所有的对象都可以用铜来制造, 也可以都不用铜来制造。

## 类别定量

请回想一下区分测量是已量化的变量沿某一特定维度的方差。区分测量图包含这些方差, 表明哪些变量沿哪一维度区分。不过, 同一个方差可以与所有分布距离适度的类别或大多数距离较近的类别相对应, 而允许少数类别与该组不同。区分图不能区别这两种情况。

类别量化图提供了显示变量区分的另一种方法, 这种方法可以识别类别关系。在该图中, 会显示每个维度上每个类别的坐标。这样便可以确定对于每个变量而言哪些类别是相似的。

图片 13-8  
类别量化



长度 (以半英寸计) 有五个类别, 其中的三个在图的顶部附近组成一组。其余两个类别位于图的下半部分, 并且类别 2\_1/2 英寸 与该组的距离非常远。长度沿维度 2 的区分较大的原因是这一类别与长度的其他类别差异很大。类似地, 对于 头部形状 而言, 类别 星形 与其他类别相距很远, 导致第二维度中区分测量较大。在区分测量图中无法演示这些模式。

变量的类别量化的分布可以反映方差, 因此可以表明变量在每个维度中的区分程度。长度的类别集中在维度 1, 而它们之间彼此远离。但是沿维度 2, 该变量的类别之间的距离却很近。因此, 长度在维度 1 中要比在维度 2 中区分得更好。相比之下, 头部形状 的类别沿两个维度的分布彼此远离, 这意味着该变量在两个维度中都区分较好。

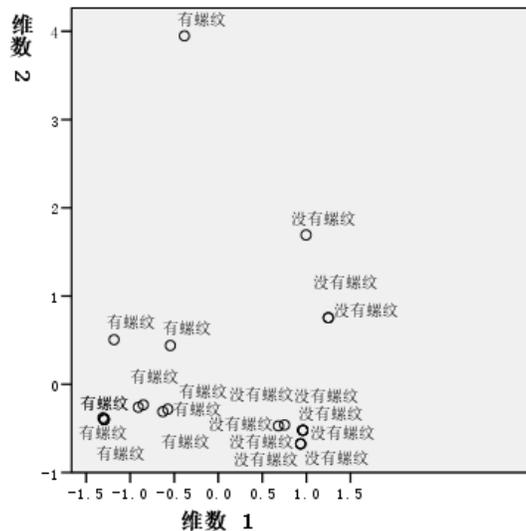
除了确定变量区分的维度和变量区分的状况之外, 类别量化图还比较变量区分。其类别彼此远离的变量要比类别彼此靠近的变量区分得更好。例如, 沿着维度 1, 铜制的两个类别彼此之间的距离要比长度的两个类别之间的距离近得多, 这表明沿着这一维度, 长度比铜制区分得好。不过, 沿着维度 2, 二者的距离很相似, 这意味着这

些变量沿着该维度的区分程度相同。之前讨论的区分测量图通过使用方差来反映类别的分布，以此来确定这些相同的关系。

## 更详细地考察对象得分

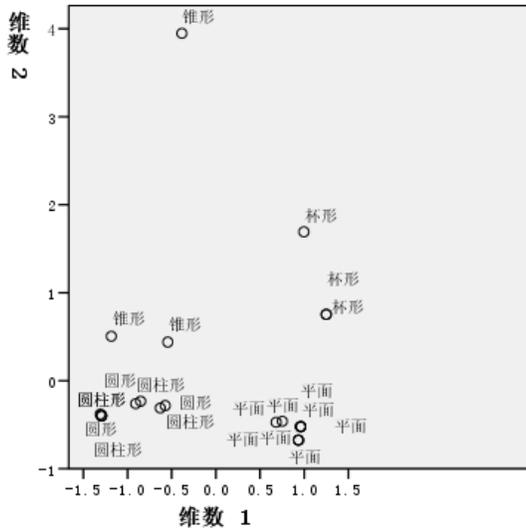
通过检查由每个变量标记的对象得分图可以更深刻地理解数据。理想的状况下，类似的对象应形成互斥的组，并且这些组应彼此远离。

图片 13-9  
用“螺纹”标记的对象得分



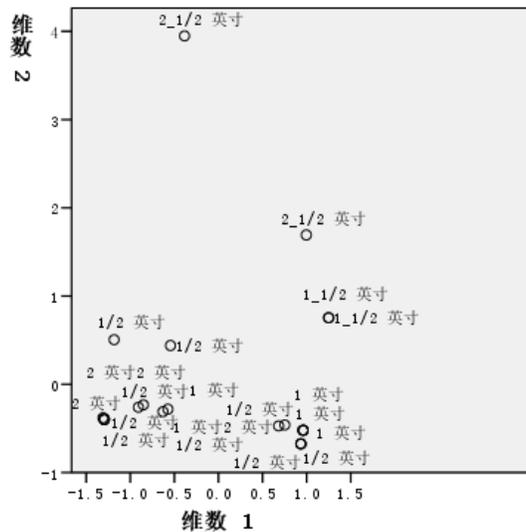
用螺纹标记的图显示第一维度将有螺纹和没有螺纹完全分隔开。所有有螺纹的对象都具有负的对象得分，而所有没有螺纹的对象都具有正的得分。尽管这两个类别没有形成紧凑的组，但是通常情况下，可将这两个类别之间的完全区分视为一个较好的结果。

图片 13-10  
用头部形状标记的对象得分



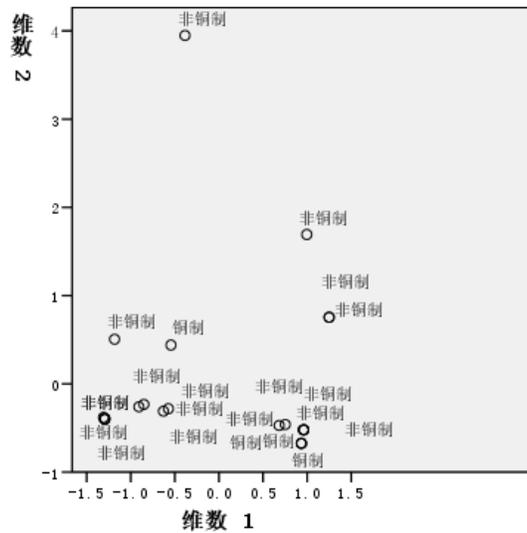
用头部形状标记的图显示该变量在两个维度中都有区分。平面对象在图的右下角组成一组，而杯形对象在右上角组成一组。所有锥形对象都位于左上角。但是，这些对象比其他组更分散，因此该组不像其他组一样是均一的。最后，圆柱形对象无法与圆形对象分隔，这两个对象都位于图的左下角。

图片 13-11  
用长度（以半英寸计）标记的对象得分



用长度（以半英寸计）标记的图显示该变量在第一维度中没有区分。在将它的类别投影到水平线上时，其类别没有显示分组。不过，长度（以半英寸计）在第二维度中确实有区分。较短的对象与正的得分相对应，而较长的对象与较高的负得分相对应。

图片 13-12  
用铜制标记的对象得分



用 铜制 标记的图显示该变量具有无法在第一或第二维度较好区分的类别。其对象得分在整个空间中分布范围较广。无法区分铜制对象与非铜制对象。

## 忽略离群值

在齐性分析中，离群值是具有过多独特特征的对象。如之前所了解到的，可以将 SCREW1 视为离群值。

若要删除该对象并再次运行分析，请从菜单中选择：  
数据 > 选择个案...

图片 13-13  
“选择个案”对话框



- ▶ 选择如果满足条件。
- ▶ 单击如果。

图片 13-14  
“如果”对话框



- ▶ 键入 `object ~= 16` 作为条件。
- ▶ 单击继续。
- ▶ 在“选择个案”对话框中单击确定。
- ▶ 最后，调用“多重对应分析”对话框，并单击确定。

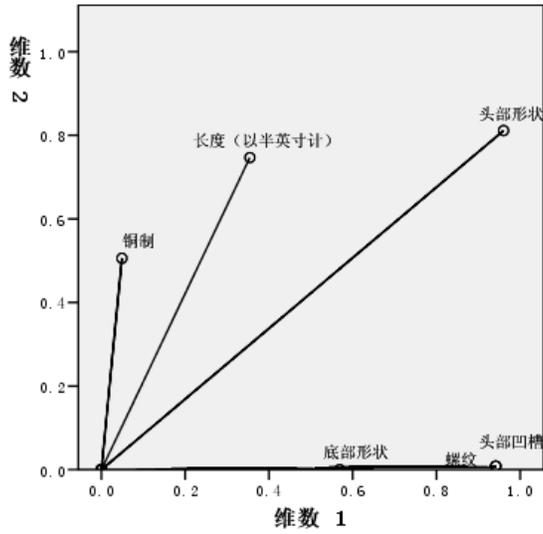
图片 13-15  
模型摘要（移去离群值）

维	Cronbach's Alpha	方差考虑情况		
		总计(特征值)	惯量	方差的 %
1	.685	3.815	.636	63.591
2	.623	2.081	.347	34.676
总计		5.896	.983	
均值	.793 <sup>a</sup>	2.948	.491	49.133

a. 总 Cronbach's Alpha 基于平均特征值。

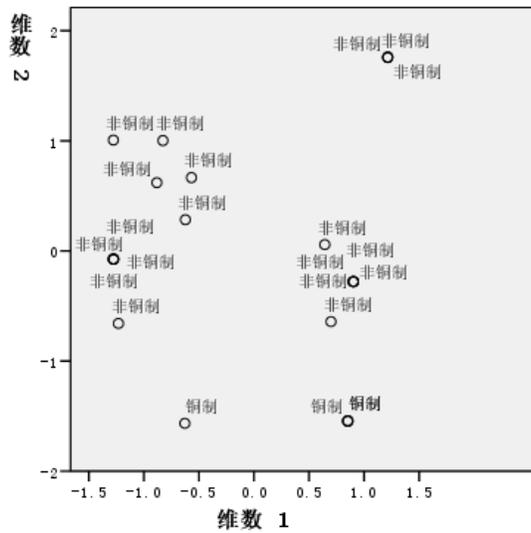
特征值会略有移位。现在第一维度可以解释略多的方差。

图片 13-16  
区分测量



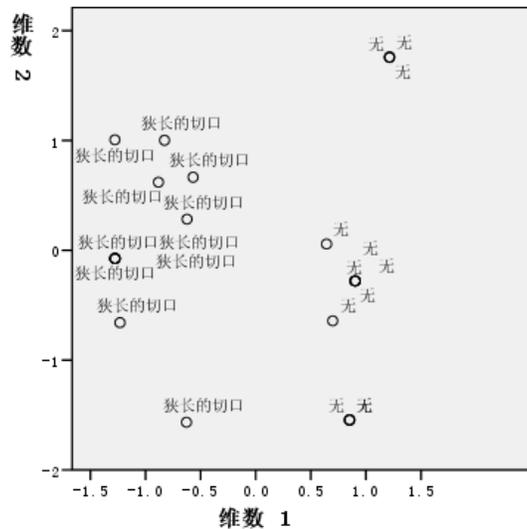
如同区分图中所示，头部凹槽不再在第二维度中区分，而铜制则从在两个维度中都没有区分改变为在第二维度中有区分。其他变量的区分基本上未发生变化。

图片 13-17  
用“铜制”标记的对象得分（移去离群值）



用铜制标记的对象得分图显示所有四个铜制对象都显示在图的底部附近（有三个对象的位置相同），这表明该对象沿第二维度高度区分。如同先前分析中螺纹的个案一样，对象未形成紧凑的组，但是该对象可以按照类别完全区分。

图片 13-18  
用“头部凹槽”标记的对象得分（移去离群值）



用头部凹槽标记的对象得分图显示第一维度可以完全区分无的对象和狭长的切口的对象，如同先前分析中的情况一样。不过，与先前的分析不同的是，第二维度现在无法区分这两个类别。

因此，忽略 SCREW1（唯一头部具有星形凹槽的对象）极大地影响了第二维度的解释。该维度现在根据铜制、头部形状和长度（以半英寸计）区分对象。

## 推荐参考

有关多重对应分析的更多信息，请参见以下内容：

Benzécri, J. P. 1992. Correspondence analysis handbook. New York: Marcel Dekker.

Guttman, L. 1941. The quantification of a class of attributes: A theory and method of scale construction. 输入: The Prediction of Personal Adjustment, P. Horst, ed. New York: Social Science Research Council.

Meulman, J. J. 1982. Homogeneity analysis of incomplete data. Leiden: DSWO Press.

Meulman, J. J. 1996. Fitting a distance model to homogeneous subsets of variables: Points of view analysis of categorical data. Journal of Classification, 13, .

Meulman, J. J., 和 W. J. Heiser. 1997. Graphical display of interaction in multiway contingency tables by use of homogeneity analysis. 输入: Visual Display of Categorical Data, M. Greenacre, 和 J. Blasius, eds. New York: Academic Press.

Nishisato, S. 1984. Forced classification: A simple application of a quantification method. *Psychometrika*, 49, .

Tenenhaus, M., 和 F. W. Young. 1985. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis, and other methods for quantifying categorical multivariate data. *Psychometrika*, 50, .

Van Rijckevorsel, J. 1987. The application of fuzzy coding and horseshoes in multiple correspondence analysis. Leiden: DSWO Press.

# 多维尺度

给定一组对象，多维尺度分析的目标是找到对象在低维空间中的表示形式。此解是使用对象间的**近似值**找到的。该过程计算初始的（也可能是转换后的）对象近似值与其低维空间中欧氏距离之间的偏移的平方的最小值。

低维空间的目的是揭示对象间的关系。通过将解限制为自变量的线性组合，您能够以这些变量来解释解的各个维。在下面的示例中，您将看到 15 个不同的亲属关系项如何以三维表示，以及如何根据每个项的性别、代和分离等级对该三维空间进行解释。

## 示例：亲属关系项的检验

Rosenberg 和 Kim (Rosenberg 和 Kim, 1975) 开始分析 15 个亲属关系项（伯母、兄弟、表兄妹、女儿、父亲、孙女、祖父、祖母、孙子、母亲、侄子或外甥、侄女或外甥女、姐妹、儿子和叔叔）。他们让四组大学生（两组女同学，两组男同学）根据相似程度将各项排序。他们让其中的两组同学（一组女同学，一组男同学）进行了两次排序，第二次排序和第一次排序采取的标准不同。这样，一共得到六组“源”，如下表所示。

表 14-1  
亲属关系数据的源结构

源	性别	条件	样本大小
1	女	单次排序	85
2	男	单次排序	85
3	女	第一次排序	80
4	女	第二次排序	80
5	男	第一次排序	80
6	男	第二次排序	80

每个源对应一个  $15 \times 15$  的近似值矩阵，其单元格中的值等于源中的人数减去此源中对象被划分的次数。此数据集可在 kinship\_dat.sav 中找到。有关详细信息，请参阅附录 A 中的样本文件中的 IBM SPSS Categories 21。

## 选择维数

您可以自由决定解应该有多少个维。碎石图可以帮助您做出决定。

- 要创建碎石图，从菜单中选择：  
分析 > 标度 > 多维尺度 (PROXSCAL)...

图片 14-1  
“数据格式”对话框



- ▶ 在“源的数目”组选项中选择多个矩阵源。
- ▶ 单击定义。

图片 14-2  
“多维尺度”对话框



- ▶ 选择伯母至叔叔作为近似值变量。
- ▶ 选择 sourceid 作为用于标识源的变量。
- ▶ 单击模型。

图片 14-3  
“模型”对话框



- ▶ 键入 10 作为维数的最大值。
- ▶ 单击继续。
- ▶ 在“多维尺度”对话框中单击限制。

图片 14-4  
“限制”对话框



- ▶ 选择自变量的线性组合。
- ▶ 单击文件选择自变量的源。
- ▶ 选择 kinship\_var.sav。

图片 14-5  
“限制”对话框



- 选择性别、代和等级作为约束变量。

注意：变量性别有一个用户定义的缺失值 — 9 = 缺失（对于表兄妹）。该过程将其视为有效类别。这样，缺省线性转换不太可能适用。改用名义转换。

图片 14-6  
“限制”对话框



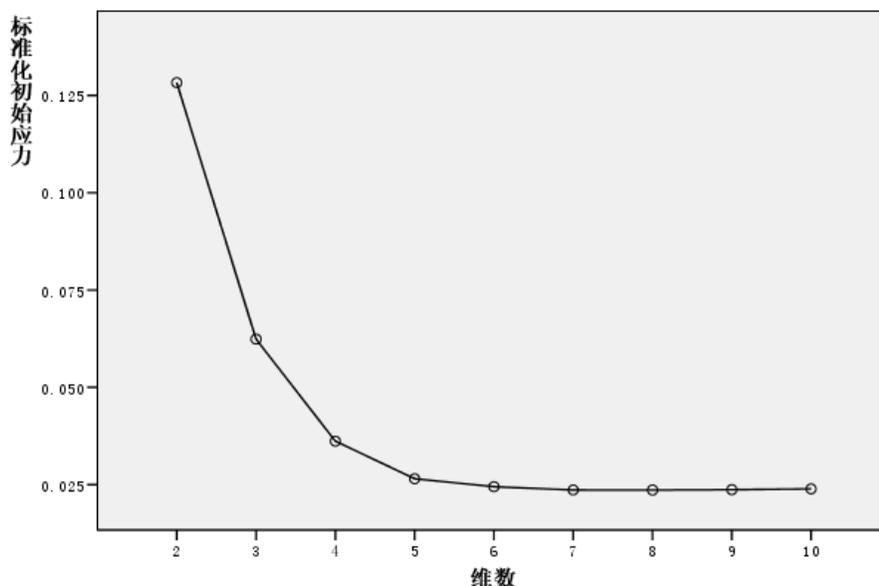
- ▶ 选择性别。
- ▶ 从“自变量转换”下拉列表中选择名义。
- ▶ 单击更改。
- ▶ 单击继续。
- ▶ 在“多维尺度”对话框中单击图。

图片 14-7  
“图”对话框



- ▶ 在“图”组选项中，选择应力。
- ▶ 单击继续。
- ▶ 在“多维尺度”对话框中单击确定。

图片 14-8  
碎石图



该过程以 10 维解开始，以 2 维解结束。碎石图显示每个维度中解的标准化初始应力。您可以在图中看到，当维数从 2 增加到 3 和从 3 增加到 4 应力会大为改进。从 4 维再增加维数，应力的改进变得很小。您将选择使用 3 维解选择分析数据，因为结果更易于解释。

### 三维解

创建自变量性别、代和（分离的）等级的目的是使用它们解释解的维度。创建的自变量如下所示：

性别	1 = 男性, 2 = 女性, 9 = 缺失（对于表兄妹）
代	距离您的代数（如果该项指您的亲属），较小数字对应较早一代。这样，祖父母就是 - 2，孙子和孙女是 2，兄弟姐妹是 0。
等级	在您的家庭树中分离的等级数。这样，父母要向上 1 个节点，子女要向下 1 个节点。兄弟姐妹要先向上 1 个节点到父母，然后向下 1 个节点到他们，分离级别为 2。和表兄妹之间的分离级别为 4 —要向上 2 个节点到祖父母，然后向下 2 个节点经过伯母/叔叔到达他们。

外部变量可以在 kinship\_var.sav 中找到。此外，kinship\_ini.sav 提供了早期分析的初始配置。有关详细信息，请参阅附录 A 中的样本文件中的 IBM SPSS Categories 21。

## 运行分析

图片 14-9  
“模型”对话框



- ▶ 要获取三维解，调用“多维尺度”对话框，然后单击模型。
- ▶ 键入 3 作为维数的最小值和最大值。
- ▶ 单击继续。
- ▶ 在“多维尺度”对话框中单击选项。

图片 14-10  
“选项”对话框



- ▶ 选择定制作为初始配置。
- ▶ 选择 kinship\_ini.sav 作为从中读取变量的文件。
- ▶ 选择 dim01、dim02 和 dim03 作为变量。
- ▶ 单击继续。
- ▶ 在“多维尺度”对话框中单击图。

图片 14-11  
“图”对话框



- ▶ 选择初始近似值与转换近似值和转换自变量。
- ▶ 单击继续。
- ▶ 在“多维尺度”对话框中单击输出。

图片 14-12  
“输出”对话框



- ▶ 选择输入数据、应力分解和变量与维数相关性。
- ▶ 单击继续。
- ▶ 在“多维尺度”对话框中单击确定。

## 应力度量标准

应力和拟合测量表明解中的距离与初始距离的近似程度。

图片 14-13  
应力和拟合测量

标准化初始应力	.06234
Stress-I	.24968 <sup>a</sup>
Stress-II	.87849 <sup>a</sup>
S-Stress	.14716 <sup>b</sup>
离散所占比例 (D. A. F.)	.93766
Tucker 冗余系数	.96833

PROXSCAL 使“标准化初始应力”最小化。

- a. 最优定标因子 = 1.066。
- b. 最优定标因子 = .984。

所有四个应力统计量测量数据的不拟合度，而离差考虑情况和 Tucker' s 同余系数则度量拟合度。较低的应力度量标准（最低为 0）和较高的拟合测量（最高为 1）指示了较好的解。

图片 14-14  
标准化初始应力的分解

对象	源						均值
	SRC_1	SRC_2	SRC_3	SRC_4	SRC_5	SRC_6	
伯母/婶母/舅母 /姑母/姨母	.0991	.0754	.0629	.0468	.0391	.0489	.0620
兄/弟	.1351	.0974	.0496	.0813	.0613	.0597	.0807
堂兄弟/堂姐妹/ 表兄弟/表姐妹	.0325	.0336	.0480	.0290	.0327	.0463	.0370
女儿	.0700	.0370	.0516	.0229	.0326	.0207	.0391
父亲	.0751	.0482	.0521	.0225	.0272	.0298	.0425
孙女	.1410	.0736	.0801	.0707	.0790	.0366	.0802
祖父	.1549	.1057	.0858	.0821	.0851	.0576	.0952
祖母	.1550	.0979	.0858	.0844	.0816	.0627	.0946
孙子	.1374	.0772	.0793	.0719	.0791	.0382	.0805
母亲	.0813	.0482	.0526	.0229	.0260	.0227	.0423
侄子/外甥	.0843	.0619	.0580	.0375	.0317	.0273	.0501
侄女/外甥女	.0850	.0577	.0503	.0353	.0337	.0260	.0480
姐/妹	.1361	.0946	.0496	.0816	.0629	.0588	.0806
儿子	.0689	.0373	.0456	.0242	.0337	.0253	.0392
伯父/叔叔/舅舅	.0977	.0761	.0678	.0489	.0383	.0498	.0631
均值	.1035	.0681	.0613	.0508	.0496	.0407	.0623

应力的分解可帮助您确定哪些源和对象对解的整体应力贡献最大。在这种情况下，源中大多数应力可归因于源 1 和源 2；而在对象中，大多数应力可归因于兄弟、孙女、祖父、祖母、孙子和姐妹。

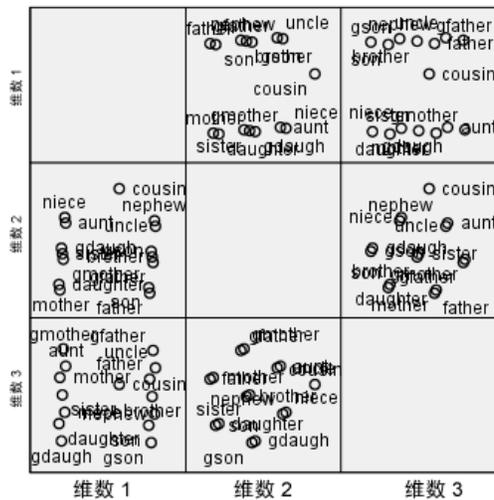
大多数应力所归因于的两个源是仅将项进行一次排序的两个组。此信息表明学生对项进行排序时考虑了多个因素。对项进行两次排序的学生在第一次排序时着重考虑其中一部分因素，然后在第二次排序时着重考虑其余的因素。

大多数应力归因于的对象的等级为 2。这些人虽然不是“核心”家庭成员（母亲、父亲、女儿、儿子），但是较其他关系更为亲近。此中间位置容易导致对这些项的排序出现差异。

## 公共空间的最终坐标

公共空间图提供了对象间关系的可视表示。

图片 14-15  
公共空间坐标



查看维度 1 和维度 3 中对象的最终坐标；这是散点图矩阵左下角的图。此图显示维度 1（在 x 轴上）与变量性别关联，维度 3（在 y 轴上）与代关联。从左至右，可以看到维度 1 将女性和男性项分隔，中间是不区分性别的项表兄妹。从图的底部到顶部，沿着轴逐渐增加的值与年长的项相对应。

查看维度 2 和维度 3 中对象的最终坐标，这是散点图矩阵中部右侧的图。从此图中，您可以看到第二维（沿 y 轴）对应于变量等级，沿着轴方向较大的值对应于和“核心”家庭成员关系较远的项。

### 使用非缺省转换的三维解

上一个解是使用近似值的缺省比率转换和自变量代与等级的区间转换来计算的。结果很不错，但是如果使用其他转换可能得到更好的结果。例如，虽然近似值代和等级都有自然排序，但是比起线性转换，有序转换能更好地建立其模型。

图片 14-16  
“模型”对话框



- ▶ 要重新运行分析，调整处于有序尺度（保留结）的代和等级的近似值，调用“多维尺度”对话框，然后单击模型。
- ▶ 选择序数作为近似值转换。
- ▶ 单击继续。
- ▶ 在“多维尺度”对话框中单击限制。

图片 14-17  
“限制”对话框



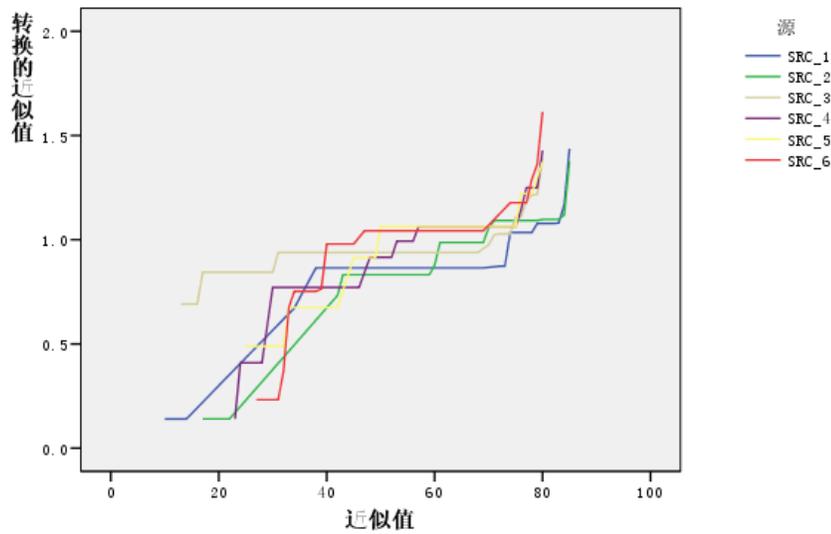
- ▶ 选择代和等级。
- ▶ 从“自变量转换”下拉列表中选择有序（保留结）。
- ▶ 单击更改。
- ▶ 单击继续。
- ▶ 在“多维尺度”对话框中单击确定。

## 变换图

变换图是第一次检验初始转换是否适用的好方法。如果图为近似线性，则线性假设适用。如果不是线性，查看应力度量标准，看看拟合度是否有改善，再查看公共空间图，看看解释是否更有用。

因为每个自变量都得到近似线性转换，所以可能适合将其作为数字进行解释。但是，因为近似值不获得线性转换，所以有序转换可能更适合近似值。

图片 14-18  
转换后的近似值



## 应力度量标准

当前解的应力支持在有序尺度调整近似值大小的参数。

图片 14-19  
应力和拟合测量

标准化初始应力	.07976
Stress-I	.28241 <sup>a</sup>
Stress-II	.79034 <sup>a</sup>
S-Stress	.21220 <sup>b</sup>
离散所占比例 (D. A. F.)	.92024
Tucker 冗余系数	.95929

PROXSCAL 使“标准化初始应力”最小化。

a. 最优定标因子 = 1.087。

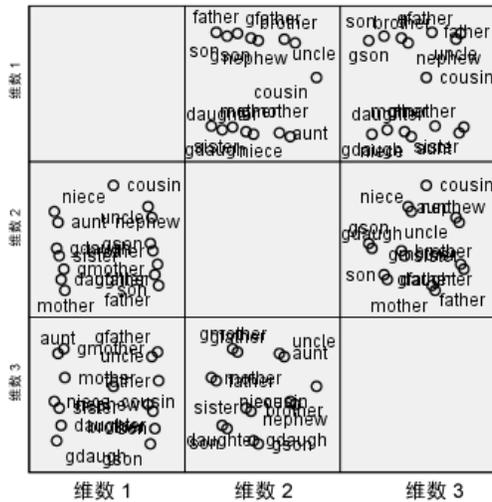
b. 最优定标因子 = .931。

前一个解的标准化初始应力是 0.06234。使用非缺省转换调整变量大小会将应力减半为 0.03137。

## 公共空间的最终坐标

公共空间图提供与前一个解完全相同的维度的解释。

图片 14-20  
公共空间坐标



## 讨论

最好将近似值作为有序变量，因为这样会使应力度量标准有很大改善。作为下一步，您可能要“解开”有序变量——即允许相等的初始变量值在转换后得到不同的值。例如，在第一个源中，伯母和儿子之间的近似值以及伯母和孙子之间的近似值为 85。有序变量的“打结的”方法强制使这些近似值转换为相同的值，但是没有特别的理由假设应该是这样。这种情况下，允许近似值解开将使您免受不必要的限制。

## 推荐参考

有关多维尺度分析的更多信息，请参见以下内容：

Commandeur, J. J. F., 和 W. J. Heiser. 1993. Mathematical derivations in the proximity scaling (PROXSCAL) of symmetric data matrices. Leiden: Department of Data Theory, University of Leiden.

De Leeuw, J., 和 W. J. Heiser. 1980. Multidimensional scaling with restrictions on the configuration. 输入: Multivariate Analysis, Vol. V, P. R. Krishnaiah, ed. Amsterdam: North-Holland.

Heiser, W. J. 1981. Unfolding analysis of proximity data. Leiden: Department of Data Theory, University of Leiden.

Heiser, W. J., 和 F. M. T. A. Busing. 2004. Multidimensional scaling and unfolding of symmetric and asymmetric proximity relations. 输入: Handbook of Quantitative Methodology for the Social Sciences, D. Kaplan, ed. Thousand Oaks, Calif.: Sage Publications, Inc..

Kruskal, J. B. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, 29, .

---

Kruskal, J. B. 1964. Nonmetric multidimensional scaling:A numerical method. Psychometrika, 29, .

Shepard, R. N. 1962. The analysis of proximities:Multidimensional scaling with an unknown distance function I. Psychometrika, 27, .

Shepard, R. N. 1962. The analysis of proximities:Multidimensional scaling with an unknown distance function II. Psychometrika, 27, .

# 多维展开

“多维展开”过程尝试查找允许您直观地检查两组对象之间关系的公共定量尺度。

## 示例：早餐食品喜好程度

在一项经典研究 (Green 和 Rao, 1972) 中, 21 名 Wharton School MBA 学生及其配偶被要求按照喜好程度顺序对 15 种早餐食品进行评价, 1 = “最喜欢”, 15 = “最不喜欢”。此信息收集在 breakfast\_overall.sav 中。有关详细信息, 请参阅附录 A 中的样本文件中的 IBM SPSS Categories 21。

该研究的结果提供了大多数多维展开算法固有的退化问题的一个典型示例。多维展开算法是通过将转换后的近似值的变异系数加以惩罚来求解的 (Busing, Groenen, 和 Heiser, 2005)。您将看到退化解以及如何使用多维展开来对问题进行求解, 从而确定不同的人是如何区别对待早餐食品的。您可以在 prefscal\_breakfast-overall.sps 中找到再次生成这些分析的语法。

## 生成退化解

- ▶ 要运行“多维展开”分析, 请从菜单中选择:  
分析 > 标度 > 多维展开 (PREFSCAL)...

图片 15-1  
“多维展开”主对话框



- ▶ 选择 Toast pop-up 到 Corn muffin and butter 作为近似值变量。

- ▶ 单击选项。

图片 15-2  
“选项”对话框



- ▶ 选择 Spearman 作为“经典”起点的归因依据。
- ▶ 在“惩罚项”组选项中，键入 1.0 作为“强度”参数的值，0.0 作为“全距”参数的值。这将关闭惩罚项。
- ▶ 单击继续。
- ▶ 在“多维展开”对话框中单击确定。

以下是由这些选择所生成的命令语法：

```

PREFSCAL
VARIABLES=TP BT EMM JD CT BMM HRB TMd BTJ TMn CB DP GD CC CMB
/INITIAL=CLASSICAL (SPEARMAN)
/TRANSFORMATION=NONE
/PROXIMITIES=DISSIMILARITIES
/CRITERIA=DIMENSIONS (2, 2) DIFFSTRESS(.000001) MINSTRESS(.0001)
MAXITER (5000)
/PENALTY=LAMBDA (1.0) OMEGA (0.0)
/PRINT=MEASURES COMMON
/PLOT=COMMON .

```

- 此语法指定对变量 tp (Toast pop-up) 到 cmb (Corn muffin and butter) 的分析。

- INITIAL 子命令指定起始值应使用 Spearman 间距归因。
- 指定给 PENALTY 子命令的值基本上将关闭惩罚项。结果是，该过程将最小化 Kruskal 的 Stress-I。这样会产生退化解。
- PLOT 子命令请求公共空间的图。
- 所有其他参数回退为其缺省值。

## 度量

图片 15-3  
退化解的测量

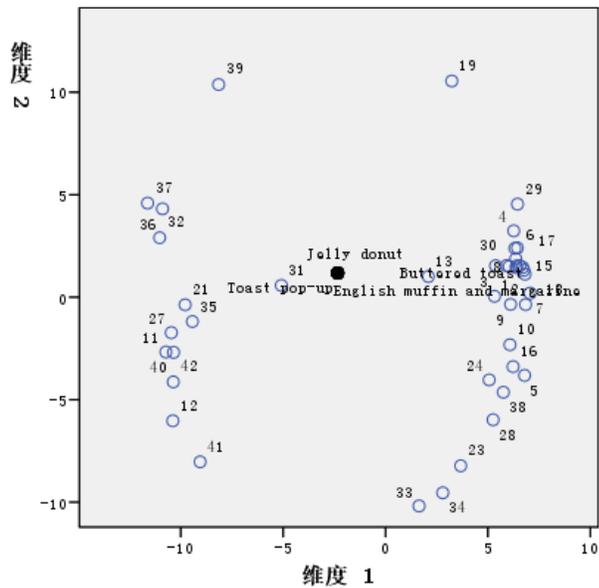
迭代		154
最终的函数值		.0000990
函数值部分	应力部分	.0000990
	惩罚部分	1.0000000
拟合的不良程度	正态化的应力	.0000000
	Kruskal 的应力-I	.0000990
	Kruskal 的应力-II	.6129749
	Young 的 S-应力-I	.0001980
	Young 的 S-应力-II	.7703817
拟合优度	考虑的离散	1.0000000
	考虑的方差	.6230788
	恢复的首选项顺序	.7074830
	Spearman 的 Rho	.7450748
	Kendall 的 Tau-b	.6218729
方差系数	方差近似值	.5590170
	转换的方差近似值	.0000924
	方差距离	.1808765
衰退指标	DeSarbo 的混合指标的平方和	117.3115413
	Shepard 的粗略非衰退索引	.0000000

该算法在第 154 次迭代后收敛，最终惩罚强度（标记的最终函数值）为 0.0000990。因为惩罚项已经关闭，惩罚强度与 Kruskal 的 Stress-I 相同（函数值的强度部分与 Kruskal 的拟合劣度量相同）。低强度值通常表示解与数据拟合得好，但是还有一些退化解的警告信号：

- 转换后的近似值的变异系数与初始近似值的变异系数的关系很小。这表示每行转换后的近似值近似于常数，因此解不提供任何对象间的区分。
- DeSarbo 混合度指标的平方和是不同集合的点的混合程度的测量。如果它们没有混合，则是解有可能退化的警告信号。越接近 0，解的混合程度就越高。报告的值很大，表示解没有混合。
- Shepard 粗略非退化指标是作为不同距离的百分比报告的，其值为 0。这是一个清晰的数字指示：不同距离不足，解可能退化。

## 公共空间

图片 15-4  
退化解的公共空间的联合图



在行与列对象的公共空间的联合图中，可以通过可视方式确认解是退化的。行对象（个人）位于中心为列对象（早餐食品）的圆周的圆周上，其坐标缩小为一个点。

## 运行非退化分析

图片 15-5  
“选项”对话框



- ▶ 要生成非退化解，单击“对话框调用”工具，然后选择多维展开。
- ▶ 在“多维展开”对话框中单击选项。
- ▶ 在“惩罚项”组选项中，键入 0.5 作为“强度”参数的值，1.0 作为“全距”参数的值。这将关闭惩罚项。
- ▶ 单击继续。
- ▶ 在“多维展开”对话框中单击确定。

以下是由这些选择所生成的命令语法：

```

PREFSCAL
VARIABLES=TP BT EMM JD CT BMM HRB TmD BTJ TmN CB DP GD CC CMB
/INITIAL=CLASSICAL (SPEARMAN)
/TRANSFORMATION=NONE
/PROXIMITIES=DISSIMILARITIES
/CRITERIA=DIMENSIONS(2, 2) DIFFSTRESS(.000001) MINSTRESS(.0001)
MAXITER(5000)
/PENALTY=LAMBDA(0.5) OMEGA(1.0)
/PRINT=MEASURES COMMON
/PLOT=COMMON .

```

- 唯一的变化是 PENALTY 子命令。LAMBDA 设置为 0.5，OMEGA 设置为 1.0，这是它们的缺省值。

## 度量

图片 15-6  
非退化解的测量

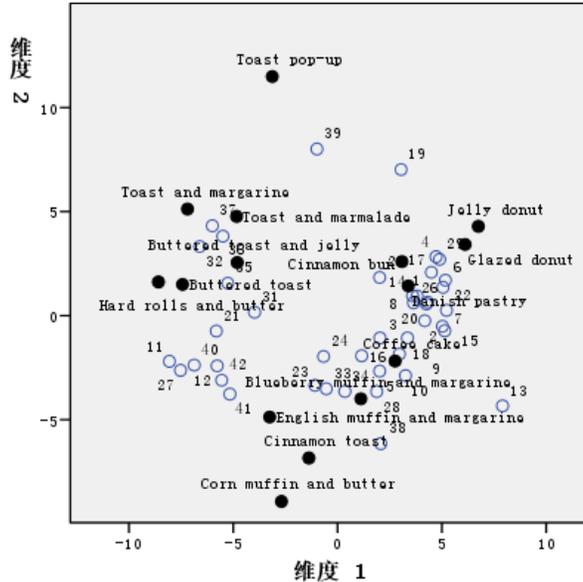
迭代		157
最终的函数值		.6848930
函数值部分	应力部分	.2428268
	惩罚部分	1.9317409
拟合的不良程度	正态化的应力	.0583589
	Kruskal 的应力-I	.2415758
	Kruskal 的应力-II	.5875599
	Young 的 S-应力-I	.3446361
	Young 的 S-应力-II	.5030127
拟合优度	考虑的离散	.9416411
	考虑的方差	.7651552
	恢复的首选项顺序	.7818594
	Spearman 的 Rho	.8179181
	Kendall 的 Tau-b	.6916725
方差系数	方差近似值	.5590170
	转换的方差近似值	.6006156
	方差距离	.4833617
衰退指标	DeSarbo 的混合指标的平方和	.1590979
	Shepard 的粗略非衰退索引	.7895692

在退化解的测量中提到的问题在这里得到了修正。

- 标准化强度不再是 0。
- 转换后的近似值的变异系数现在有一个和初始近似值的变异系数相似的值。
- DeSarbo 混合度指标更接近 0，表示解得到了较好的混合。
- Shepard 粗略非退化指标是作为不同距离的百分比报告的，现在接近 80%。有足够的不同距离，解可能是非退化的。

## 公共空间

图片 15-7  
非退化解的公共空间的联合图



通过公共空间的联合图，可以对各维度进行解释。水平维度在图中区分柔软 bread 或 toast 和坚硬 bread 或 toast，越靠近水平轴的右侧，食品越软。垂直维度没有清晰的解释，尽管可能是出于方便而进行的区分。沿垂直轴向下，食品越来越“正式”。

这样就产生了早餐食品的数个聚类。例如，donut、cinnamon bun 和 Danish pastry 组成了柔软且不太正式的食品的一个聚类。Muffin 和 cinnamon toast 组成了较硬但较正式的食品的一个聚类。其他的 toast 和 hard roll 组成了坚硬且不太正式的食品的一个聚类。Toast pop-up 是极不正式的坚硬食品。

由行对象所表示的个人根据对软硬食品的喜好程度清晰划分成聚类，而沿垂直维度有一定的聚类内的变异。

## 示例：早餐食品喜好程度的三维展开

在一项经典研究 (Green 等., 1972) 中, 21 名 Wharton School MBA 学生及其配偶被要求按照喜好程度顺序对 15 种早餐食品进行评价, 1 = “最喜欢”, 15 = “最不喜欢”。他们的喜好根据六种不同的情况加以记录, 从“全部喜欢”到“只带饮料的快餐”。此信息收集在 breakfast.sav。有关详细信息, 请参阅附录 A 中的样本文件中的 IBM SPSS Categories 21。

六种情况可视为各自独立的源。使用 PREFSCAL 进行行、列和源的三维展开。您可以在 prefscal\_breakfast.sps 中找到再次生成这些分析的语法。

## 运行分析

- ▶ 要运行“多维展开”分析，请从菜单中选择：  
分析 > 标度 > 多维展开 (PREFSCAL)...

图片 15-8  
“多维展开”主对话框



- ▶ 选择 Toast pop-up 到 Corn muffin and butter 作为近似值变量。
- ▶ 选择 Menu scenarios 作为源变量。
- ▶ 单击模型。

图片 15-9  
“模型”对话框



- ▶ 选择加权欧几里德作为尺度模型。
- ▶ 单击继续。
- ▶ 在“多维展开”对话框中单击选项。

图片 15-10  
“选项”对话框



- ▶ 选择 Spearman 作为“经典”起点的归因依据。
- ▶ 单击继续。
- ▶ 在“多维展开”对话框中单击图。

图片 15-11  
“图”对话框



- ▶ 在“图”组选项中选择私有空间。
- ▶ 单击继续。
- ▶ 在“多维展开”对话框中单击确定。

以下是由这些选择所生成的命令语法：

```

PREFSCAL
VARIABLES=TP BT EMM JD CT BMM HRB TMd BTJ TMn CB DP GD CC CMB
/INPUT=SOURCES(srcid)
/INITIAL=CLASSICAL (SPEARMAN)
/CONDITION=ROW
/TRANSFORMATION=NONE
/PROXIMITIES=DISSIMILARITIES
/MODEL=WEIGHTED
/CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)

```

```

MAXITER(5000)
/PENALTY=LAMBDA(0.5) OMEGA(1.0)
/PRINT=MEASURES COMMON
/PLOT=COMMON WEIGHTS INDIVIDUAL ( ALL ) .

```

- 此语法指定对变量 tp (Toast pop-up) 到 cmb (Corn muffin and butter) 的分析。变量 srcid 用于标识源。
- INITIAL 子命令指定起始值应使用 Spearman 间距归因。
- MODEL 子命令指定一个加权 Euclidean 模型，该模型使每个私有空间都可以对公共空间的维数进行不同的加权。
- PLOT 子命令请求公共空间、私有空间和私有空间权重的图。
- 所有其他参数回退为其缺省值。

## 度量

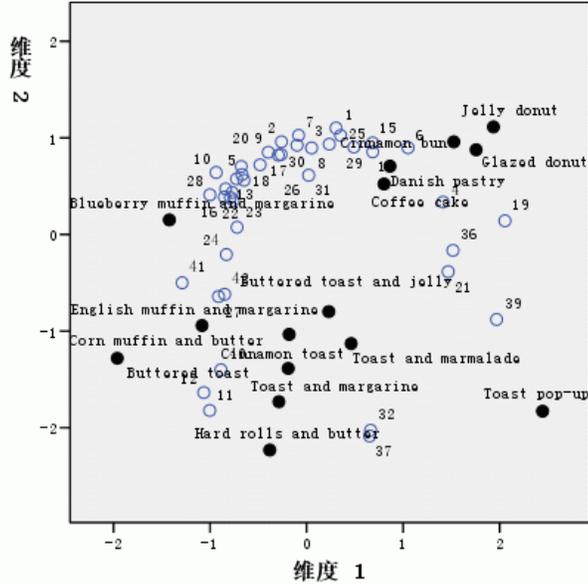
图片 15-12  
度量

迭代		408
最终的函数值		.8890680
函数值部分	应力部分	.3815927
	惩罚部分	2.0714286
拟合的不良程度	正态化的应力	.1456130
	Kruskal 的应力-I	.4212509
	Kruskal 的应力-II	1.8458204
	Young 的 S-应力-I	.5467283
	Young 的 S-应力-II	.8110237
拟合优度	考虑的离散	.8543870
	考虑的方差	.3502584
	恢复的首选项顺序	.7199546
	Spearman 的 Rho	.5763322
	Kendall 的 Tau-b	.4399093
方差系数	方差近似值	.5590170
	转换的方差近似值	.5590170
	方差距离	.3004635
衰退指标	DeSarbo 的混合指标的平方和	3.0076344
	Shepard 的粗略非衰退索引	.6230915

算法在 481 次迭代后收敛，最终惩罚强度为 0.8199642。变异系数和 Shepard 指标足够大，并且 DeSarbo 指标足够低，表示退化没有问题。

## 公共空间

图片 15-13  
公共空间的联合图



公共空间的联合图显示与整体喜好的二维分析十分相似的最终配置，解在 45 度线附近。这样，垂直维度在图中区分柔软和坚硬的 bread 或 toast，沿垂直轴向上，食品越来越柔软。水平维度没有清晰的解释，尽管可能是出于方便而进行区分。沿垂直轴向左，食品越来越“正式”。

行对象所表示的个人仍然根据对“坚硬”或“柔软”食品的喜好程度清晰划分成聚类，沿水平维度有一定的聚类内的变异。

## 私有空间

图片 15-14  
维度权重

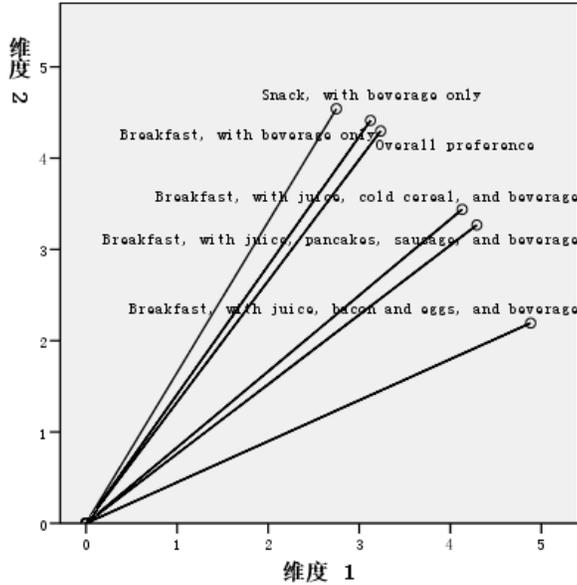
源		维度		特征 <sup>a</sup>
		1	2	
Overall preference		3.235	4.297	.186
Breakfast, with juice, bacon and eggs, and beverage		4.883	2.193	.457
Breakfast, with juice, cold cereal, and beverage		4.131	3.438	.109
Breakfast, with juice, pancakes, sausage, and beverage		4.291	3.267	.164
Breakfast, with beverage only		3.124	4.413	.223
Snack, with beverage only		2.750	4.541	.313
重要性 <sup>b</sup>		.504	.496	

- a. 特征表明源的典型性。特征的范围介于 0 与 1 之间，0 表明具有相同维度权重的均匀源，1 表明某一维度权重特别大而其他权重接近于零的特殊源。
- b. 每个维度的相对重要性，以一个维度的平方和与总平方和的比的形式给出。

为每个源计算一个私有空间。维度权重表明私有空间如何加载到公共空间的维度上。较大的权重表示私有空间中较大的距离，从而使得私有空间该维度中对象间区别较大。

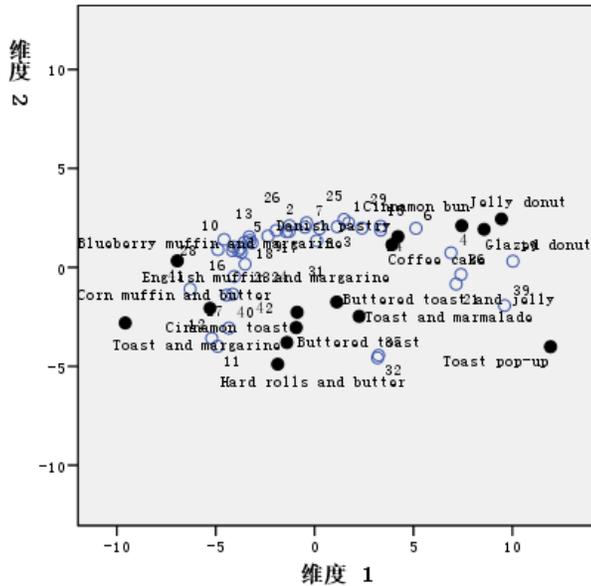
- **特异性**是对私有空间与公共空间的差异程度的测量。与公共空间相同的私有空间有相同的维度权重，其特异性为 0，而特定于某一特定维度的私有空间有单一的较大维度权重，其特异性为 1。在这种情况下，分歧最大的源是 Breakfast, with juice, bacon and eggs, and beverage 和 Snack, with beverage only。
- **重要性**是各个维度对解的相对贡献的测量。在这种情况下，各维度同等重要。

图片 15-15  
维度权重



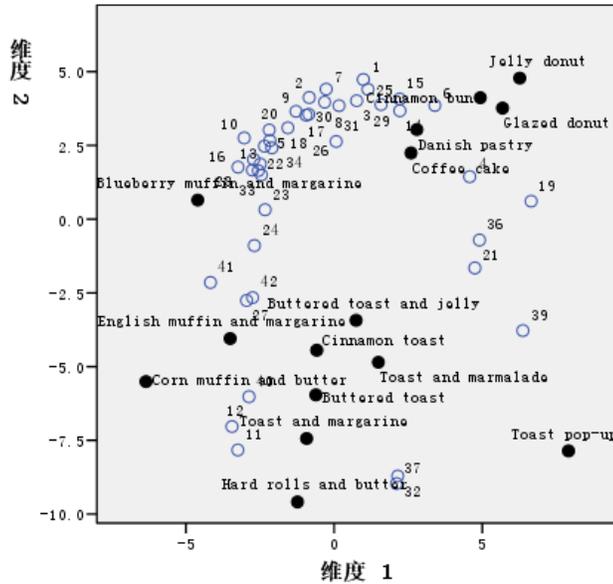
维度权重图提供了权重表的可视表达。Breakfast, with juice, bacon and eggs, and beverage 和 Snack, with beverage only 与维度轴最近，但是它们都不与特定的维度具有强烈的关联度。

图片 15-16  
私有空间“Breakfast, with juice, bacon and eggs, and beverage”的联合图



私有空间的联合图 Breakfast, with juice, bacon and eggs, and beverage 显示了此情况下喜好程度的效果。因为此源更多地加载在第一个维度上，所以食品的区别主要来自第一个维度。

图片 15-17  
私有空间 “Snack, with beverage only” 的联合图



私有空间 Snack, with beverage only 的联合图显示了此情况下喜好程度的效果。因为此源更多地加载在第二个维度上，所以食品的区别主要来自第二个维度。但是，仍有一定程度的区别来自第一个维度，因为此源的特异性较低。

## 使用不同的初始配置

最终配置可由给予算法的起始点来确定。理想状态下，解的常规结构应保持一致；否则很难确定哪个解是正确的。但是，当您尝试不同的初始配置时，详细信息可能会汇聚到更小的焦点上，如在早餐数据的三维分析上使用对应起点。

- ▶ 要使用对应起点生成一个解，单击“对话框调用”工具，然后选择多维展开。

- ▶ 在“多维展开”对话框中单击选项。

图片 15-18  
“选项”对话框



- ▶ 在“初始配置”组选项中选择对应。
- ▶ 单击继续。
- ▶ 在“多维展开”对话框中单击确定。

以下是由这些选择所生成的命令语法：

```

PREFSCAL
VARIABLES=TP BT EMM JD CT BMM HRB TMd BTJ TMn CB DP GD CC CMB
/INPUT=SOURCES(srcid)
/INITIAL=CORRESPONDENCE
/TRANSFORMATION=NONE
/PROXIMITIES=DISSIMILARITIES
/CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
MAXITER(5000)
/PENALTY=LAMBDA(0.5) OMEGA(1.0)
/PRINT=MEASURES COMMON
/PLOT=COMMON WEIGHTS INDIVIDUAL ( ALL ) .

```

- 唯一的变化是 INITIAL 子命令。起始配置已设置为 CORRESPONDENCE，它使用反向数据（相似性而非相异性）的对应分析的结果，且对行和列的得分进行对称的标准化。

## 度量

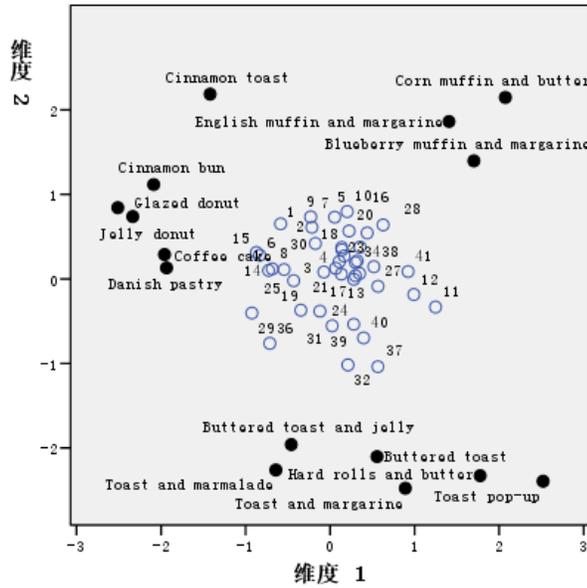
图片 15-19  
对应初始配置的测量

迭代		118
最终的函数值		.6893480
函数值部分	应力部分	.3818331
	惩罚部分	2.0714286
拟合的不良程度	正态化的应力	.1457965
	Kruskal 的应力-I	.4158619
	Kruskal 的应力-II	1.8055362
	Young 的 S-应力-I	.5448719
	Young 的 S-应力-II	.8082700
拟合优度	考虑的离散	.8542035
	考虑的方差	.3493448
	恢复的首选项顺序	.7243008
	Spearman 的 Rho	.5843396
	Kendall 的 Tau-b	.4486017
方差系数	方差近似值	.5590170
	转换的方差近似值	.5590170
	方差距离	.2967839
衰退指标	DeSarbo 的混合指标的平方和	2.9892074
	Shepard 的粗略非衰退索引	.6432351

算法在 385 次迭代后收敛，最终惩罚强度为 0.8140741。此统计量、拟合劣度、拟合优度、变异系数和 Shepard 指标都与使用经典 Spearman 起点的解很相似。DeSarbo 指标有一些不同，其值为 1.7571887 而不是 0.2199287，这表明使用对应起点的解没有很好地混合。要了解其对解的影响，请参见公共空间的联合图。

### 公共空间

图片 15-20  
对应初始配置的公共空间的联合图



此公共空间的联合图显示与使用经典 Spearman 初始配置的分析相似的最终配置；但是列对象（早餐食品）位于行对象（个人）的四周，而没有与行对象混合。

### 私有空间

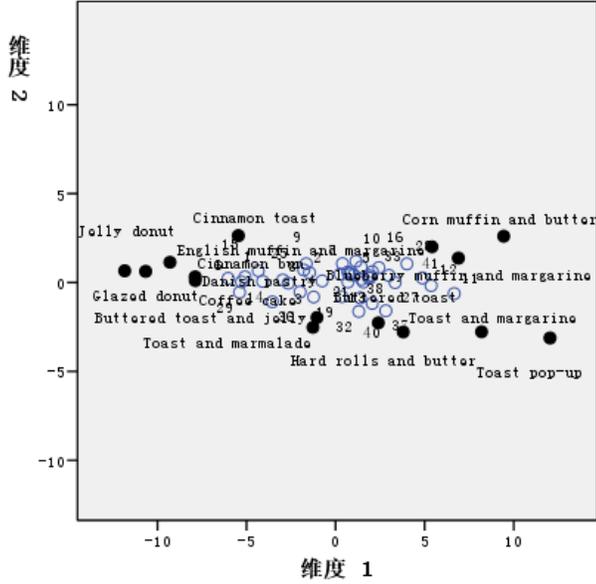
图片 15-21  
对应初始配置的维度权重

		维度		特征 <sup>a</sup>
		1	2	
源	Overall preference	2.836	3.877	.279
	Breakfast, with juice, bacon and eggs, and beverage	4.727	1.207	.636
	Breakfast, with juice, cold cereal, and beverage	4.183	2.377	.263
	Breakfast, with juice, pancakes, sausage, and beverage	4.412	1.993	.389
	Breakfast, with beverage only	2.605	4.050	.351
	Snack, with beverage only	1.864	4.415	.552
重要性 <sup>b</sup>		.556	.444	

- a. 特征表明源的典型性。特征的范围介于 0 与 1 之间，0 表明具有相同维度权重的均匀源，1 表明某一维度权重特别大而其他权重接近于零的特殊源。
- b. 每个维度的相对重要性，以一个维度的平方和与总平方和的比的形式给出。

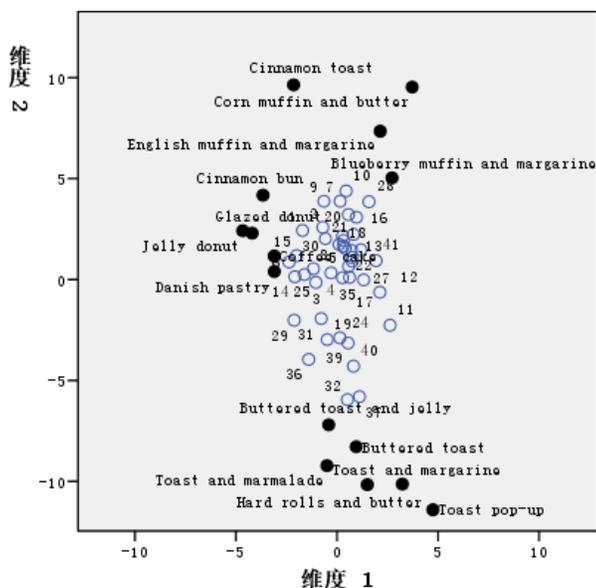
在对应初始配置中，每个私有空间都有一个较高的特异性；也就是说，参与者对早餐食品归类的每个情景都更紧密地与特定维度相关联。分歧最大的源还是 Breakfast, with juice, bacon and eggs, and beverage 和 Snack, with beverage only。

图片 15-22  
对应初始配置的私有空间 “Breakfast, with juice, bacon and eggs, and beverage” 的联合图



较高的特异性在私有空间 Breakfast, with juice, bacon and eggs, and beverage 的联合图中很明显。源在第一个维度上的荷载甚至比其在经典 Spearman 起点下的荷载更大，因此行和列的对象在垂直轴上显示较少的变异，在水平轴上显示较多的变异。

图片 15-23  
对应初始配置的私有空间 “Snack, with beverage only” 的联合图



私有空间 Snack, with beverage only 的联合图显示行和列对象比在经典 Spearman 起点下离垂直线更近。

## 示例：检查行为情况的得体性

在一个经典示例中(Price 和 Bouffard, 1974), 52 名学生被要求以 10 点的标度对 15 种情况和 15 种行为的组合进行评价, 该 10 点的标度从 0 = “极得体” 到 9 = “极不得体”。平均值在个人值之上, 值被视为相异性。

这些信息收集在 behavior.sav 中。[有关详细信息, 请参阅附录 A 中的样本文件中的 IBM SPSS Categories 21。](#)使用“多维展开”得到相似情况的聚类 and 与其关联最紧密的行为。您可以在 prefscal\_behavior.sps 中找到再次生成这些分析的语法。

## 运行分析

- ▶ 要运行“多维展开”分析, 请从菜单中选择:  
分析 > 标度 > 多维展开 (PREFSCAL)...

图片 15-24  
“多维展开”主对话框



- ▶ 选择 Run 到 Shout 作为近似值变量。
- ▶ 选择 ROWID 作为行变量。
- ▶ 单击模型。

图片 15-25  
“模型”对话框



- ▶ 选择线性作为近似值转换，然后选择包括截距。
- ▶ 选择“应用转换”为同时跨所有源。
- ▶ 单击继续。
- ▶ 在“多维展开”对话框中单击选项。

图片 15-26  
“选项”对话框



- ▶ 在“初始配置”组选项中选择定制。
- ▶ 浏览至并选择 behavior\_ini.sav 作为包含定制初始配置的文件。有关详细信息，请参阅附录 A 中的样本文件中的 IBM SPSS Categories 21。
- ▶ 选择 dim1 和 dim2 作为指定初始配置的变量。
- ▶ 单击继续。
- ▶ 在“多维展开”对话框中单击图。

图片 15-27  
“图”对话框



- ▶ 在“图”组选项选择转换图。
- ▶ 单击继续。
- ▶ 在“多维展开”对话框中单击确定。

以下是由这些选择所生成的命令语法：

```

PREFSCAL
VARIABLES=Run Talk Kiss Write Eat Sleep Mumble Read Fight Belch Argue Jump
Cry Laugh Shout
/INPUT=ROWS (ROWID )
/INITIAL=( 'samplesDirectory/behavior_ini.sav' )
dim1 dim2
/CONDITION=UNCONDITIONAL
/TRANSFORMATION=LINEAR (INTERCEPT)
/PROXIMITIES=DISSIMILARITIES

```

```

/MODEL=IDENTITY
/CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
MAXITER(5000)
/PENALTY=LAMBDA(0.5) OMEGA(1.0)
/PRINT=MEASURES COMMON
/PLOT=COMMON TRANSFORMATIONS .

```

- 此语法指定了从变量 run 到 shout 的分析。变量 rowid 用来标识行。
- INITIAL 子命令指定了起始值需从文件 behavior\_ini.sav 中获得。行坐标和列坐标堆叠在一起，列坐标跟着行坐标。
- CONDITION 子命令指定了所有的近似值可以互相比。在此分析中确实如此，因为您应该能够比较在公园里跑和在教堂里跑的近似值，并得出一种行为比另一种行为得体的结论。
- TRANSFORMATION 子命令指定了带截距的近似值的线性转换。如果近似值的一个 1 点的区别在整个 10 点的标度中是相同的，那么这是合适的。也就是说，如果学生指定他们的得分使得 0 和 1 之间的区别与 5 和 6 之间的区别相同，那么线性转换是合适的。
- PLOT 子命令请求了公共空间的图和转换图。
- 所有其他参数回退为其缺省值。

## 度量

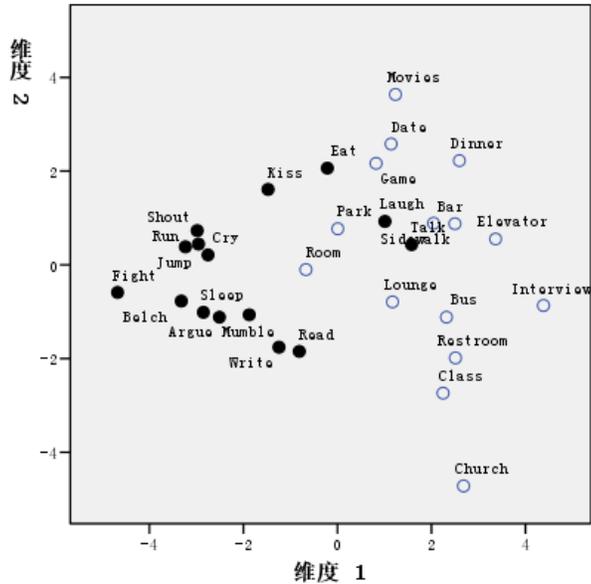
图片 15-28  
度量

迭代		169
最终的函数值		.6427725
函数值部分	应力部分	.1900001
	惩罚部分	2.1745069
拟合的不良程度	正态化的应力	.0361000
	Kruskal 的应力-I	.1900001
	Kruskal 的应力-II	.5224668
	Young 的 S-应力-I	.2760971
	Young 的 S-应力-II	.4525933
拟合优度	考虑的离散	.9639000
	考虑的方差	.8082862
	恢复的首选项顺序	.8608333
	Spearman 的 Rho	.8961120
	Kendall 的 Tau-b	.7202452
方差系数	方差近似值	.5138436
	转换的方差近似值	.4751934
	方差距离	.3912592
衰退指标	DeSarbo 的混合指标的平方和	.4957969
	Shepard 的粗略非衰退索引	.7173810

算法在 169 次迭代后收敛，最终惩罚强度为 0.6427725。变异系数和 Shepard 指标足够大，并且 DeSarbo 指标足够低，表示退化没有问题。

## 公共空间

图片 15-29  
公共空间的联合图

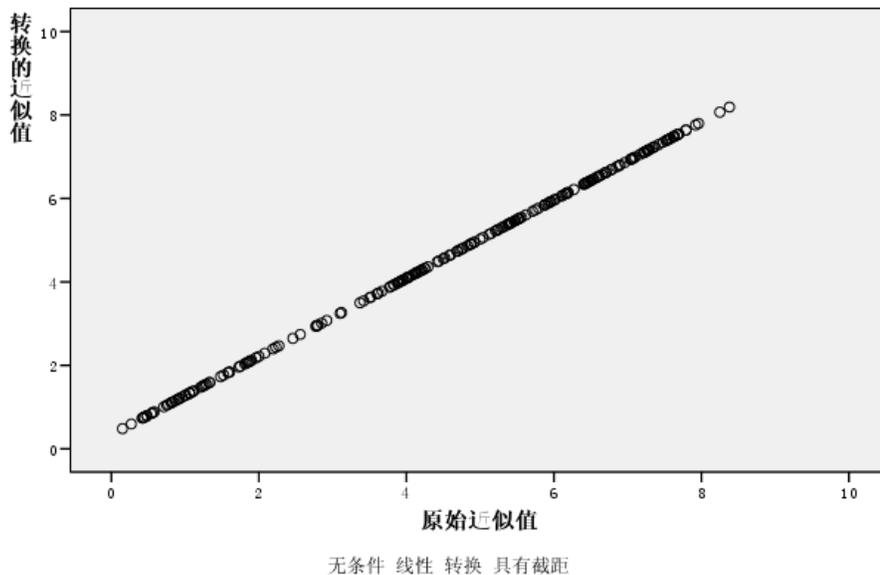


水平维度在图中与列对象（行为）的关联更紧密，且区分了不得体的行为（fighting、belching）和较得体行为。垂直维度在图中与行对象（情况）的关联更紧密，且定义了不同的情况行为限制。

- 朝着垂直维度的底部方向是将行为限制为较安静/内省型行为（read、write）的情况（church、class）。这样，这些行为被拉到垂直轴下部。
- 朝着垂直维度顶部的方向是将行为限制为社会/外向型行为（eat、kiss、laugh）的各种情况（movies、game、date）。这样，这些行为被拉到垂直轴下部。
- 在垂直维度的中间，各种情况则基于情况的常规限制性在水平维度上分隔开来。离行为较远的情况（interview）是最受限制的，而离行为较近的情况（room、park）通常受较小的限制。

## 近似值转换

图片 15-30  
变换图



因为在此分析中近似值被视为线性的，所以转换后的值和初始近似值的图形成了一条直线。此解的拟合较好，但是不同的近似值的转换可能得到更好的拟合。

## 更改近似值转换（有序）

- ▶ 要使用有序的近似值转换生成解，单击“对话框调用”工具，然后选择“多维展开”。

- ▶ 在“多维展开”对话框中单击模型。

图片 15-31  
“模型”对话框



- ▶ 选择序数作为近似值转换。
- ▶ 单击继续。
- ▶ 在“多维展开”对话框中单击确定。

以下是由这些选择所生成的命令语法：

```
PREFSCAL
VARIABLES=Run Talk Kiss Write Eat Sleep Mumble Read Fight Belch Argue Jump
Cry Laugh Shout
/INPUT=ROWS (ROWID )
/INITIAL=( 'samplesDirectory/behavior_ini.sav' )
dim1 dim2
/CONDITION=UNCONDITIONAL
/TRANSFORMATION=ORDINAL (KEEP TIES)
/PROXIMITIES=DISSIMILARITIES
/MODEL=IDENTITY
/CRITERIA=DIMENSIONS (2, 2) DIFFSTRESS (.000001) MINSTRESS (.0001)
MAXITER (5000)
/PENALTY=LAMBDA (0.5) OMEGA (1.0)
/PRINT=MEASURES COMMON
/PLOT=COMMON TRANSFORMATIONS .
```

- 唯一的变化是 TRANSFORMATION 子命令。转换设置为 ORDINAL，它保留了近似值的顺序，但不要求转换后的值与初始值成比例。

## 度量

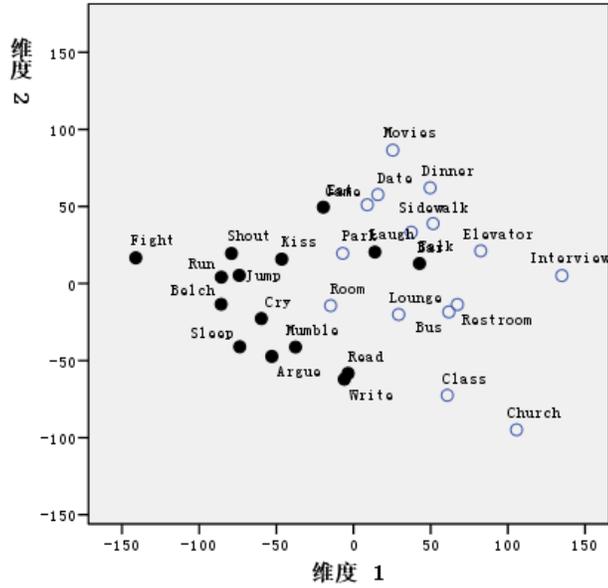
图片 15-32  
使用有序转换的解的测量

迭代		268
最终的函数值		.6044671
函数值部分	应力部分	.1747239
	惩罚部分	2.0911875
拟合的不良程度	正态化的应力	.0305285
	Kruskal 的应力-I	.1747239
	Kruskal 的应力-II	.4444641
	Young 的 S-应力-I	.2707147
	Young 的 S-应力-II	.3978003
拟合优度	考虑的离散	.9694715
	考虑的方差	.8454488
	恢复的首选项顺序	.8574206
	Spearman 的 Rho	.9032676
	Kendall 的 Tau-b	.7532788
方差系数	方差近似值	.5138436
	转换的方差近似值	.4930018
	方差距离	.4284849
衰退指标	DeSarbo 的混合指标的平方和	.3610680
	Shepard 的粗略非衰退索引	.7469048

算法在 268 次迭代后收敛，最终惩罚强度为 .6044671。此统计量和其他度量对于这个解来说比使用线性近似值转换的解稍好。

## 公共空间

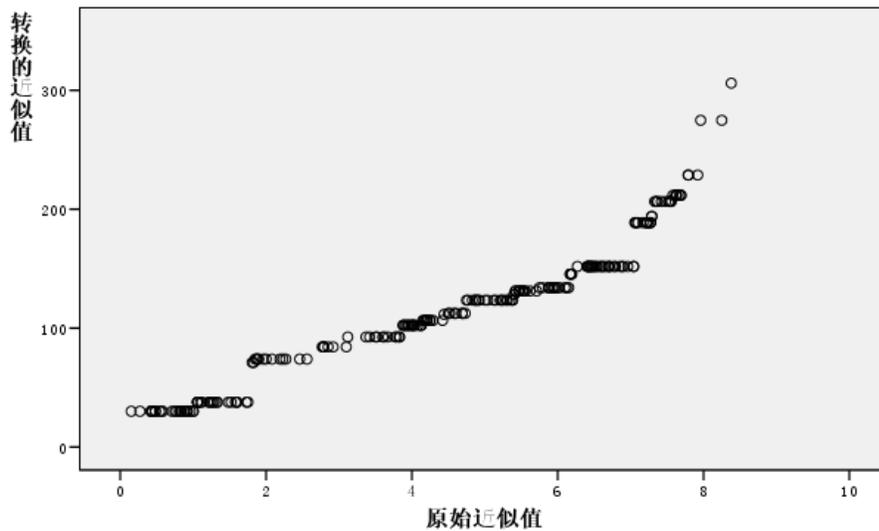
图片 15-33  
使用有序转换的解的公共空间的联合图



公共空间的解释在两个解中是一样的。可能此解（使用有序转换）在垂直维度的变异比在水平维度上的变异相对较少。而水平维度在使用线性转换的解中很明显。

## 近似值转换

图片 15-34  
使用有序转换的解的变换图



无条件 序数 转换 具有保持约束的结

除近似值最大的几个从其他值向上弯曲的值外，近似值的有序转换基本上是线性的。这些近似值可能导致有序解和线性解的大部分区别；但是，这里没有足够的信息用以决定较高的值的非线性趋势是真的趋势还是反常现象。

## 推荐参考

有关更多信息，请参见以下内容：

Busing, F. M. T. A., P. J. F. Groenen, 和 W. J. Heiser. 2005. Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. *Psychometrika*, 70, .

Green, P. E., 和 V. Rao. 1972. *Applied multidimensional scaling*. Hinsdale, Ill.: Dryden Press.

Price, R. H., 和 D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30, .

# 样本文件

随产品一起安装的样本文件可以在安装目录的 Samples 子目录中找到。对于以下每种语言在“样本”子目录中有单独的文件夹：英语、法语、德语、意大利语、日语、韩语、波兰语、俄语、简体中文、西班牙语和繁体中文。

并非所有样本文件均提供此处的全部语言版本。如果样本文件未提供某种语言的版本，则相应语言文件夹中包含该样本文件的英语版本。

## 描述

以下是对在整个文档的各种示例中使用的样本文件的简要描述。

- **accidents.sav**。该假设数据文件涉及某保险公司，该公司正在研究给定区域内汽车事故的年龄和性别风险因子。每个个案对应一个年龄类别和性别类别的交叉分类。
- **adl.sav**。该假设数据文件涉及在确定针对脑卒中患者的建议治疗类型的优点方面的举措。医师将女性脑卒中患者随机分配到两组中的一组。第一组患者接受标准的物理治疗，而第二组患者则接受附加的情绪治疗。在进行治疗的三个月时间里，将为每个患者进行一般日常生活行为的能力评分并作为原始变量。
- **advert.sav**。该假设数据文件涉及某零售商在检查广告支出与销售业绩之间的关系方面的举措。为此，他们收集了过去的销售数据以及相关的广告成本。
- **aflatoxin.sav**。该假设数据文件涉及对谷物的黄曲霉毒素的检测，该毒素的浓度会因谷物产量的不同（不同谷物之间及同种谷物之间）而有较大变化。谷物加工机从 8 个谷物产量的每一个中收到 16 个样本并以十亿分之几 (PPB) 为单位来测量黄曲霉毒素的水平。
- **anorectic.sav**。在研究厌食/暴食行为的标准症状参照时，研究人员 (Van der Ham, Meulman, Van Strien, 和 Van Engeland, 1997) 对 55 名已知存在进食障碍的青少年进行了调查。其中每名患者每年都将进行四次检查，因此总观测数为 220。在每次观测期间，将对这些患者按 16 种症状逐项评分。但 71 号和 76 号患者的症状得分均在时间点 2 缺失，47 号患者的症状得分在时间点 3 缺失，因此有效观测数为 217。
- **bankloan.sav**。该假设数据文件涉及某银行在降低贷款拖欠率方面的举措。该文件包含 850 位过去和潜在客户的财务和人口统计信息。前 700 个个案是以前曾获得贷款的客户。剩下的 150 个个案是潜在客户，银行需要按高或低信用风险对他们进行分类。
- **bankloan\_binning.sav**。该假设数据文件包含 5,000 位过去客户的财务和人口统计信息。
- **behavior.sav**。在一个经典示例中 (Price 和 Bouffard, 1974)，52 名学生被要求以 10 分的标度对 15 种情况和 15 种行为的组合进行评价，该 10 分的标度介于 0 = 平均值在个人值之上，值被视为相异性。
- **behavior\_ini.sav**。该数据文件包含 behavior.sav 的二维解的初始配置。

- **brakes.sav**。该假设数据文件涉及某生产高性能汽车盘式制动器的工厂的质量控制。该数据文件包含对 8 台专用机床中每一台的 16 个盘式制动器的直径测量。盘式制动器的目标直径为 322 毫米。
- **breakfast.sav**。在一项经典研究中 (Green 和 Rao, 1972), 21 名 Wharton School MBA 学生及其配偶被要求按照喜好程度顺序对 15 种早餐食品进行评价, 从 1 = 他们的喜好根据六种不同的情况加以记录, 从“全部喜欢”到“只带饮料的快餐”。
- **breakfast-overall.sav**。该数据文件只包含早餐食品喜好的第一种情况, 即“全部喜欢”。
- **broadband\_1.sav**。该假设数据文件包含各地区订制了全国宽带服务的客户的数量。该数据文件包含 4 年期间 85 个地区每月的订户数量。
- **broadband\_2.sav**。该数据文件和 broadband\_1.sav 一样, 但包含另外三个月的数据。
- **car\_insurance\_claims.sav**。在别处被提出和分析的 (McCullagh 和 Nelder, 1989) 关于汽车损坏赔偿的数据集。平均理赔金额可以当作其具有 gamma 分布来建模, 通过使用逆联接函数将因变量的均值与投保者年龄、车辆类型和车龄的线性组合关联。提出理赔的数量可以作为尺度权重。
- **car\_sales.sav**。该数据文件包含假设销售估计值、订价以及各种品牌和型号的车辆的物理规格。订价和物理规格可以从 edmunds.com 和制造商处获得。
- **car\_sales\_uprepared.sav**。这是 car\_sales.sav 的修改版本, 不包含字段的任何已转换版本。
- **carpet.sav**。在一个常用示例中 (Green 和 Wind, 1973), 一家公司非常重视一种新型地毯清洁用品的市场营销, 希望检验以下五种因素对消费者偏好的影响—包装设计、品牌名称、价格、优秀家用品标志和退货保证。包装设计有三个因子水平, 每个因子水平因刷体位置而不同; 有三个品牌名称 (K2R、Glory 和 Bissell); 有三个价格水平; 最后两个因素各有两个级别 (有或无)。十名消费者对这些因素所定义的 22 个特征进行了排序。变量优选包含对每个特征的平均等级的排序。低排序与高偏好相对应。此变量反映了对每个特征的偏好的总体度量。
- **carpet\_prefs.sav**。该数据文件所基于的示例和在 carpet.sav 中所描述的一样, 但它还包含从 10 位消费者的每一位中收集到的实际排列顺序。消费者被要求按照从最喜欢到最不喜欢的顺序对 22 个产品特征进行排序。carpet\_plan.sav 中定义了变量 PREF1 到 PREF22 包含相关特征的标识符。
- **catalog.sav**。该数据文件包含某编目公司出售的三种产品的假设每月销售数据。同时还包括 5 个可能的预测变量的数据。
- **catalog\_seasfac.sav**。除添加了一组从“季节性分解”过程中计算出来的季节性因子和附带的日期变量外, 该数据文件和 catalog.sav 是相同的。
- **cellular.sav**。该假设数据文件涉及某便携式电话公司在减少客户流失方面的举措。客户流失倾向分被应用到帐户, 分数范围从 0 到 100。得到 50 分或更高分数的帐户可能会更换提供商。
- **ceramics.sav**。该假设数据文件涉及某制造商在确定新型优质合金是否比标准合金具有更高的耐热性方面的举措。每个个案代表对一种合金的单独检验; 个案中会记录合金的耐热极限。
- **cereal.sav**。该假设数据文件涉及一份 880 人参与的关于早餐喜好的民意调查, 该调查记录了参与者的年龄、性别、婚姻状况以及生活方式是否积极 (根据他们是否每周至少做两次运动)。每个个案代表一个单独的调查对象。

- **clothing\_defects.sav**。这是关于某服装厂的质量控制过程的假设数据文件。检验员要对工厂中每次大批量生产的服装进行抽样检测并清点不合格的服装的数量。
- **coffee.sav**。这是关于六种冰咖啡的认知品牌形象 (Kennedy, Riquier, 和 Sharp, 1996) 的数据文件。对于 23 种冰咖啡特征属性中的每种属性, 人们选择了由该属性所描述的所有品牌。为保密起见, 六种品牌用 AA、BB、CC、DD、EE 和 FF 来表示。
- **contacts.sav**。该假设数据文件涉及一组公司计算机销售代表的联系方式列表。根据这些销售代表所在的公司部门及其公司的秩来对每个联系方式进行分类。同时还记录了最近一次的销售量、最近一次销售距今的时间和所联系公司的规模。
- **creditpromo.sav**。该假设数据文件涉及某百货公司在评价最新信用卡促销的效果方面的举措。为此, 随机选择了 500 位持卡人。其中一半收到了宣传关于在接下来的三个月内降低消费利率的广告。另一半收到了标准的季节性广告。
- **customer\_dbase.sav**。该假设数据文件涉及某公司在使用数据仓库中的信息来为最有可能回应的客户提供特惠商品方面的举措。随机选择客户群的子集并为其提供特惠商品, 同时记录下他们的回应。
- **customer\_information.sav**。该假设数据文件包含客户邮寄信息, 如姓名和地址。
- **customer\_subset.sav**。来自 customer\_dbase.sav 的拥有 80 个个案的子集。
- **debate.sav**。该假设数据文件涉及在某政治辩论前后对该辩论的参与者所做的调查的成对回答。每个个案对应一个单独的调查对象。
- **debate\_aggregate.sav**。该假设数据文件分类汇总了 debate.sav 中的回答。每个个案对应一个辩论前后的偏好的交叉分类。
- **demo.sav**。这是关于购物客户数据库的假设数据文件, 用于寄出每月的商品。将记录客户对商品是否有回应以及各种人口统计信息。
- **demo\_cs\_1.sav**。该假设数据文件涉及某公司在汇编调查信息数据库方面的举措的第一步。每个个案对应不同的城市, 并记录地区、省、区和城市标识。
- **demo\_cs\_2.sav**。该假设数据文件涉及某公司在汇编调查信息数据库方面的举措的第二步。每个个案对应来自第一步中所选城市的不同的家庭单元, 并记录地区、省、区、市、子区和单元标识。还包括设计前两个阶段的抽样信息。
- **demo\_cs.sav**。该假设数据文件包含用复杂抽样设计收集的调查信息。每个个案对应不同的家庭单元, 并记录各种人口统计和抽样信息。
- **dmdata.sav**。该假设数据文件包含直销公司的人口统计学和购买信息。dmdata2.sav 包含收到测试邮件的联系人子集的信息, dmdata3.sav 包含未收到测试邮件的其余联系人的信息。
- **dietstudy.sav**。该假设数据文件包含对 “Stillman diet” (Rickman, Mitchell, Dingman, 和 Dalen, 1974) 的研究结果。每个个案对应一个单独的主体, 并记录其在实行饮食方案前后的体重 (磅) 以及甘油三酸酯的水平 (毫克/100 毫升)。
- **dvdplayer.sav**。这是关于开发新的 DVD 播放器的假设数据文件。营销团队用原型收集了焦点小组数据。每个个案对应一个单独的被调查用户, 并记录他们的人口统计信息及其对原型问题的回答。
- **german\_credit.sav**。该数据文件取自加州大学欧文分校的 Repository of Machine Learning Databases (Blake 和 Merz, 1998) 中的 “German credit” 数据集。
- **grocery\_1month.sav**。该假设数据文件是在数据文件 grocery\_coupons.sav 的基础上加上了每周购物 “累计”, 所以每个个案对应一个单独的客户。所以, 一些每周更改的变量消失了, 而且现在记录的消费金额是为期四周的研究过程中的消费金额之和。

- **grocery\_coupons.sav**。该假设数据文件包含由重视顾客购物习惯的杂货连锁店收集的调查数据。对每位顾客调查四周，每个个案对应一个单独的顾客周，并记录有关顾客购物地点和方式的信息（包括那一周里顾客在杂货上的消费金额）。
- **guttman.sav**。Bell (Bell, 1961) 创建了一个表，用来阐释可能的社会群体。Guttman (Guttman, 1968) 引用了该表的一部分，其中包括五个变量，用于描述以下七个理论社会群体的社会交往、对群体的归属感、成员的物理亲近度以及关系正式性：观众（比如在足球比赛现场的人们）、听众（比如在剧院或听课堂讲座的人们）、公众（比如报纸或电视观众）、组织群体（与观众类似但具有紧密的关系）、初级群体（关系密切）、次级群体（自发组织）及现代社区（因在物理上亲近而导致关系松散并需要专业化服务）。
- **health\_funding.sav**。该假设数据文件包含关于保健基金（每 100 人的金额）、发病率（每 10,000 人的比率）以及保健提供商拜访率（每 10,000 的比率）的数据。每个个案代表不同的城市。
- **hivassay.sav**。该假设数据文件涉及某药物实验室在开发用于检测 HIV 感染的快速化验方面的举措。化验结果为八个加深的红色阴影，如果有更深的阴影则表示感染的可能性很大。用 2,000 份血液样本来进行实验室试验，其中一半受到 HIV 感染而另一半没有受到感染。
- **hourlywagedata.sav**。该假设数据文件涉及在政府机关和医院工作的具有不同经验水平的护士的时薪。
- **insurance\_claims.sav**。该假设数据文件涉及某保险公司，该公司希望构建一个模型用于标记可疑的、具有潜在欺骗性的理赔。每个个案代表一次单独的理赔。
- **insure.sav**。该假设数据文件涉及某保险公司，该公司正在研究指示客户是否会根据 10 年的人寿保险合同提出理赔的风险因子。数据文件中的每个个案代表一副根据年龄和性别进行匹配的合同，其中一份记录了一次理赔而另一份则没有。
- **judges.sav**。该假设数据文件涉及经过训练的裁判（加上一个体操爱好者）对 300 次体操表演给出的分数。每行代表一次单独的表演；裁判们观看相同的表演。
- **kinship\_dat.sav**。Rosenberg 和 Kim (Rosenberg 和 Kim, 1975) 开始分析 15 个亲属关系项（伯母、兄弟、表兄妹、女儿、父亲、孙女、祖父、祖母、孙子、母亲、侄子或外甥、侄女或外甥女、姐妹、儿子和叔叔）。他们让四组大学生（两组女同学，两组男同学）根据相似程度将各项排序。他们让其中的两组同学（一组女同学，一组男同学）进行了两次排序，第二次排序和第一次排序采取的标准不同。这样，一共得到六组“源”。每个源对应一个  $15 \times 15$  的相似性矩阵，其单元格中的值等于源中的人数减去此源中对象被划分的次数。
- **kinship\_ini.sav**。该数据文件包含 kinship\_dat.sav 的三维解的初始配置。
- **kinship\_var.sav**。该数据文件包含自变量 gender、gener(ation) 和 degree (of separation)，这些变量可用于解释 kinship\_dat.sav 的解的维数。具体而言，它们可用来将解的空间限制为这些变量的线性组合。
- **marketvalues.sav**。该数据文件涉及 1999 - 2000 年间 Algonquin, Ill. 地区新的房屋开发中的住房销售。这些销售仅仅来自公众记录。
- **nhis2000\_subset.sav**。美国健康访问调查 (NHIS) 是针对美国全体公民的大型人口调查。该调查对美国的具有全国代表性的家庭样本进行了面对面的访问，并获取了每个家庭的成员的健康行为和健康状态的人口统计信息和观察数据。该数据文件包含取自 2000 年调查信息的子集。国家健康统计中心。2000 年美国健康访问调查。公用数据文件和文档。

ftp://ftp.cdc.gov/pub/Health\_Statistics/NCHS/Datasets/NHIS/2000/。2003 年发布。

- **ozone.sav**。这些数据包含了用来根据其余变量预测臭氧浓度的六个气象变量的 330 个观察值。在以前的研究人员中, (Breiman 和 Friedman(F), 1985) 和 (Hastie 和 Tibshirani, 1990) 发现了这些变量之间的非线性, 这妨碍了标准回归方法。
- **pain\_medication.sav**。该假设数据文件包含用于治疗慢性关节炎疼痛的抗炎药的临床试验结果。我们感兴趣的是该药见效的时间以及它和现有药物的比较。
- **patient\_los.sav**。该假设数据文件包含被医院确诊为疑似心肌梗塞(即 MI 或“心脏病发作”)的患者的治疗记录。每个个案对应一位单独的患者, 并记录与其住院期有关的一些变量。
- **patlos\_sample.sav**。该假设数据文件包含在治疗心肌梗塞(即 MI 或“心脏病发作”)期间收到溶解血栓剂的患者样本的治疗记录。每个个案对应一位单独的患者, 并记录与其住院期有关的一些变量。
- **poll\_cs.sav**。该假设数据文件涉及民意测验专家在确定正式立法前公众对法案的支持水平方面的举措。个案对应注册的选民。每个个案记录选民居住的县、镇、区。
- **poll\_cs\_sample.sav**。该假设数据文件包含在 poll\_cs.sav 中列出的选民的样本。该样本是根据 poll\_csplan 中指定的设计来选取的, 而且该数据文件记录包含概率和样本权重。请注意, 由于该抽样计划使用与大小成正比(PPS)方法, 因此, 还有一个文件(poll\_jointprob.sav)包含联合选择概率。在选取了样本之后, 对应于选民人群统计信息及其对提交法案的意见的附加变量将被收集并添加到数据文件。
- **property\_assess.sav**。该假设数据文件涉及某县资产评估员在利用有限的资源不断更新资产价值评估方面的举措。个案对应过去一年中县里所出售的资产。数据文件中的每个个案记录资产所在的镇、最后评估资产的评估员、该次评估距今的时间、当时的估价以及资产的出售价格。
- **property\_assess\_cs.sav**。该假设数据文件涉及某州资产评估员在利用有限的资源不断更新资产价值评估方面的举措。个案对应该州的资产。数据文件中的每个个案记录资产所在的县、镇和区, 最后一次评估距今的时间以及当时的估价。
- **property\_assess\_cs\_sample.sav**。该假设数据文件包含在 property\_assess\_cs.sav 中列出的资产的样本。该样本是根据 property\_assess\_csplan 中指定的设计来选取的, 而且该数据文件记录包含概率和样本权重。在选取了样本之后, 附加变量 Current value 将被收集并添加到数据文件。
- **recidivism.sav**。该假设数据文件涉及某政府执法机构在了解其管辖区域内的屡犯率方面的举措。每个个案对应先前的一名罪犯, 并记录其人口统计信息和第一次犯罪的详细资料; 如果在第一次被捕后两年内又第二次被捕, 则还将记录两次被捕间隔的时间。
- **recidivism\_cs\_sample.sav**。该假设数据文件涉及某政府执法机构在了解其管辖区域内的屡犯率方面的举措。每个个案对应应在 2003 年 6 月期间第一次被捕释放的先前的一名罪犯, 并记录其人口统计信息和第一次犯罪的详细资料, 及其第二次被捕的数据(如果发生在 2006 年 6 月底之前)。根据 recidivism\_csplan 中指定的抽样计划从抽样部门选择罪犯; 该计划使用与大小成正比(PPS)方法, 因此, 还有一个文件(recidivism\_cs\_jointprob.sav)包含联合选择概率。
- **rfm\_transactions.sav**。此假设数据文件包含购买交易数据, 即每笔交易的购买日期、购买商品和消费金额。

- **salesperformance.sav**。这是关于评估两个新的销售培训课程的假设数据文件。60 名员工被分成 3 组且都接受标准的培训。另外，组 2 接受技术培训；组 3 接受实践教程。在培训课程结束时，对每名员工进行测验并记录他们的分数。数据文件中的每个个案代表一名单独的受训者，并记录其被分配到的组以及测验的分数。
- **satisf.sav**。该假设数据文件涉及某零售公司在 4 个商店位置所进行的满意度调查。总共对 582 位客户进行了调查，每个个案代表一位单独客户的回答。
- **screws.sav**。该数据文件包含关于螺钉、螺栓、螺母和图钉的特征的信息 (Hartigan, 1975)。
- **shampoo\_ph.sav**。这是关于某发制品厂的质量控制的假设数据文件。在规定的间隔对六批独立输出的产品进行检测并记录它们的 pH 值。目标范围是 4.5 - 5.5。
- **ships.sav**。在别处被提出和分析的 (McCullagh 等., 1989) 关于波浪对货船造成的损坏的数据集。在给定了船的类型、建造工期和服务期后，可以根据泊松比率发生来为事件计数建模。在因子交叉分类构成的表格中，每个单元格的分类汇总服务月数提供遇到风险的值。
- **site.sav**。该假设数据文件涉及某公司在为扩展业务而选择新址方面的举措。该公司聘请了两名顾问分别对选址进行评估，除了提供长期报告外，他们还要以“前景颇佳”、“前景良好”或“前景不佳”来对每个选址进行总结。
- **smokers.sav**。该数据文件摘自 1998 年全国家庭药物滥用调查并且是美国家庭的概率样本。(<http://dx.doi.org/10.3886/ICPSR02934>) 因此，分析该数据文件的第一步应该是对数据进行加权以反映总体趋势。
- **stocks.sav** 该假设数据文件包含某一年的股票价格和成交量。
- **stroke\_clean.sav**。该假设数据文件包含某医学数据库在经过“数据准备”选项中的过程清理后的状态。
- **stroke\_invalid.sav**。该假设数据文件包含某医学数据库的初始状态及一些数据输入错误。
- **stroke\_survival**。此假设数据文件涉及正在研究结束缺血性中风后复元计划的患者存活时间的研究人员面临着很多挑战。中风后，记录心肌梗塞、缺血性中风或出血性中风的发生及其时间。样本为左侧截短，因为只包含在中风后管理的复元计划结束后存活的患者。
- **stroke\_valid.sav**。该假设数据文件包含在使用“验证数据”过程检查值后，某医学数据库的状态。它仍包含潜在异常个案。
- **survey\_sample.sav**。此数据文件包含调查数据，包括人口统计学数据和各种态度测量。它基于 1998 NORC 综合社会调查的变量子集，但某些数据值已经过修改，并添加了其他虚拟变量以供演示用途。
- **telco.sav**。该假设数据文件涉及某电信公司在减少客户群中的客户流失方面的举措。每个个案对应一个单独的客户，并记录各类人口统计和服务用途信息。
- **telco\_extra.sav**。该数据文件与 telco.sav 数据文件类似，但删除了“tenure”和经对数转换的客户消费变量，代替它们的是标准化的对数转换客户消费变量。
- **telco\_missing.sav**。该数据文件是 telco.sav 数据文件的子集，但某些人口统计数据值已被缺失值替换。

- **testmarket.sav**。该假设数据文件涉及某快餐连锁店为其菜单添加新项目的计划。有三种可能的促销新产品的活动，所以会在多个随机选择的地点引入新的项目。在每个地点采用不同的促销方式，并记录新项目前四周的每周销售情况。每个个案对应单独地点的一周。
- **testmarket\_1month.sav**。该假设数据文件是在数据文件 `testmarket.sav` 的基础上加上了每周销售“累计”，所以每个个案对应一个单独的地点。所以，一些每周更改的变量消失了，而且现在记录的销售是为期四周的研究过程中的销售之和。
- **tree\_car.sav**。该假设数据文件包含人口统计和车辆购买价格数据。
- **tree\_credit.sav**。该假设数据文件包含人口统计和银行贷款历史数据。
- **tree\_missing\_data.sav**。该假设数据文件包含具有大量缺失值的人口统计和银行贷款历史数据。
- **tree\_score\_car.sav**。该假设数据文件包含人口统计和车辆购买价格数据。
- **tree\_textdata.sav**。这是一个只有两个变量的样本数据文件，主要打算在指定测量级别和值标签之前显示变量的默认状态。
- **tv-survey.sav**。该假设数据文件涉及由某电视演播室进行的一项关于是否要继续制作一档成功的节目的调查。906 位调查对象被问及他们在各种情况下是否会收看该节目。每行代表一位单独的调查对象；每列代表一种单独的情况。
- **ulcer\_recurrence.sav**。此文件包含某项研究的部分信息，该研究旨在比较两种用来防止溃疡复发的治疗的功效。它提供了区间数据的优秀示例并且已在别处被提出和分析 (Collett, 2003)。
- **ulcer\_recurrence\_recoded.sav**。该文件重新组织 `ulcer_recurrence.sav` 中的信息以允许为研究的每个区间的事件概率建模而不是简单地研究结束事件概率建模。它已在别处被提出和分析 (Collett 等., 2003)。
- **verd1985.sav**。该数据文件涉及某项调查 (Verdegaal, 1985)。该调查记录了 15 个主体对 8 个变量的响应。需要处理的变量被分成 3 个集。数据集 1 包含年龄和婚姻；数据集 2 包含宠物和新闻；数据集 3 包含音乐和居住。宠物被尺度化为多名义而年龄被尺度化为有序；所有其他变量都被尺度化为单名义。
- **virus.sav**。该假设数据文件涉及某因特网服务提供商 (ISP) 在确定病毒对其网络的影响方面的举措。他们从发现病毒到威胁得以遏制这段时间内跟踪其网络上受感染的电子邮件的流量的 (近似) 百分比。
- **wheeze\_steubenville.sav**。这是关于空气污染对儿童健康影响的纵向研究的一个子集 (Ware, Dockery, Spiro III, Speizer, 和 Ferris Jr., 1984)。这些数据包含儿童的气喘状况的重复二分类测量 (这些儿童来自 Steubenville, Ohio, 年龄为 7 到 10 岁)，以及母亲在研究的第一年中是否为吸烟者的固定记录。
- **workprog.sav**。该假设数据文件涉及一份尝试为弱势群体提供较好的工作的政府工作计划。文件后还有一个潜在计划参与者的样本，其中一些参与者是被随机选择来参加该计划的，而其他参与者则不是。每个个案代表一位单独的计划参与者。
- **worldsales.sav**。该假设数据文件包含按不同大洲和产品列出的销售收入。

# 注意事项

这些信息开发用于在全球提供的产品和服务。

IBM 可能在其他国家/地区中不提供在本文档中讨论的产品、服务或功能。请咨询您当地的 IBM 代表以了解有关您所在地区当前可用产品和服务的信息。任何对 IBM 产品、程序或服务的引用，并不意味着仅可使用这些 IBM 产品、程序或服务。作为替代，可以使用任何功能相当的产品、程序或服务，前提是不侵犯任何 IBM 知识产权。不过，用户应自行负责评估和验证任何非 IBM 产品、程序或服务的工作情况。

在本文档中介绍的主题可能涉及 IBM 的专利或申请中的专利。您获得该文档并不意味着授予您任何这些专利许可。您可以将书面许可查询函件发送至：

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785, U. S. A.

要查询双字节字符集 (DBCS) 相关许可证信息，请联系所在国家/地区中的 IBM 知识产权部门，或者以书面形式将查询函件发送至：

Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Japan.

**以下段落不适用于英国或此类条款与当地法律不符的其他国家/地区：** INTERNATIONAL BUSINESS MACHINES 公司“按原样”提供本出版物，不保证任何明示或暗示，包括但不限于对非侵权性、适销性或对特定用途适用性的暗示担保。有些州/省不允许针对某些交易的明示或暗示免责条款，因此本声明可能不适用于您。

本信息可能含有技术误差或排版错误。此处的信息会定期进行更改；这些更改会体现在本出版物的新版本中。IBM 可能随时对本出版物中所述的产品和/或程序进行改进和/或更改，恕不另行通知。

在本信息材料中对任何非 IBM 网站的引用仅为了方便用户，并不以任何方式表明对这些网站的认可。这些网站上的材料并非本 IBM 产品材料的一部分，您对这些网站的使用需自担风险。

IBM 可以自认为适当并且不会对您构成任何约束的任何方式使用或分发您提供的任何信息。

如果本程序的受许可方试图了解有关程序的信息以启用：(i) 在独立创建的程序和其他程序（包括本程序）之间交换信息；(ii) 相互使用交换的信息，则应联系：

IBM Software Group, Attention:Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.

此类信息的提供应遵照相关条款和条件，其中包括在某些情况下支付适当费用。

在本文档中介绍的受许可保护程序，及其所有受许可保护材料由 IBM 在双方签署的“IBM 客户协议”、“IBM 国际程序许可证协议”或任何其他等同协议下提供。

非 IBM 产品的相关信息来自这些产品的供应商，及其发布的公告或其他公开来源。IBM 并未测试过这些产品，无法确认有关非 IBM 产品的性能准确性、兼容性或任何其他声明。有关非 IBM 产品功能的问题应由这些产品的供应商负责。

本信息包含日常业务运营中使用的数据和报告的示例。为了尽可能详尽地对其进行说明，示例中包含了人员的姓名、公司、品牌和产品的名称。所有这些名称均为虚构，与真实商业企业使用的名称和地址的任何雷同纯属巧合。

如果您正在查阅此信息的软拷贝，照片和彩色插图可能不会显示。

## 商标

IBM、IBM 徽标、ibm.com 和 SPSS 是 IBM Corporation 的商标，在全球许多司法辖区注册。有关最新的 IBM 商标列表，请访问网页 <http://www.ibm.com/legal/copytrade.shtml>。

Adobe、Adobe 徽标、PostScript 和 PostScript 徽标是 Adobe Systems Incorporated 在美国和/或其他国家或地区的注册商标或商标。

Intel、Intel 徽标、Intel Inside、Intel Inside 徽标、Intel Centrino、Intel Centrino 徽标、Celeron、Intel Xeon、Intel SpeedStep、Itanium 和 Pentium 是 Intel Corporation 或其子公司在美国和其他国家或地区的商标或注册商标。

Java 和所有基于 Java 的商标和徽标是 Sun Microsystems, Inc. 在美国、其他国家或地区或两者的商标。

Linux 是 Linus Torvalds 在美国、其他国家或地区或两者的注册商标。

Microsoft、Windows、Windows NT 和 Windows 徽标是 Microsoft Corporation 在美国、其他国家或地区或两者的商标。

UNIX 是 The Open Group 在美国和其他国家或地区的注册商标。

此产品使用 Polar 工程咨询公司的 WinWrap Basic，版权 1993 年-2007 年，<http://www.winwrap.com>。

其他产品和服务名称可能是 IBM 或其他公司的商标。

Adobe 产品屏幕截图重印已获得 Adobe Systems Incorporated 的许可。

Microsoft 产品屏幕截图重印已获得 Microsoft Corporation 的许可。



---

# 参考书目

- Barlow, R. E., D. J. Bartholomew, D. J. Bremner, 和 H. D. Brunk. 1972. Statistical inference under order restrictions. New York: John Wiley and Sons.
- Bell, E. H. 1961. Social foundations of human behavior: Introduction to the study of sociology. New York: Harper & Row.
- Benzécri, J. P. 1969. Statistical analysis as a tool to make patterns emerge from data. 输入: Methodologies of Pattern Recognition, S. Watanabe, ed. New York: Academic Press.
- Benzécri, J. P. 1992. Correspondence analysis handbook. New York: Marcel Dekker.
- Bishop, Y. M., S. E. Feinberg, 和 P. W. Holland. 1975. Discrete multivariate analysis: Theory and Practice. Cambridge, Mass.: MIT Press.
- Blake, C. L., 和 C. J. Merz. 1998. "UCI Repository of machine learning databases." Available at <http://www.ics.uci.edu/~mllearn/MLRepository.html>.
- Breiman, L., 和 J. H. Friedman(F). 1985. Estimating optimal transformations for multiple regression and correlation. Journal of the American Statistical Association, 80, .
- Buja, A. 1990. Remarks on functional canonical variates, alternating least squares methods and ACE. Annals of Statistics, 18, .
- Busing, F. M. T. A., P. J. F. Groenen, 和 W. J. Heiser. 2005. Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. Psychometrika, 70, .
- Carroll, J. D. 1968. Generalization of canonical correlation analysis to three or more sets of variables. 输入: Proceedings of the 76th Annual Convention of the American Psychological Association, 3, Washington, D.C.: American Psychological Association.
- Collett, D. 2003. Modelling survival data in medical research, 2 ed. Boca Raton: Chapman & Hall/CRC.
- Commandeur, J. J. F., 和 W. J. Heiser. 1993. Mathematical derivations in the proximity scaling (PROXSCAL) of symmetric data matrices. Leiden: Department of Data Theory, University of Leiden.
- De Haas, M., J. A. Algera, H. F. J. M. Van Tuijl, 和 J. J. Meulman. 2000. Macro and micro goal setting: In search of coherence. Applied Psychology, 49, .
- De Leeuw, J. 1982. Nonlinear principal components analysis. 输入: COMPSTAT Proceedings in Computational Statistics, Vienna: Physica Verlag.
- De Leeuw, J. 1984. Canonical analysis of categorical data, 2nd ed. Leiden: DSWO Press.
- De Leeuw, J. 1984. The Gifi system of nonlinear multivariate analysis. 输入: Data Analysis and Informatics III, E. Diday, et al., ed..
- De Leeuw, J., 和 W. J. Heiser. 1980. Multidimensional scaling with restrictions on the configuration. 输入: Multivariate Analysis, Vol. V, P. R. Krishnaiah, ed. Amsterdam: North-Holland.

- De Leeuw, J., 和 J. Van Rijkevorsel. 1980. HOMALS and PRINCALS—Some generalizations of principal components analysis. 输入: Data Analysis and Informatics, E. Diday, et al., ed. Amsterdam: North-Holland.
- De Leeuw, J., F. W. Young, 和 Y. Takane. 1976. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, 41, .
- De Leeuw, J. 1990. Multivariate analysis with optimal scaling. 输入: Progress in Multivariate Analysis, S. Das Gupta, 和 J. Sethuraman, eds. Calcutta: Indian Statistical Institute.
- Eckart, C., 和 G. Young. 1936. The approximation of one matrix by another one of lower rank. *Psychometrika*, 1, .
- Fisher, R. A. 1938. Statistical methods for research workers. Edinburgh: Oliver and Boyd.
- Fisher, R. A. 1940. The precision of discriminant functions. *Annals of Eugenics*, 10, .
- Gabriel(G), K. R. 1971. The biplot graphic display of matrices with application to principal components analysis. *Biometrika*, 58, .
- Gifi, A. 1985. PRINCALS. Research Report UG-85-02. Leiden: Department of Data Theory, University of Leiden.
- Gifi, A. 1990. Nonlinear multivariate analysis. Chichester: John Wiley and Sons.
- Gilula, Z., 和 S. J. Haberman. 1988. The analysis of multivariate contingency tables by restricted canonical and restricted association models. *Journal of the American Statistical Association*, 83, .
- Gower, J. C., 和 J. J. Meulman. 1993. The treatment of categorical information in physical anthropology. *International Journal of Anthropology*, 8, .
- Green, P. E., 和 V. Rao. 1972. Applied multidimensional scaling. Hinsdale, Ill.: Dryden Press.
- Green, P. E., 和 Y. Wind. 1973. Multiattribute decisions in marketing: A measurement approach. Hinsdale, Ill.: Dryden Press.
- Guttman, L. 1941. The quantification of a class of attributes: A theory and method of scale construction. 输入: The Prediction of Personal Adjustment, P. Horst, ed. New York: Social Science Research Council.
- Guttman, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points. *Psychometrika*, 33, .
- Hartigan, J. A. 1975. Clustering algorithms. New York: John Wiley and Sons.
- Hastie, T., 和 R. Tibshirani. 1990. Generalized additive models. London: Chapman and Hall.
- Hastie, T., R. Tibshirani, 和 A. Buja. 1994. Flexible discriminant analysis. *Journal of the American Statistical Association*, 89, .
- Hayashi, C. 1952. On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics*, 2, .

- Heiser, W. J. 1981. Unfolding analysis of proximity data. Leiden: Department of Data Theory, University of Leiden.
- Heiser, W. J., 和 F. M. T. A. Busing. 2004. Multidimensional scaling and unfolding of symmetric and asymmetric proximity relations. 输入: Handbook of Quantitative Methodology for the Social Sciences, D. Kaplan, ed. Thousand Oaks, Calif.: Sage Publications, Inc..
- Heiser, W. J., 和 J. J. Meulman. 1994. Homogeneity analysis: Exploring the distribution of variables and their nonlinear relationships. 输入: Correspondence Analysis in the Social Sciences: Recent Developments and Applications, M. Greenacre, 和 J. Blasius, eds. New York: Academic Press.
- Heiser, W. J., 和 J. J. Meulman. 1995. Nonlinear methods for the analysis of homogeneity and heterogeneity. 输入: Recent Advances in Descriptive Multivariate Analysis, W. J. Krzanowski, ed. Oxford: Oxford University Press.
- Horst, P. 1961. Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology*, 17, .
- Horst, P. 1961. Relations among m sets of measures. *Psychometrika*, 26, .
- Israëls, A. 1987. Eigenvalue techniques for qualitative data. Leiden: DSWO Press.
- Kennedy, R., C. Riquier, 和 B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research. *Journal of Targeting, Measurement, and Analysis for Marketing*, 5, .
- Kettenring, J. R. 1971. Canonical analysis of several sets of variables. *Biometrika*, 58, .
- Kruskal, J. B. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, .
- Kruskal, J. B. 1964. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29, .
- Kruskal, J. B. 1965. Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society Series B*, 27, .
- Kruskal, J. B. 1978. Factor analysis and principal components analysis: Bilinear methods. 输入: International Encyclopedia of Statistics, W. H. Kruskal, 和 J. M. Tanur, eds. New York: The Free Press.
- Kruskal, J. B., 和 R. N. Shepard. 1974. A nonmetric variety of linear factor analysis. *Psychometrika*, 39, .
- Krzanowski, W. J., 和 F. H. C. Marriott. 1994. Multivariate analysis: Part I, distributions, ordination and inference. London: Edward Arnold.
- Lebart, L., A. Morineau, 和 K. M. Warwick. 1984. Multivariate descriptive statistical analysis. New York: John Wiley and Sons.
- Lingoes, J. C. 1968. The multivariate analysis of qualitative data. *Multivariate Behavioral Research*, 3, .
- McCullagh, P., 和 J. A. Nelder. 1989. Generalized Linear Models, 2nd ed. London: Chapman & Hall.

- Meulman, J. J. 1982. Homogeneity analysis of incomplete data. Leiden: DSWO Press.
- Meulman, J. J. 1986. A distance approach to nonlinear multivariate analysis. Leiden: DSWO Press.
- Meulman, J. J. 1992. The integration of multidimensional scaling and multivariate analysis with optimal transformations of the variables. *Psychometrika*, 57, .
- Meulman, J. J. 1993. Principal coordinates analysis with optimal transformations of the variables:Minimizing the sum of squares of the smallest eigenvalues. *British Journal of Mathematical and Statistical Psychology*, 46, .
- Meulman, J. J. 1996. Fitting a distance model to homogeneous subsets of variables:Points of view analysis of categorical data. *Journal of Classification*, 13, .
- Meulman, J. J. 2003. Prediction and classification in nonlinear data analysis:Something old, something new, something borrowed, something blue. *Psychometrika*, 4, .
- Meulman, J. J., 和 W. J. Heiser. 1997. Graphical display of interaction in multiway contingency tables by use of homogeneity analysis. 输入: *Visual Display of Categorical Data*, M. Greenacre, 和 J. Blasius, eds. New York: Academic Press.
- Meulman, J. J., 和 P. Verboon. 1993. Points of view analysis revisited:Fitting multidimensional structures to optimal distance components with cluster restrictions on the variables. *Psychometrika*, 58, .
- Meulman, J. J., A. J. Van der Kooij, 和 A. Babinec. 2000. New features of categorical principal components analysis for complicated data sets, including data mining. 输入: *Classification, Automation and New Media*, W. Gaul, 和 G. Ritter, eds. Berlin: Springer-Verlag.
- Meulman, J. J., A. J. Van der Kooij, 和 W. J. Heiser. 2004. Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. 输入: *Handbook of Quantitative Methodology for the Social Sciences*, D. Kaplan, ed. Thousand Oaks, Calif.: Sage Publications, Inc..
- Nishisato, S. 1980. Analysis of categorical data:Dual scaling and its applications. Toronto: University of Toronto Press.
- Nishisato, S. 1984. Forced classification:A simple application of a quantification method. *Psychometrika*, 49, .
- Nishisato, S. 1994. Elements of dual scaling:An introduction to practical data analysis. Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc.
- Pratt, J. W. 1987. Dividing the indivisible:Using simple symmetry to partition variance explained. 输入: *Proceedings of the Second International Conference in Statistics*, T. Pukkila, 和 S. Puntanen, eds. Tampere, Finland: University of Tampere.
- Price, R. H., 和 D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30, .

- Ramsay, J. O. 1989. Monotone regression splines in action. *Statistical Science*, 4, .
- Rao, C. R. 1973. *Linear statistical inference and its applications*, 2nd ed. New York: John Wiley and Sons.
- Rao, C. R. 1980. Matrix approximations and reduction of dimensionality in multivariate statistical analysis. 输入: *Multivariate Analysis*, Vol. 5, P. R. Krishnaiah, ed. Amsterdam: North-Holland.
- Rickman, R., N. Mitchell, J. Dingman, 和 J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. *Journal of the American Medical Association*, 228, .
- Rosenberg, S., 和 M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research*, 10, .
- Roskam, E. E. 1968. Metric analysis of ordinal data in psychology. Voorschoten: VAM.
- Shepard, R. N. 1962. The analysis of proximities:Multidimensional scaling with an unknown distance function I. *Psychometrika*, 27, .
- Shepard, R. N. 1962. The analysis of proximities:Multidimensional scaling with an unknown distance function II. *Psychometrika*, 27, .
- Shepard, R. N. 1966. Metric structures in ordinal data. *Journal of Mathematical Psychology*, 3, .
- Tenenhaus, M., 和 F. W. Young. 1985. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis, and other methods for quantifying categorical multivariate data. *Psychometrika*, 50, .
- Theunissen, N. C. M., J. J. Meulman, A. L. Den Ouden, H. M. Koopman, G. H. Verrips, S. P. Verloove-Vanhorick, 和 J. M. Wit. 2003. Changes can be studied when the measurement instrument is different at different time points. *Health Services and Outcomes Research Methodology*, 4, .
- Tucker, L. R. 1960. Intra-individual and inter-individual multidimensionality. 输入: *Psychological Scaling:Theory & Applications*, H. Gulliksen, 和 S. Messick, eds. New York: John Wiley and Sons.
- Van der Burg, E. 1988. *Nonlinear canonical correlation and some related techniques*. Leiden: DSWO Press.
- Van der Burg, E., 和 J. De Leeuw. 1983. Nonlinear canonical correlation. *British Journal of Mathematical and Statistical Psychology*, 36, .
- Van der Burg, E., J. De Leeuw, 和 R. Verdegaal. 1988. Homogeneity analysis with k sets of variables:An alternating least squares method with optimal scaling features. *Psychometrika*, 53, .
- Van der Ham, T., J. J. Meulman, D. C. Van Strien, 和 H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents:A longitudinal perspective. *British Journal of Psychiatry*, 170, .
- Van der Kooij, A. J., 和 J. J. Meulman. 1997. MURALS:Multiple regression and optimal scaling using alternating least squares. 输入: *Softstat ' 97*, F. Faulbaum, 和 W. Bandilla, eds. Stuttgart: Gustav Fisher.

- Van Rijckevorsel, J. 1987. The application of fuzzy coding and horseshoes in multiple correspondence analysis. Leiden: DSWO Press.
- Verboon, P., 和 I. A. Van der Lans. 1994. Robust canonical discriminant analysis. *Psychometrika*, 59, .
- Verdegaal, R. 1985. Meer sets analyse voor kwalitatieve gegevens (in Dutch). Leiden: Department of Data Theory, University of Leiden.
- Vlek, C., 和 P. J. Stallen. 1981. Judging risks and benefits in the small and in the large. *Organizational Behavior and Human Performance*, 28, .
- Wagenaar, W. A. 1988. Paradoxes of gambling behaviour. London: Lawrence Erlbaum Associates, Inc.
- Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, 和 B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. *American Review of Respiratory Diseases*, 129, .
- Winsberg, S., 和 J. O. Ramsay. 1980. Monotonic transformations to additivity using splines. *Biometrika*, 67, .
- Winsberg, S., 和 J. O. Ramsay. 1983. Monotone spline transformations for dimension reduction. *Psychometrika*, 48, .
- Wolter, K. M. 1985. Introduction to variance estimation. Berlin: Springer-Verlag.
- Young, F. W. 1981. Quantitative analysis of qualitative data. *Psychometrika*, 46, .
- Young, F. W., J. De Leeuw, 和 Y. Takane. 1976. Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 41, .
- Young, F. W., Y. Takane, 和 J. De Leeuw. 1978. The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika*, 43, .
- Zeijl, E., Y. te Poel, M. du Bois-Reymond, J. Ravesloot, 和 J. J. Meulman. 2000. The role of parents and peers in the leisure activities of young adolescents. *Journal of Leisure Research*, 32, .
- 最大值, J. 1960. Quantizing for minimum distortion. *Proceedings IEEE (Information Theory)*, 6, .

- ANOVA  
在“分类回归”中, 19
- Cronbach 的 alpha  
在“分类主成份分析”中, 127
- DeSarbo 混合度指标  
在“多维展开”中, 242, 245, 251, 257, 265
- PREFSCAL, 70
- $R^2$   
在“分类回归”中, 92
- Ridge 回归  
在“分类回归”中, 18
- Shepard 图  
在“多维展开”中, 75
- Shepard 粗略非退化指标  
在“多维展开”中, 242, 245, 251, 257, 265
- 三标图  
在“分类主成份分析”中, 31
- 三维展开  
在“多维展开”中, 246
- 不严格的更新  
在多维尺度中, 65
- 互相关性  
在“分类回归”中, 91
- 公共空间  
在多维尺度中, 233, 237  
在“多维展开”中, 243, 246, 252, 258, 266, 270
- 公共空间图  
在多维尺度中, 66  
在“多维展开”中, 75
- 公共空间坐标  
在多维尺度中, 67  
在“多维展开”中, 77
- 公共空间的约束  
在“多维展开”中, 72
- 公共空间的联合图  
在“多维展开”中, 243, 246, 252, 258, 266, 270
- 分类主成份分析, 23, 27, 122, 134  
保存变量, 30  
“分类主成份分析”中的, 153  
命令附加功能, 34  
定量, 128, 147  
对象得分, 129, 133, 151  
成份载荷, 131, 134, 149  
最优尺度级别, 25  
模型摘要, 127, 133, 149  
迭代历史记录, 127  
“分类主成份分析”中的  
在“分类主成份分析”中, 153
- 分类回归, 12, 80  
互相关性, 91  
保存, 20  
变换图, 94  
命令附加功能, 22  
图, 12  
最优尺度级别, 13  
模型拟合, 92  
残差, 95  
相关, 92 - 93  
统计量, 12  
规则化, 18  
重要性, 93  
“分类回归”中的  
在“分类回归”中, 93
- 初始公共空间图  
在“多维展开”中, 75
- 初始配置  
在“分类回归”中, 16  
在多维尺度中, 65  
在“多维展开”中, 73  
在“非线性典型相关性分析”中, 38
- 加权欧几里得模型  
在“多维展开”中, 71
- 区分测量  
在“多重对应分析”中, 52, 211
- 区分测量图  
在“多重对应分析”中, 55
- 双标图  
在“分类主成份分析”中, 31  
在“多重对应分析”中, 54  
在“对应分析”中, 45
- 变异系数  
在“多维展开”中, 242, 245, 251, 257, 265
- 变换图  
在“分类主成份分析”中, 32  
在“分类回归”中, 94  
在多维尺度中, 66, 236  
在“多维展开”中, 75, 267, 270  
在“多重对应分析”中, 55
- 变量权重  
在“分类主成份分析”中, 25  
在“多重对应分析”中, 49
- 商标, 280
- 回归系数  
在“分类回归”中, 19
- 图  
在“分类回归”中, 21

## 索引

- 在多维尺度中, 66 - 67
- 在“对应分析”中, 45
- 在“非线性典型相关性分析”中, 38
- 复相关系数
  - 在“分类回归”中, 19
- 多维尺度, 57, 59 - 62, 221
  - 公共空间, 233, 237
  - 变换图, 236
  - 命令附加功能, 69
  - 图, 57, 66 - 67
  - 应力度量标准, 232, 237
  - 模型, 63
  - 约束, 64
  - 统计量, 57
  - 输出, 67
  - 选项, 65
- 多维展开, 70, 240, 260
  - 三维展开, 246
  - 公共空间, 243, 246, 252, 258, 266, 270
  - 公共空间的约束, 72
  - 命令附加功能, 78
  - 图, 70, 75
  - 度量, 242, 245, 251, 257, 265, 269
  - 模型, 71
  - 私有空间, 253, 258
  - 统计量, 70
  - 输出, 77
  - 近似值转换, 267, 270
  - 退化解, 240
  - 选项, 73
- 多起点图
  - 在“多维展开”中, 75
- 多重对应分析, 47, 51, 206
  - 保存变量, 53
  - 区分测量, 211
  - 命令附加功能, 56
  - 对象得分, 210, 213
  - 最优尺度级别, 49
  - 模型摘要, 209
  - 离群值, 215
  - 类别量化, 212
- 套索
  - 在“分类回归”中, 18
- 定量
  - 在“分类主成份分析”中, 128, 147
  - 在“非线性典型相关性分析”中, 182
- 对应分析, 40 - 42, 44 - 45, 195 - 196
  - 命令附加功能, 46
  - 图, 40
  - 对应分析”中的, 202
  - “对应分析”中的, 202
- 标准化, 195
- 统计量, 40
- 维度, 200
- 贡献, 201
- 对应分析”中的
  - 在“对应分析”中, 202
  - “对应分析”中的
    - 在“对应分析”中, 195, 202
- 对象得分
  - 在“分类主成份分析”中, 29, 129, 133, 151
  - 在“多重对应分析”中, 52, 210, 213
  - 在“非线性典型相关性分析”中, 38
- 对象点图
  - 在“分类主成份分析”中, 31
  - 在“多重对应分析”中, 54
- 尺度模型
  - 在“多维展开”中, 71
- 广义欧几里得模型
  - 在“多维展开”中, 71
- 应力图
  - 在多维尺度中, 66
  - 在“多维展开”中, 75
- 应力度量标准
  - 在多维尺度中, 67, 232, 237
  - 在“多维展开”中, 77
- 弹性网络
  - 在“分类回归”中, 18
- 恒等模型
  - 在“多维展开”中, 71
- 惩罚强度
  - 在“多维展开”中, 242, 251, 257, 265, 269
- 惩罚项
  - 在“多维展开”中, 73
- 惯量
  - 在“对应分析”中, 44
- 成份载荷
  - 在“分类主成份分析”中, 29, 131, 134, 149
  - 在“非线性典型相关性分析”中, 38, 181
- 成份载荷图
  - 在“分类主成份分析”中, 33
- 投影质心
  - 在“非线性典型相关性分析”中, 185
- 投影质心图
  - 在“分类主成份分析”中, 32
- 拟合
  - 在“非线性典型相关性分析”中, 38

- 拟合值  
在“非线性典型相关性分析”中, 178
- 拟合散点图  
在“多维展开”中, 75
- 损失值  
在“非线性典型相关性分析”中, 178
- 描述统计  
在“分类回归”中, 19
- 方差解释情况  
在“分类主成份分析”中, 29, 127, 149
- 最优尺度级别  
在“分类主成份分析”中, 25  
在“多重对应分析”中, 49
- 最终公共空间图  
在“多维展开”中, 75
- 权重  
在“非线性典型相关性分析”中, 38, 179
- 标准化  
在“对应分析”中, 42, 195
- 样本文件  
位置, 272
- 模型摘要  
在“多重对应分析”中, 209
- 残差  
在“分类回归”中, 95
- 残差图  
在“多维展开”中, 75
- 法律注意事项, 279
- 特征值  
在“分类主成份分析”中, 127, 133, 149  
在“非线性典型相关性分析”中, 178
- 相关  
在多维尺度中, 67
- 相关性图  
在多维尺度中, 66
- 相关矩阵  
在“分类主成份分析”中, 29  
在“多重对应分析”中, 52
- 离散化  
在“分类主成份分析”中, 26  
在“分类回归”中, 14  
在“多重对应分析”中, 49
- 离群值  
在“多重对应分析”中, 215
- 私有空间  
在“多维展开”中, 253, 258
- 私有空间图  
在多维尺度中, 66  
在“多维展开”中, 75
- 私有空间坐标  
在“多维展开”中, 77
- 私有空间权重  
在多维尺度中, 67  
在“多维展开”中, 77
- 私有空间权重图  
在多维尺度中, 66  
在“多维展开”中, 75
- 私有空间的联合图  
在“多维展开”中, 253, 258
- 空间权重图  
在“多维展开”中, 75
- 类别图  
在“分类主成份分析”中, 32  
在“多重对应分析”中, 55
- 类别坐标  
在“非线性典型相关性分析”中, 185
- 类别量化  
在“分类主成份分析”中, 29  
在“分类回归”中, 19  
在“多重对应分析”中, 52, 212  
在“非线性典型相关性分析”中, 38
- 系数  
在“分类回归”中, 92
- 约束  
在多维尺度中, 64
- 维度  
在“对应分析”中, 42, 200
- 维度权重  
在“多维展开”中, 253, 258
- 缺失值  
在“分类主成份分析”中, 26  
在“分类回归”中, 15  
在“多重对应分析”中, 50
- 置信统计量  
在“对应分析”中, 44
- 联合类别图  
在“分类主成份分析”中, 32  
在“多重对应分析”中, 55
- 补充对象  
在“分类回归”中, 16

## 索引

## 贡献

在“对应分析”中, 201

## 质心

在“非线性典型相关性分析”中, 38, 185

## 距离

在多维尺度中, 67

在“多维展开”中, 77

## 距离测量

在“对应分析”中, 42

## 转换后的自变量

在多维尺度中, 67

## 转换后的近似值

在多维尺度中, 67

在“多维展开”中, 77

## 近似值转换

在“多维展开”中, 71

## 迭代历史记录

在“分类主成份分析”中, 29, 127

在多维尺度中, 67

在“多维展开”中, 77

在“多重对应分析”中, 52

## 迭代标准

在多维尺度中, 65

在“多维展开”中, 73

## 重要性

在“分类回归”中, 93

## 零阶相关

在“分类回归”中, 93

## 非线性典型相关性分析, 35, 37 - 38, 171

分析摘要, 178

命令附加功能, 39

图, 35

定量, 182

成份载荷, 179, 181

权重, 179

类别坐标, 185

统计量, 35

质心, 185