版本 17

多元方法

“真正的发现之旅，并不在于寻求新的景观，
而在于拥有新的眼光。”

Marcel Proust

JMP Statistical Discovery LLC
SAS Campus Drive
Cary, North Carolina 27513-2414

17.1

The correct bibliographic citation for this manual is as follows: JMP Statistical Discovery LLC 2022–2023. *JMP® 17 Multivariate Methods*. Cary, NC: JMP Statistical Discovery LLC

JMP® 17 Multivariate Methods

JMP Statistical Discovery LLC, SAS Campus Drive, Cary, North Carolina 27513-2414.

October 2022

March 2023

充分利用 JMP

无论您是新手还是经验丰富的用户，都有需要了解的 JMP 知识。

访问 JMP.com 获取以下资源：

- JMP 入门知识网络讲座直播和录像
- 新功能和高级技巧视频演示和网络讲座
- 关于注册参加 JMP 培训的详细信息
- 您所在地区举办的研讨会安排
- 其他人使用 JMP 的成功案例
- JMP 用户社区、提供给用户的资源（包括插件和脚本示例）、论坛、博客、会议信息等

jmp.com/getstarted

目录
多元方法

第 1 章

了解 JMP

文档和其他资源

了解 JMP 文档，例如手册规范、每个 JMP 文档的说明、“帮助”系统以及哪里可以查找到其他的支持。

目录

JMP 文档中的格式规范

以下规范有助于将书面材料与您在屏幕上看到的信息相联系：

- 样本数据表名称、列名、路径名称、文件名、文件扩展名和文件夹采用 Helvetica（或无衬线联机）字体显示。
- 代码采用 `Lucida Sans Typewriter`（或固定宽度联机）字体显示。
- 代码输出采用 *`Lucida Sans Typewriter`* 斜体（或固定宽度斜体联机）字体显示，并且相对于之前的代码缩进显示。
- Helvetica **粗体**格式（或粗体无衬线联机）表示为了完成某个任务而选择的项：
  - 按钮
  - 复选框
  - 命令
  - 可供选择的列表名称
  - 菜单
  - 选项
  - 选项卡名称
  - 文本框
- 下列项采用**黑体**字体显示：
  - 重要的或具有特定 JMP 定义的字词
  - 变量
- 仅适用于 JMP Pro 的功能使用 JMP Pro 图标 JMP PRO 加以注释。对于 JMP Pro 功能概述，请访问 jmp.com/software/pro。

注意：特殊信息和适用的局限性将在“注意”中显示。

提示：实用信息将在“提示”中显示。

JMP 帮助

使用 “帮助”菜单中的 “JMP 帮助”，您可以搜索有关 JMP 功能、统计方法以及 “JMP 脚本语言”（或 JSL）的信息。打开 “JMP 帮助”有以下几种方式：

- 在 Windows 系统中，通过选择**帮助** > **JMP 帮助**搜索和查看“JMP 帮助”。
- 在 Windows 系统中，按 F1 键可在默认浏览器中打开“帮助”系统。
- 获取有关数据表或报表窗口特定部分的帮助。从**工具**菜单选择“帮助”工具  ，然后点击数据表或报表窗口中的任意位置，查看该部分的帮助。
- 在 JMP 窗口中，点击**帮助**按钮。

注意：“JMP 帮助”可供具有 Internet 连接的用户使用。没有 Internet 连接的用户可以通过选择**帮助** > **JMP 文档库**搜索 PDF 文件格式的所有手册。详细信息，请参见“JMP 文档库”。

JMP 文档库

“帮助”系统内容也在一个称为 JMP 文档库的 PDF 文件中提供。选择**帮助** > **JMP 文档库**可以打开该文件。若您更喜欢搜索 JMP 库中对应于每个文档的单个 PDF 文件，也可以下载“文档 PDF 文件”插件。从 community.jmp.com 下载可用的插件。

下表说明了 JMP 库中每个文档的用途和内容。

文档标题	文档用途	文档内容
发现 JMP	若您不熟悉 JMP，则从该文档开始学习。	介绍 JMP 以及指导如何创建和分析数据。还可学习如何共享结果。
使用 JMP	了解 JMP 数据表以及如何执行基本操作。	涵盖了贯穿整个 JMP 的常规 JMP 概念和功能，包括导入数据、修改列属性、数据排序以及连接至 SAS。

文档标题	文档用途	文档内容
基本分析	使用该文档执行基本分析。	介绍以下“分析”菜单平台： • 分布 • 以 X 拟合 Y • 制表 • 文本分析器 涵盖如何通过“分析”>“以 X 拟合 Y”执行二元分析、单因子方差分析和列联分析。还包括如何使用 Bootstrapping 来近似估计抽样分布以及如何使用“模拟”平台执行参数再抽样。
基本绘图	为您的数据找到最理想的图形展示方法。	介绍以下“图形”菜单平台： • 图形生成器 • 三维散点图 • 等高线图 • 气泡图 • 平行图 • 方格图 • 散点图矩阵 • 三元图 • 矩形树图 • 图表 • 叠加图 手册还涵盖如何创建背景地图和定制地图。
刻画器指南	学习如何使用交互式刻画工具，它们可以让您查看任何响应曲面的横截面。	涵盖“图形”菜单中列出的所有刻画器。还包括了对噪声因子的分析以及使用随机输入运行模拟。
实验设计指南	学习如何设计实验和确定适当的样本大小。	涵盖“实验设计”菜单中的所有主题。

文档标题	文档用途	文档内容
拟合线性模型	学习“拟合模型”平台及其许多特质。	介绍以下特质，它们都位于“分析”菜单“拟合模型”平台中： • 标准最小二乘法 • 逐步 • 广义回归 • 混合模型 • 广义线性混合模型 • 多元方差分析 • 对数线性方差 • 名义型 Logistic • 有序型 Logistic • 广义线性模型

文档标题	文档用途	文档内容
预测和专业建模	学习其他建模技巧。	介绍以下“分析”>“预测建模”菜单平台： • 神经 • 分割 • Bootstrap 森林法 • 提升树 • K 最近邻 • 朴素 Bayes • 支持向量机 • 模型比较 • 模型筛选 • 生成验证列 • 公式存储库 介绍以下“分析”>“专业建模”菜单平台： • 拟合曲线 • 非线性 • 函数数据分析器 • 高斯过程 • 时间序列 • 配对 介绍以下“分析”>“筛选”菜单平台： • 探索离群值 • 探索缺失值 • 探索模式 • 响应筛选 • 预测变量筛选 • 关联分析 • 过程历史分析器

文档标题	文档用途	文档内容
多元方法	获悉同时分析多个变量的技巧。	介绍以下“分析”>“多元方法”菜单平台： • 多元 • 主成分 • 判别 • 偏最小二乘 • 多重对应分析 • 结构化方程模型 • 因子分析 • 多维尺度化 • 多元嵌入 • 项目分析 介绍以下“分析”>“聚类”菜单平台： • 层次聚类 • K 均值聚类 • 正态混合 • 潜在类分析 • 聚类变量
质量和过程方法	了解用于评估和改进过程的工具。	介绍以下“分析”>“质量和过程”菜单平台： • 控制图生成器和单个控制图 • 测量系统分析（EMP 和 1 型量具） • 变异性/计数量具图 • 过程筛选 • 过程能力 • 模型驱动的多元控制图 • 传统控制图 • Pareto 图 • 关系图 • 管理限值 • OC 曲线

文档标题	文档用途	文档内容
可靠性和生存方法	学习评估和改进产品或系统的可靠性以及分析人或产品的生存数据。	介绍以下“分析”>“可靠性和生存”菜单平台： • 寿命分布 • 以 X 拟合寿命 • 累积损坏 • 复发分析 • 退化 • 重复测量退化 • 破坏性退化 • 可靠性预测 • 可靠性增长 • 可靠性框图 • 可修复系统模拟 • 生存 • 拟合参数生存 • 拟合比例风险
消费者研究	学习研究消费者偏好的方法以及如何通过获得的洞察力创造更好的产品和服务。	介绍以下“分析”>“消费者研究”菜单平台： • 分类 • 选择 • MaxDiff • 提升 • 多重因子分析
遗传学	了解 JMP 中提供的方法，帮助您分析遗传数据，并使用该数据模拟育种计划，以预测要进行的最优遗传杂交。	说明以下“分析”>“遗传学”菜单平台： • 标记统计量 • 标记模拟
Scripting Guide	学习如何使用功能强大的 JMP 脚本语言 (JSL)。	涵盖多方面主题，例如编写和调试脚本、操作数据表、构造显示框以及创建 JMP 应用程序。

文档标题	文档用途	文档内容
JSL Syntax Reference	获悉许多 JSL 函数及其变元和发送至对象和显示框的消息。	包括 JSL 命令的语法、示例和注释。

学习 JMP 的其他资源

除了阅读 JMP 帮助外，还可以使用以下资源学习 JMP：

- “搜索 JMP”
- “JMP 教程”
- “样本数据表”
- “学习统计和 JSL 术语”
- “学习 JMP 的提示与技巧”
- “JMP 工具提示”
- “JMP 用户社区”
- “免费在线统计思维课程”
- “JMP 新用户欢迎套件”
- “统计学知识门户”
- “JMP 培训”
- “用户编写的 JMP 手册”
- ““JMP 起始页”窗口”

搜索 JMP

若您不确定在哪里查找统计步骤，可以在 JMP 中进行搜索。结果会根据启动搜索的窗口进行调整，例如数据表或报表。

1. 点击**帮助** > **搜索 JMP**。或者，按 Ctrl+ 逗号。
2. 输入您的搜索文本。
3. 点击包含所需过程的结果。

   在右侧，您可以看到步骤的说明和位置。
4. 点击相应的按钮可打开或导航至结果。

JMP 教程

可以选择**帮助** > **教程**来访问 JMP 教程。

若您不熟悉 JMP，则从**初学者教程**开始。它分步介绍了 JMP 界面并解释了使用 JMP 的基本操作。

其余教程有助于您了解 JMP 的特定方面，例如设计实验以及将样本均值与常数比较。

样本数据表

JMP 文档系列中的所有示例使用的都是样本数据。选择**帮助 > 样本数据文件夹**以打开样本数据目录。

要查看按字母顺序列出的样本数据表或查看不同分类下的样本数据，选择**帮助 > 样本索引**。

样本数据表安装在以下目录：

在 Windows 上：C:\Program Files\SAS\JMP\17\Samples\Data

在 macOS 上：\Library\Application Support\JMP\17\Samples\Data

在 JMP Pro 中，样本数据安装在 JMPPRO（而不是 JMP）目录中。

要查看使用样本数据的示例，选择**帮助 > 样本索引**并导航到“教学资源”部分。

学习统计和 JSL 术语

有关统计术语的帮助，请选择 “帮助 > 统计索引”。有关 JSL 脚本和示例的帮助，请选择**帮助 > 脚本索引**。

统计索引 提供统计术语定义。

脚本索引 使您可以搜索有关 JSL 函数、对象和显示框的信息。您还可以从 “脚本索引”编辑和运行样本脚本以及获取有关命令的帮助。

学习 JMP 的提示与技巧

首次启动 JMP 时会看到“今日提示”窗口。该窗口提供使用 JMP 的一些小技巧。

要关闭“今日提示”，清除**启动时显示提示**复选框。要再次查看它，请选择**帮助 > 今日提示**。或者可以使用“首选项”窗口将其关闭。

JMP 工具提示

若您将鼠标悬停在下列项之上， JMP 会提供说明性工具提示 （或**悬停标签**）：

- 菜单或工具栏选项
- 图形中的标签
- 报表窗口中的文本结果（在结果上以圆圈的方式移动光标可显示提示）
- “主窗口”中的文件或窗口
- “脚本编辑器”中的代码

提示：在 Windows 上，可在“JMP 首选项”中隐藏工具提示。选择**文件 > 首选项 > 常规**，然后取消选择**显示菜单提示**。在 macOS 上该选项不可用。

JMP 用户社区

“JMP 用户社区”提供了多种选项，帮助您更好地学习 JMP 以及与其他 JMP 用户建立联系。学习资源库包含一页的指南、教程和演示，您可以从这里入手开始学习。之后您可以注册各种 JMP 培训课程以获得 JMP 进阶学习。

其他资源包括论坛、文件交换库（样本数据，脚本等）、网络学习视频以及社交网络小组等。

要访问网站上的 JMP 资源，选择**帮助 > 网络 JMP > JMP 用户社区**或访问 https://community.jmp.com。

免费在线统计思维课程

通过本免费在线课程的相关主题（例如探索性数据分析、质量方法以及相关和回归）学习实用统计技能。课程包括短片、演示、练习等等。访问 jmp.com/statisticalthinking。

JMP 新用户欢迎套件

“JMP 新用户欢迎套件”旨在帮助您快速熟悉 JMP 的基本操作。您将完成它的 30 个演示短片与活动，树立使用本软件的信心，以及与全球最大 JMP 用户网上社区建立联系。访问 jmp.com/welcome。

统计学知识门户

“统计学知识门户”使简明的统计解释与启发性的示例和图形相结合，帮助访问者奠定牢固的基础来逐步增强统计能力。访问 jmp.com/skp。

JMP 培训

SAS 提供各种由经验丰富的 JMP 专家团队领导的主题培训。提供公开课、网上实况课程和现场课程。您也可以选择网上在线学习订阅以在您方便的时候进行学习。访问 jmp.com/training。

用户编写的 JMP 手册

您可以从以下 JMP 网站获取 JMP 用户编写的有关使用 JMP 的其他手册。访问 jmp.com/books。

“JMP 起始页”窗口

若您不熟悉 JMP 或数据分析，可以从“JMP 起始页”窗口开始操作。在该窗口中，各种选项进行了分类并伴有说明，通过点击按钮即可启动。“JMP 起始页”窗口包括分析、图形、表和文件菜单中的许多选项。该窗口还列出 JMP Pro 功能和平台。

- 要打开“JMP 起始页”窗口，请选择**视图**（在 macOS 中选择**窗口**）> **JMP 起始页**。
- 在 Windows 中，要在打开 JMP 时自动显示“JMP 起始页”，请选择**文件** > **首选项** > **常规**，然后从“初始 JMP 窗口”列表中选择 **JMP 起始页**。在 macOS 中，选择 **JMP** > **首选项** > **常规** > **初始 JMP 起始页窗口**。

JMP 技术支持

JMP 技术支持由在 SAS 和 JMP 接受过培训的统计学家和工程师提供，其中很多人具有统计学或其他技术学科的研究生学位。

许多技术支持选项在 jmp.com/support 中提供，包括技术支持电话。

第 2 章

多元分析介绍

多元方法概述

《多元方法》说明以下用于同时分析若干变量的方法：

- “多元”平台检查多个变量以查看它们之间如何彼此相关。请参见“相关性和多元方法”。
- “主成分”平台从一组测量变量中得到少数几个相互独立的线性组合（主成分），使用它们来捕获原始变量中尽可能多的变异性。这是一种有用的探索性方法，可帮助您创建预测模型。请参见“主成分”。
- “判别”平台尝试找到一种根据已知的连续响应 (Y) 来预测分类 (X) 变量（名义型或有序型）的方法。它可以视为多元方差分析 (MANOVA) 的逆预测。请参见“判别分析”。
- “偏最小二乘”平台根据因子即解释变量 (X) 的线性组合来拟合线性模型。“偏最小二乘”利用 X 和 Y 之间的相关性来揭示底层的潜在结构。请参见“偏最小二乘模型”。
- “多重对应分析”(MCA) 平台适用于多个分类变量，并力求确定这些变量各水平之间的关联。MCA 常用在社会科学中，在法国和日本尤为普遍。可将其用在调查分析中，找出测试对象对不同问题的态度一致性。请参见“多重对应分析”。
- JMP PRO “结构化方程模型”支持您拟合各种模型，包括确认性因子分析、具有或不具有潜在变量的路径模型、测量值误差模型以及潜在变量增长曲线模型。请参见“结构化方程模型”。
- “因子分析”平台支持您从更大的一组观测变量中构造因子。这些因子可以表示为观测变量子集的线性组合。通过因子分析，您可以探索由一组测量的观测变量解释的因子数量，以及因子与变量之间关系的强度。请参见“因子分析”。
- “多维尺度化 (MDS)”平台支持您分析一组对象间的邻近性（相似性、相异性或距离），将它们之间的模式通过图形直观呈现出来。请参见“多维尺度化”。
- JMP PRO “多元嵌入”平台支持您将数据从极高维空间映射到低维空间。请参见“多元嵌入”。
- “项目分析”平台支持您拟合项目反应理论模型。项目反应理论 (IRT) 方法用于对测量手段（如：测试和问卷）进行分析和评分。IRT 使用模型体系将个体的特征与该个体正面或正确响应某个项目的概率相关联。IRT 可用于研究标准化测验、认知发展和消费者偏好。请参见“项目分析”。
- “层次聚类”平台将在几个变量上享有相似值的行分组在一起。这是一种有用的探索性方法，可帮助您理解数据的聚簇结构。请参见“层次聚类”。
- “K 均值聚类”平台将在几个变量上享有相似值的观测分组在一起。请参见“K 均值聚类”。
- “正态混合”平台可使您在数据来自重叠的正态分布时对观测聚类。请参见“正态混合”。
- “潜在类分析”平台可找到分类响应变量的观测聚类。该模型采取多项式混合模型的形式。请参见“潜在类分析”。

- “聚类变量”平台将相似变量分到典型组中。您可以将“聚类变量”用作降维方法。您不用在建模中使用大量变量，聚类中最典型的变量的聚类成分可用于解释数据中的大部分变异。请参见“聚类变量”。

第 3 章

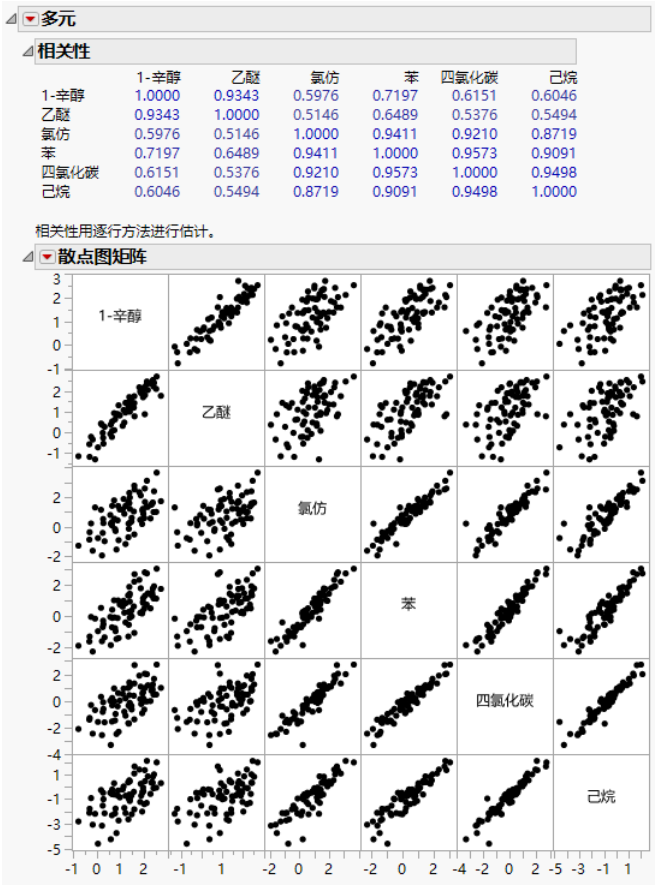
相关性和多元方法

探索变量的多维行为

多元数据涉及多个变量，而不是一个变量（一元）或两个变量（二元）。使用“多元”平台可探索多个变量如何彼此相关。“多元”平台提供许多方法来汇总和检验每对响应变量之间的线性关系的强度。该平台提供了参数和非参数相关性检验。还可以使用图形功能（例如散点图矩阵和色图）来标识变量之间的相依性、离群值和聚类。

还有其他的多元分析方法来进一步检验变量之间的关系，包括主成分分析、离群值分析和项目信度。这些方法可以通过“多元”报表获得。您还可以使用 JMP 中的“主成分分析”和“离群值分析”平台更深入地实施这些方法。

图 3.1 “多元”报表的示例



目录

“多元”平台的示例

本例在 “多元”平台中创建相关性矩阵和色图，以检查针对身体的不同测量值之间的关系。

1. 选择**帮助 > 样本数据文件夹**，然后打开 Body Measurements.jmp。
2. 选择**分析 > 多元方法 > 多元**。
3. 选择除**体重**之外的所有列并点击 **Y，列**。
4. 点击**确定**。

   初始多元报表包含相关性矩阵和散点图矩阵。还有一条注释告知您使用的是“逐行”方差估计方法。所有变量都是正相关，但强度不同。
5. 点击 “多元”红色小三角并取消选择**散点图矩阵**。
6. 点击 “多元”红色小三角并选择**相关性概率**。
7. 点击 “多元”红色小三角并选择**色图 > 相关性色图**。

图 3.2 针对人体测量数据的多元报表

多元

相关性

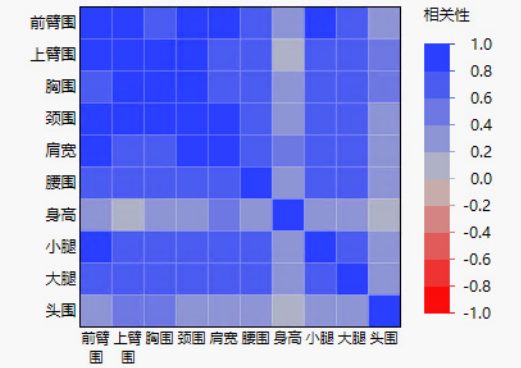
	前臂围	上臂围	胸围	颈围	肩宽	腰围	身高	小腿	大腿	头围
前臂围	1.0000	0.8134	0.7671	0.8134	0.8670	0.7577	0.3537	0.8130	0.7877	0.3029
上臂围	0.8134	1.0000	0.8552	0.8240	0.7202	0.6652	0.0657	0.6421	0.6974	0.4199
胸围	0.7671	0.8552	1.0000	0.8279	0.7999	0.7154	0.2017	0.6436	0.7802	0.4379
颈围	0.8134	0.8240	0.8279	1.0000	0.8312	0.7530	0.3912	0.6095	0.7055	0.3223
肩宽	0.8670	0.7202	0.7999	0.8312	1.0000	0.7462	0.4839	0.6847	0.6825	0.2258
腰围	0.7577	0.6652	0.7154	0.7530	0.7462	1.0000	0.3513	0.6893	0.7403	0.2997
身高	0.3537	0.0657	0.2017	0.3912	0.4839	0.3513	1.0000	0.3977	0.2860	0.1576
小腿	0.8130	0.6421	0.6436	0.6095	0.6847	0.6893	0.3977	1.0000	0.7497	0.2742
大腿	0.7877	0.6974	0.7802	0.7055	0.6825	0.7403	0.2860	0.7497	1.0000	0.2363
头围	0.3029	0.4199	0.4379	0.3223	0.2258	0.2997	0.1576	0.2742	0.2363	1.0000

相关性用逐行方法进行估计。

相关性概率

	前臂围	上臂围	胸围	颈围	肩宽	腰围	身高	小腿	大腿	头围
前臂围	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.1064	<.0001	<.0001	0.1706
上臂围	<.0001	<.0001	<.0001	<.0001	0.0002	0.0007	0.7715	0.0013	0.0003	0.0517
胸围	<.0001	<.0001	<.0001	<.0001	<.0001	0.0002	0.3681	0.0012	<.0001	0.0415
颈围	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0718	0.0026	0.0002	0.1436
肩宽	<.0001	0.0002	<.0001	<.0001	<.0001	<.0001	0.0225	0.0004	0.0005	0.3124
腰围	<.0001	0.0007	0.0002	<.0001	<.0001	<.0001	0.1089	0.0004	<.0001	0.1753
身高	0.1064	0.7715	0.3681	0.0718	0.0225	0.1089	<.0001	0.0668	0.1969	0.4836
小腿	<.0001	0.0013	0.0012	0.0026	0.0004	0.0004	0.0668	<.0001	<.0001	0.2168
大腿	<.0001	0.0003	<.0001	0.0002	0.0005	<.0001	0.1969	<.0001	<.0001	0.2898
头围	0.1706	0.0517	0.0415	0.1436	0.3124	0.1753	0.4836	0.2168	0.2898	<.0001

相关性色图

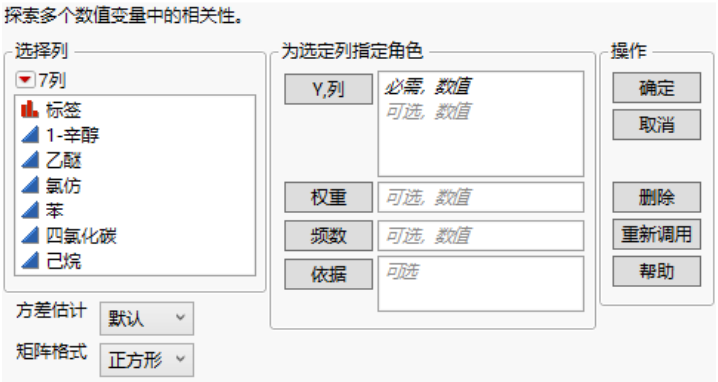


“相关性色图”报表提供“相关性矩阵”中信息的更简明版本。图中大部分是深蓝色，表明大多数变量高度相关。两个浅色的行和列表示身高和头围测量值与其他变量的相关性不高。“相关性概率”表中的“身高”和“头围”的大部分非显著 p 值都进一步支持了这一点。

启动“多元”平台

通过选择**分析** > **多元方法** > **多元**来启动“多元”平台。

图 3.3 “多元”启动窗口



有关“选择列”红色小三角菜单中选项的详细信息，请参见《使用 JMP》。

Y，列 标识一个或多个响应列。响应列必须具有数值数据类型，但建模类型可以是连续型或有序型。

注意：若您指定有序型响应变量，则在“启动”窗口中点击“确定”后会出现 JMP 警报。该警报指示哪些变量是有序型变量，并确认您打算在分析中包含有序型变量。

权重 标识一列，该列的数值为分析中的每一行都分配一个权重。

频数 标识一列，该列的数值为分析中的每一行都分配一个频数。

依据 为“依据”变量的每个水平生成单独的报表。若指定了多个“依据”变量，将为“依据”变量水平的每种可能组合生成单独的报表。

方差估计 指定计算相关性的方法。“REML”和“配对”是最常用的方法。其中有些方法解决了缺失数据的处理。您也可以使用“逐行”以外的方法估计缺失值，然后选择“补缺缺失数据”命令。请参见“补缺缺失数据”。

默认 默认选项使用“逐行”、“配对”或“REML”方法。

- 逐行估计用于不含缺失值的数据表。
- 配对估计用于包含缺失值并且多于 10 列、多于 5,000 行或列数多于行数的数据表。
- 在其他情况下使用 REML 估计。

注意：若“默认”选项在其他情况下导致 REML 估计，但拟合不能正确收敛，则平台转换为“配对”方法。当数据表中有缺失值且以下情形全部成立时就会出现这种情况：您的数据表少于 10 列、少于 5,000 行以及列数少于行数。若显示的方差估计为“配对”，这意味着 REML 拟合未收敛。

REML 限制最大似然 (REML) 估计使用所有数据，即使存在缺失值。由于存在偏倚修正因子，若您的数据集很大且包含许多缺失值，该方法会很慢。因此，REML 最适用于较小的数据集。若数据中不含缺失单元格，则 REML 和 ML 估计值等价于样本协方差矩阵。若存在缺失单元格，与 ML 估计相比，REML 的方差和协方差估计值的偏倚更小。有关统计详细信息，请参见“REML”。

ML 最大似然 (ML) 估计使用所有数据，即使存在缺失值。由于 ML 的估计值生成速度更快，该方法最适用于包含缺失数据的大数据表。

稳健 稳健估计使用所有数据，即使存在缺失值。该方法降低了极值的权重，因此最适用于可能具有离群值的数据表。有关统计详细信息，请参见“稳健”。

逐行 逐行估计为每对列计算 Pearson 相关性系数。有关统计详细信息，请参见“Pearson 乘积矩相关系数的统计详细信息”。逐行估计不使用包含缺失值的行。该方法适用于排除具有缺失数据的观测。

配对 配对估计使用所有数据，即使存在缺失值。该方差估计方法使用这两列中不含缺失值的所有观测为每对列计算 Pearson 相关性系数。有关统计详细信息，请参见“Pearson 乘积矩相关系数的统计详细信息”。配对估计最适用于包含缺失值并且列数多于行数、多于 10 列或多于 5,000 行的数据表。

注意：若您选择“REML”、“ML”或“稳健”，而您的数据表中的列数多于行数并且具有缺失值， JMP 会将 “方差估计”切换为 “配对”。

矩阵格式 选择用于 “散点图矩阵”的格式选项。“正方形”选项为所有排序列组合显示图。“下三角”在对角线下方显示图，前 n - 1 个列显示在水平轴上。“上三角”在对角线上方显示图，前 n - 1 个列显示在垂直轴上。

“多元”报表

默认的多元报表显示标准相关性矩阵、散点图矩阵，以及一条指示估计相关性所用的方法的注释。在某些情况下，还会显示解释为何使用给定方法的信息。平台菜单还列有其他的相关性选项以及分析方法用于研究多个变量之间的关系。请参见““多元”平台选项”。

“多元”平台选项

“多元”红色小三角菜单包含以下选项：

多元相关性 显示或隐藏“相关性”表。该表是一个相关系数矩阵，该矩阵汇总了每对响应 (*Y*) 变量之间的线性关系的强度。默认启用该选项。请参见 “Pearson 乘积矩相关系数的统计详细信息”。

注意：该相关性矩阵使用您在启动窗口中选择的方法来计算。

相关性概率 显示或隐藏“相关性概率”表。该表是一个 p 值矩阵。每个 p 值都对应变量之间的真实相关性为 0 这一原假设下的检验。这是针对两个响应变量之间不存在线性关系的检验。

相关性置信区间 显示或隐藏相关性的双侧置信区间。

提示：默认置信系数为 95%。使用设置 α 水平选项可更改置信系数。

逆相关性 显示或隐藏“逆相关性”报表，该报表是逆相关性矩阵。矩阵的对角线元素是一个函数，它表示为一个变量在多大程度上接近由其他变量构成的线性函数。在逆相关性表中，对角线为 $1/(1 - R^2)$ 的值。R^2 从对其他所有变量回归该变量的简单线性模型计算得出。若多重相关性为 0，则对角线逆元素为 1。若多重相关性为 1，则逆元素变为无穷大，所以报告为缺失。有关逆相关性的统计详细信息，请参见 “逆相关性矩阵的统计详细信息”。

偏相关性 显示或隐藏“偏相关性”报表，该报表是偏相关性矩阵。偏相关性矩阵显示了考虑其他所有变量的效应进行调整之后，两个变量之间关系的测度。该表是按照单位对角线统一尺度的负的逆相关性矩阵。这意味着矩阵经过统一尺度，以使对角线元素等于 1。

偏相关性概率 显示或隐藏“偏相关性概率”报表，该报表是 p 值的矩阵。每个 p 值都对应变量之间的真正偏相关性为 0 这一原假设下的检验。这是在对其他变量的效应进行调整之后，针对两个响应变量之间不存在线性关系的检验。

注意：在没有足够的自由度时，“偏相关性概率”选项不可用。当变量数多于观测数时就可能出现这种情况。

协方差矩阵 显示或隐藏协方差矩阵，该矩阵测量的是一对变量同时改变的程度。

配对相关性 显示或隐藏“配对相关性”表，该表列出每对 Y 变量的 Pearson 乘积矩相关性。相关性通过配对删除方法计算。若任何变量对中的任一个变量含有缺失值，计数值将有所不同。“配对相关性”报表还显示显著性概率并在条形图中比较相关性。所有结果都基于配对方法。

注意：一旦所考虑的两个变量中任意一个变量存在缺失值，该选项会排除相应的行。

Hotelling T^2 检验 支持您针对作为 Y 输入的变量执行多元分布的单样本均值检验。在选择“Hotelling T^2 检验”选项后，会出现一个窗口，该窗口支持您指定原假设下的均值向量。输入每个变量的假设均值。该检验假定 Y 变量服从多元正态分布。“Hotelling T^2 检验”报表提供以下信息：

变量 作为 Y 输入的变量。

均值 每个变量的样本均值。

假设均值 您指定的原假设均值。

检验统计量 Hotelling T^2 统计量的值。

F 比 检验统计量的值。若您有 n 行和 k 个变量，则 F 比的计算公式定义如下：

$$\frac{n-k}{k(n-1)}T^2$$

概率 > F 检验的 p 值。在原假设下，F 比服从自由度为 k 和 n - k 的 F 分布。

注意：要删除报表，请点击 T^2 检验旁边的红色小三角并选择**删除检验**。

简单统计量 该菜单包含两个选项，每个选项都可以显示或隐藏每列的简单统计量（均值、标准差、总和、最小值和最大值）。若含有缺失值或使用了“稳健”方法，一元和多元简单统计量会有所不同。

一元简单统计量 显示对每列计算的统计量，而不考虑其他列中的值或缺失值。这些值与“分布”平台生成的那些值相同。

多元简单统计量 显示与在启动窗口中选定的方差估计方法对应的统计量，以及是否存在缺失数据。若没有缺失观测，该选项仅可用于稳健方法。若有缺失观测，该选项可用于“配对”之外的所有方差估计方法。对于“REML”、“ML”或“稳健”方法，选定的方法估计均值向量和协方差矩阵。对于“逐行”方法，则从均值和方差的计算中排除至少包含一个缺失值的所有行。

非参数相关性 显示配对相关性的非参数测度的子菜单。每个选项显示或隐藏一个非参数报表，其中提供所选关联测度的显著性概率并且在条形图上显示关联值。共有三个非参数相关性测度。

Spearman Rho 一个基于数据值的秩而不是数据值本身计算的相关系数。

Kendall Tau 基于观测的一致对和不一致对的数目。若具有较大 X 值的观测也具有较大的 Y 值，则为**一致**对。若具有较大 X 值的观测具有较小的 Y 值，则为**不一致**对。同分对（即具有相等的 X 值或相等的 Y 值的观测对）会进行校正。

Hoeffding D 范围从 –0.5 到 1 的统计尺度。较大的正值说明较强的相关性。该统计量近似于 2 x 2 分类表的卡方统计量的观测加权和。通过将每个数据值设置为阈值生成 2 x 2 表。该统计量可以检测到更一般的独立性偏离。

注意：使用“配对”方法计算非参数相关性，即便在启动窗口中选定了其他方差估计方法。

注意：当指定了“权重”变量时，缺失值和零值权重会从非参数相关性计算中排除。所有其他权重值均视为 1。

有关这三种方法的统计详细信息，请参见 “关联的非参数测度的统计详细信息”。

设置 α 水平 您可以为相关性置信区间指定任意 alpha 值。共列出四个 alpha 值：0.01、0.05、0.10 和 0.50。选择**其他**可输入任何其他值。

散点图矩阵 显示或隐藏每对响应变量的散点图矩阵。默认启用该选项。请参见“散点图矩阵”。

色图 “色图”菜单包含三种色图。该菜单中的每个选项显示或隐藏相应类型的色图。提供以下类型的色图：

相关性色图 生成一个方格图，该图采用由蓝 (+1) 到红 (-1) 的尺度显示变量之间的相关性。

p 值色图 生成一个方格图，该图采用由 $p = 0$（红）到 $p = 1$（蓝）的尺度显示相关性的显著性。

对相关性进行聚类 生成一个方格图，该图将相似变量聚类在一起。该相关性与“相关性色图”的相关性相同，但变量的位置可能有所不同。

配对相关性色图 生成一个方格图，该图采用由蓝 (+1) 到红 (-1) 的尺度显示变量之间的配对相关性。

Spearman ρ 色图 生成一个方格图，该图采用由蓝 (+1) 到红 (-1) 的尺度显示变量之间的 Spearman ρ 非参数相关性。

Kendall τ 色图 生成一个方格图，该图采用由蓝 (+1) 到红 (-1) 的尺度显示变量之间的 Kendall τ 非参数相关性。

Hoeffding D 色图 生成一个方格图，该图采用由蓝 (+1) 到红 (-1) 的尺度显示变量之间的 Hoeffding D 非参数相关性。

平行坐标图 显示或隐藏变量的平行坐标图。

三维椭圆图 显示或隐藏带 95% 置信椭圆的三维散点图。选择该选项时，系统会提示您指定三个变量及其相应轴。

偏相关性关系图 显示或隐藏“偏相关性关系图”报表。该选项对偏相关性矩阵执行特征值分解，并使用结果直观表示偏相关性。有几个选项可用于更改关系图的显示。请参见“偏相关性关系图”。

离群值分析 该菜单包含的每个选项均用于显示或隐藏特定图，这些图使用以下方法之一测量多元意义上的距离：Mahalanobis 距离、Jackknife 距离和 T^2 统计量。请参见“离群值分析”。

项目信度 该菜单包含的每个选项都显示或隐藏一个项目信度报表。这些报表使用 Cronbach α 值或标准化 α 指示一组工具在测量总响应上的一致性程度。请参见“项目信度”。

补缺缺失数据 （仅当数据表包含缺失值时才可用。）生成一个新数据表，它复制您的数据表并用估计值替换所有缺失值。插补值是以每行的非缺失值为前提条件的期望值。通过启动窗口中选择的方法来估计的均值和协方差矩阵用于补缺计算。所有多元检验和选项随后可用于补缺数据集。

保存插补公式 （仅当数据表包含缺失值时才可用。）对于包含缺失值的列，将包含用于估计缺失值的公式的新列保存到数据表中。新列称为**补缺 _< 列名 >**。

请参见《使用 JMP》获取有关下列选项的信息：

本地数据过滤器 显示或隐藏支持您过滤特定报表中使用的数据的本地数据过滤器。

重新运行 包含使您可以重复或重新启动分析的选项。在支持该功能的平台中，“自动重新计算”选项立即在相应报表窗口中反映您对数据表所做的更改。

平台首选项 包含的选项支持您查看当前平台首选项或更新平台首选项以匹配当前 JMP 报表中的设置。

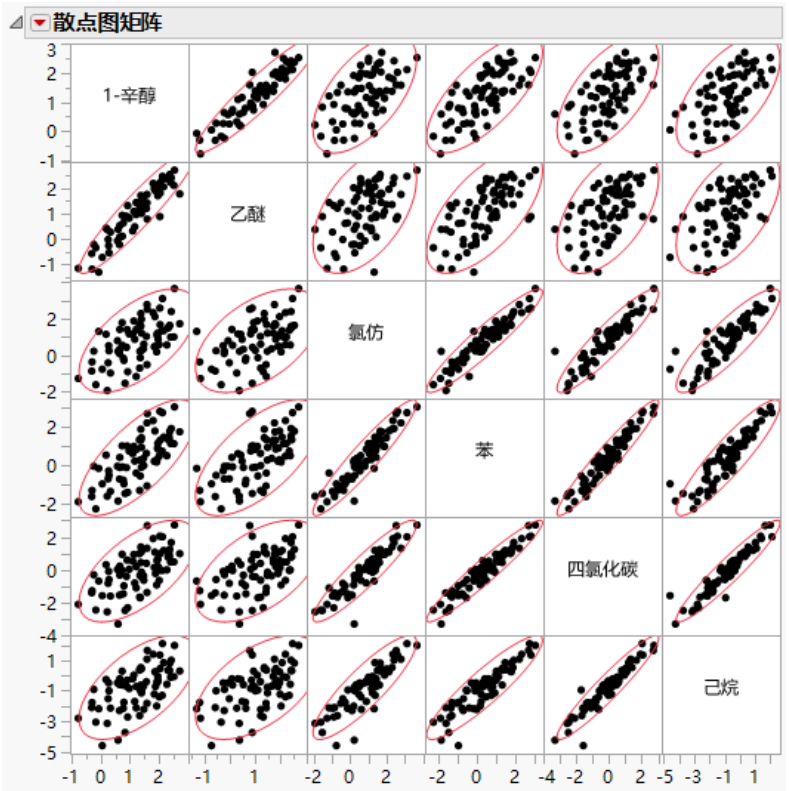
保存脚本 包含的选项支持您保存可将报表重现到若干目标的脚本。

保存 “依据”组脚本 包含使您可以保存脚本的选项，可将为 “依据”变量的所有水平重新生成平台报表的脚本保存到多个不同的位置。仅当在启动窗口中指定 “依据”变量时才可用。

散点图矩阵

在“多元”平台中，“散点图矩阵”报表帮助您直观演示每对响应变量之间的相关性。

图 3.4 散点图矩阵



启用“密度椭圆”选项可在每个散点图中都显示一个 95% 二元正态密度椭圆。假定每对变量都服从二元正态分布，该椭圆将覆盖大约 95% 的点。椭圆的窄度反映变量的相关度。若椭圆较圆，不沿着对角方向延伸，则变量之间不相关。若该椭圆较窄且向对角方向延伸，则变量之间相关。

提示：

- 调整任意单元的大小将调整所有单元的大小。
- 将标签单元拖至另一个标签单元可对矩阵重新排序。
- 查找散点图矩阵中的模式时，您可以看到变量根据其相关性聚类到组中。图 3.4 显示了两个相关性聚类：前两个变量（顶部，左侧）和后四个变量（底部，右侧）。

散点图矩阵选项

“散点图矩阵”红色小三角菜单中的选项支持您使用颜色和密度椭圆以及设置 α 水平来定制矩阵。

显示点 在散点图中显示或隐藏点。

拟合线 显示或隐藏回归线和拟合回归线的 95% 水平置信曲线。

密度椭圆 在散点图中显示或隐藏 95% 密度椭圆。使用椭圆 α 菜单可更改 α 水平。

着色椭圆 为每个椭圆着色。使用椭圆透明度和椭圆颜色菜单可更改透明度和颜色。

显示相关性 在每个散点图的左上角显示或隐藏每对变量的相关性。

矩阵选项 （仅当在启动窗口中选定“正方形”矩阵格式时才可用。）显示用于更改散点图矩阵右上小三角形外观的选项子菜单。一次只能选择以下选项之一。

显著性圆圈 在散点图矩阵的右上三角中显示或隐藏相关性圆圈。每个圆圈的颜色采用由红 (+1) 到蓝 (-1) 的尺度来表示每对变量之间的相关性。每个圆圈的大小表示变量之间的显著性检验。较大的圆圈指示关系更为显著。

热图 在散点图矩阵的右上三角中显示或隐藏相关性热图。热图中每个方格的颜色采用由红 (+1) 到蓝 (-1) 的尺度来表示每对变量之间的相关性。

显示直方图 在标签单元中显示或隐藏水平或垂直直方图。一旦添加直方图，选择**显示计数**可在直方图的每个直条上标记其计数。选择**水平**或**垂直**可更改直方图的方向或删除直方图。

椭圆 α 设置用于椭圆的 α 水平。在菜单中选择一个标准 α 水平，或选择**其他**输入不同值。

椭圆透明度 若椭圆已着色，则设置其透明度。选择一个默认水平，或选择**其他**输入不同值。默认值为 0.2。

椭圆颜色 若椭圆已着色，则设置其颜色。在调色板中选择一种颜色，或选择**其他**使用另一种颜色。默认值为红色。

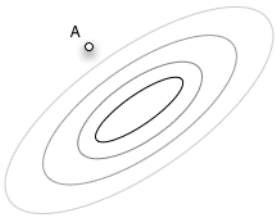
非参数密度 基于描述数据点密度的平滑非参数二元曲面显示或隐藏着色密度等高线。显示非参数曲面的 10% 和 50% 分位数的等高线。

离群值分析

在“多元”红色小三角菜单中，“离群值分析”选项包含一个子菜单，其中包含用于离群值检测的三个距离测度。子菜单中的每个选项都显示或隐藏一个图，该图测量多元意义上相对于相关性结构的距离。检验是在图底部显示的 alpha 水平下执行的。

在图 3.5 中，“点 A”是一个离群值，这是因为它位于相关性结构之外，即便它在任意坐标方向上都不是一个离群值。

图 3.5 离群值示例



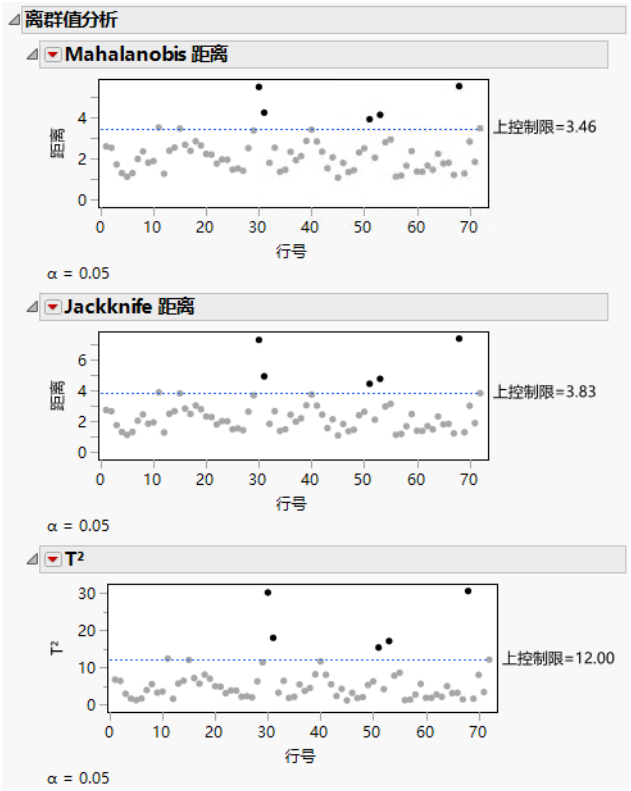
提供以下距离测度选项：

Mahalanobis 距离 显示或隐藏每个点与多元均值（重心）的 Mahalanobis 距离。标准 Mahalanobis 距离取决于数据的均值、标准差和相关性的估计值。为每个观测编号都标绘距离。通过突出显示具有最大距离值的点来标识极端多元离群值。请参见“Mahalanobis 距离测度”。

Jackknife 距离 显示或隐藏使用 Jackknife 方法计算的距离。使用均值、标准差和相关性矩阵（不含观测本身）的估计值计算每个观测的距离。当存在离群值时 Jackknife 距离会很有用。在这种情况下，Mahalanobis 距离会失真，往往会隐藏离群值或令其他点看起来比实际上离群。请参见“Jackknife 距离测度”。

T^2 显示或隐藏 Mahalanobis 距离平方后的距离。该图更适用于多元控制图，它包含计算的 T^2 统计量的值及其上控制限。落在该限值之外的值可能为离群值。请参见“T^2 距离测度”。

图 3.6 离群值分析图



保存距离和值

您可以通过从该图的红色小三角菜单中选择**保存**选项，将所有距离保存至数据表。

注意：没有任何公式与 Jackknife 距离列一同保存。这意味着您修改数据表后**不**会重新计算该距离。若在数据表中添加/删除列或更改值，请再次选择**分析 > 多元方法 > 多元**以计算新的 Jackknife 距离。

除了保存每行的距离值，还会创建一个列属性，用于对指定的“离群值分析”类型提供上控制限 (UCL) 值。

项目信度

您可以在“多元”平台中直接执行项目信度分析。项目信度指示一组工具测量总响应的一致性。Cronbach α 值 (Cronbach 1951) 是一种信度测度。Cronbach α 值的两个主要应用分别为工具信度和问卷分析。

Cronbach α 值基于在测量值尺度内项目的相关性平均值。它相当于计算数据表中所有折半相关性的平均值。若项目的方差变化很大，则可以请求计算“标准化 α”。

注意：Cronbach α 值与显著性水平 α 没有关系。此外，项目信度与生存时间信度分析无关。

要查看单个项目的影响，JMP 将从计算中排除该项目，并显示 Cronbach α 值的影响。若 α 值在您排除某个变量（项目）后增大，则该变量不与其他变量高度相关。若 α 值减小，您可以得出结论：该变量在该尺度上与其他项目相关。

有关计算的详细信息，请参见“Cronbach α 的统计详细信息”。

偏相关性关系图

“多元”平台中的“偏相关性关系图”报表提供对偏相关性的直观表示。在该关系图中，每个变量对应一个节点，每对变量对应一个边。节点在关系图中的位置由“节点方向”选项确定。提供两个选项：圆形和特征向量。“圆形”方向将各个节点均匀分布在一个圆中（若关系图不是正方形，则为椭圆）。“特征向量”方向基于特征向量来排布各个节点。每个特征向量乘以相应特征值的平方根可创建一组统一尺度的特征向量。关系图上的坐标是在关系图下方指定的维的统一尺度的特征向量。

每个变量节点都通过一条线连接到其他节点。这条线的颜色表示变量之间正相关（红色）或负相关（蓝色）。这条线的宽度表示变量之间相关性的强度。这条线越粗，相关性越强。

使用以下选项调整该关系图：

节点方向 将节点的方向改为“特征向量”或“圆形”。

旋转 使用旋转按钮  旋转关系图。

偏相关性绝对值 隐藏偏相关性绝对值低于指定数字的行。您可以通过在“偏相关性绝对值”旁边的框中输入值或使用滑块来指定截止值。

颜色分离 调整线条颜色。

宽度 调整线条的粗细。

透明度 调整线条的透明度。

重置 将所有显示选项重置为默认值。

选择维 （仅适用于特征值方向。）指定用于节点间距的特征向量维。

“偏相关性关系图”选项

“偏相关性关系图”红色小三角菜单包含以下选项：

特征值 显示或隐藏从偏相关性矩阵分解而来的排序特征值表。

特征向量 显示或隐藏从偏相关性矩阵分解而来的特征向量表。

显示偏相关性 显示或隐藏关系图上相应弧顶部的偏相关性值。

颜色主题 支持您为关系图选择颜色主题。

保存坐标 将统一尺度的特征向量 （维）保存至新的数据表。统一尺度的特征向量是特征向量乘以相应特征值的平方根所得的结果，是在偏相关性关系图中使用的坐标。

“多元”平台的更多示例

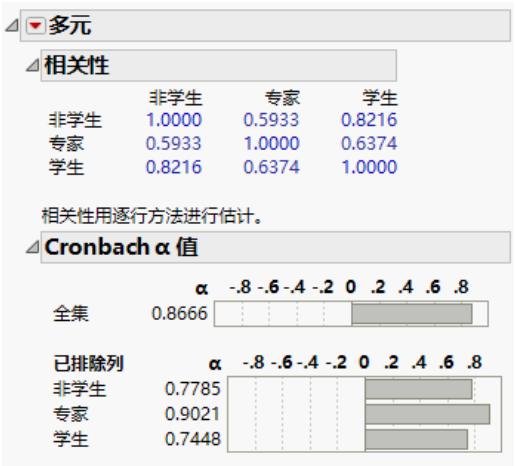
本节包含使用“多元”平台的示例。

“项目信度”示例

在本例中，您使用“多元”平台中的项目信度分析，以三组人员根据感知的危险程度对 30 个项进行排名的方式评估一致性。请注意，在这类示例中，值对每组而言都是相同的一组排名，将数据标准化没有任何影响。

1. 选择**帮助** > **样本数据文件夹**，然后打开 Danger.jmp。
2. 选择**分析** > **多元方法** > **多元**。
3. 选择除**活动**之外的所有列并点击 **Y, 列**。
4. 点击**确定**。
5. 点击 “多元”红色小三角并选择**项目信度** > Cronbach α 值。
6. （可选）点击 “多元”红色小三角并选择**散点图矩阵**以隐藏该图。

图 3.7 “Cronbach α 值”报表



“Cronbach α 值”结果显示总 α 值为 0.8666，该值指示三个组中的排名值具有较高的相关性。不仅如此，当您从分析中排除专家后，**非学生**和**学生**对危险的排名几乎相同，二者的 Cronbach α 值得分分别为 0.7785 和 0.7448。

偏相关性示例

使用“多元”平台检查金属棒不同间隔处涂层厚度之间的关系。

1. 选择**帮助 > 样本数据文件夹**，然后打开 Quality Control/Thickness.jmp。
2. 选择**分析 > 多元方法 > 多元**。
3. 选择所有**厚度**列并点击 **Y，列**。
4. 点击**确定**。
5. （可选）点击 “多元”红色小三角并选择**散点图矩阵**以便从报表中删除散点图矩阵。
   这样做是为了清理报表，因为我们不会用到散点图矩阵。
6. 点击 “多元”红色小三角并选择**偏相关性**。

图 3.8 针对 “厚度”的相关性和偏相关性

多元

相关性

	厚度 01	厚度 02	厚度 03	厚度 04	厚度 05	厚度 06	厚度 07	厚度 08	厚度 09	厚度 10	厚度 11	厚度 12
厚度 01	1.0000	0.9815	0.9624	0.9590	0.9669	0.9793	0.9825	0.9803	0.9807	0.9521	0.9615	0.8250
厚度 02	0.9815	1.0000	0.9720	0.9568	0.9598	0.9678	0.9633	0.9558	0.9513	0.9198	0.9304	0.8490
厚度 03	0.9624	0.9720	1.0000	0.9942	0.9928	0.9867	0.9726	0.9671	0.9613	0.9127	0.9215	0.8872
厚度 04	0.9590	0.9568	0.9942	1.0000	0.9949	0.9878	0.9776	0.9719	0.9665	0.9158	0.9241	0.8880
厚度 05	0.9669	0.9598	0.9928	0.9949	1.0000	0.9961	0.9858	0.9824	0.9782	0.9237	0.9335	0.8663
厚度 06	0.9793	0.9678	0.9867	0.9878	0.9961	1.0000	0.9926	0.9891	0.9854	0.9291	0.9401	0.8462
厚度 07	0.9825	0.9633	0.9726	0.9776	0.9858	0.9926	1.0000	0.9977	0.9919	0.9392	0.9497	0.8383
厚度 08	0.9803	0.9558	0.9671	0.9719	0.9824	0.9891	0.9977	1.0000	0.9960	0.9496	0.9588	0.8312
厚度 09	0.9807	0.9513	0.9613	0.9665	0.9782	0.9854	0.9919	0.9960	1.0000	0.9673	0.9746	0.8237
厚度 10	0.9521	0.9198	0.9127	0.9158	0.9237	0.9291	0.9392	0.9496	0.9673	1.0000	0.9986	0.8259
厚度 11	0.9615	0.9304	0.9215	0.9241	0.9335	0.9401	0.9497	0.9588	0.9746	0.9986	1.0000	0.8247
厚度 12	0.8250	0.8490	0.8872	0.8880	0.8663	0.8462	0.8383	0.8312	0.8237	0.8259	0.8247	1.0000

相关性用逐行方法进行估计。

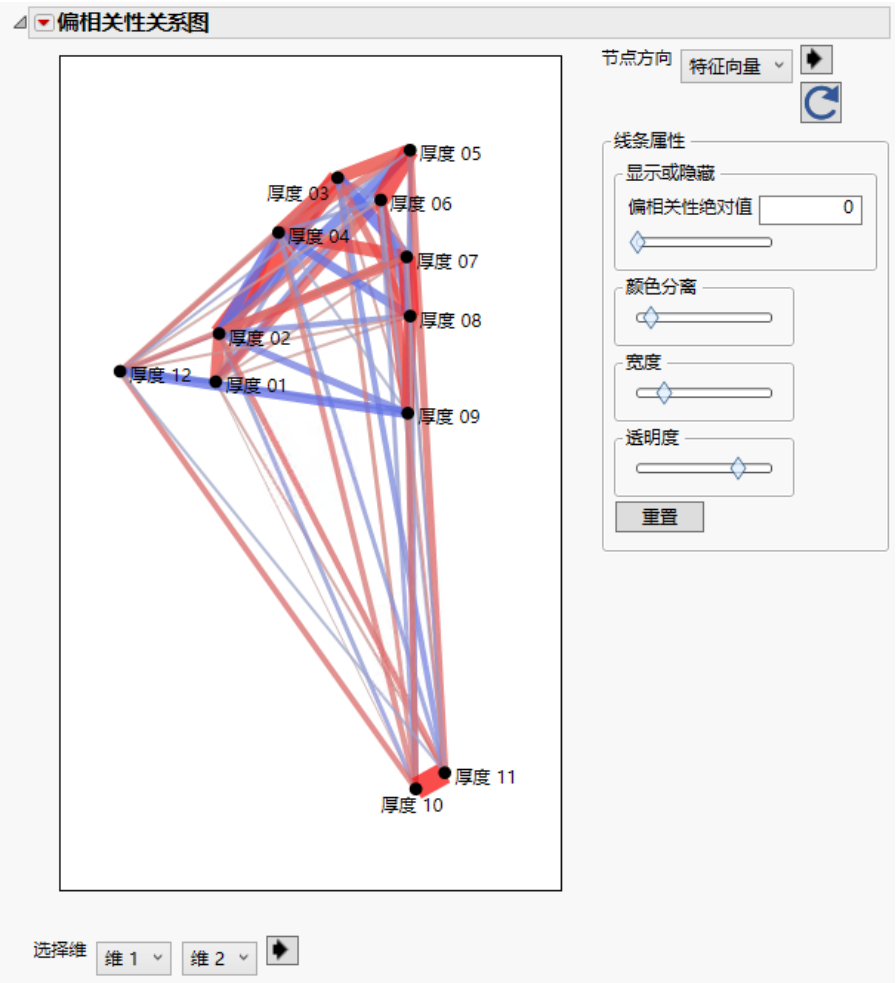
偏相关性

	厚度 01	厚度 02	厚度 03	厚度 04	厚度 05	厚度 06	厚度 07	厚度 08	厚度 09	厚度 10	厚度 11	厚度 12
厚度 01	.	0.4726	-0.1404	0.1830	-0.2349	0.3387	0.0529	0.0493	-0.0563	0.0331	0.0609	-0.1690
厚度 02	0.4726	.	0.7829	-0.4282	-0.3929	0.2099	0.4214	-0.1933	-0.2868	-0.1195	0.2325	-0.0784
厚度 03	-0.1404	0.7829	.	0.5479	0.5054	-0.1240	-0.4829	0.2775	0.0827	0.1277	-0.2094	0.1728
厚度 04	0.1830	-0.4282	0.5479	.	0.2432	-0.1254	0.5071	-0.3655	-0.0370	0.1735	-0.1360	0.1292
厚度 05	-0.2349	-0.3929	0.5054	0.2432	.	0.7058	0.0297	0.0909	-0.1508	-0.1659	0.2545	-0.0860
厚度 06	0.3387	0.2099	-0.1240	-0.1254	0.7058	.	0.1242	-0.2577	0.4602	-0.1098	-0.0539	0.0399
厚度 07	0.0529	0.4214	-0.4829	0.5071	0.0297	0.1242	.	0.7569	0.0580	-0.1731	0.0644	0.2092
厚度 08	0.0493	-0.1933	0.2775	-0.3655	0.0909	-0.2577	0.7569	.	0.5104	0.0341	-0.0850	0.0711
厚度 09	-0.0563	-0.2868	0.0827	-0.0370	-0.1508	0.4602	0.0580	0.5104	.	0.1882	0.0916	-0.4119
厚度 10	0.0331	-0.1195	0.1277	0.1735	-0.1659	-0.1098	-0.1731	0.0341	0.1882	.	0.9449	0.2184
厚度 11	0.0609	0.2325	-0.2094	-0.1360	0.2545	-0.0539	0.0644	-0.0850	0.0916	0.9449	.	-0.0516
厚度 12	-0.1690	-0.0784	0.1728	0.1292	-0.0860	0.0399	0.2092	0.0711	-0.4119	0.2184	-0.0516	.

偏相关性矩阵表明变量之间既有正相关也有负相关。这不同于标准相关性矩阵，该矩阵仅显示正相关。这是因为偏相关性测量的是调整其他所有变量的效应之后，一对变量之间关系的强度。这样可以更清楚地表明各对变量之间的真实关系。

7. 点击 “多元”红色小三角并选择**偏相关性关系图**。

图 3.9 偏相关性关系图



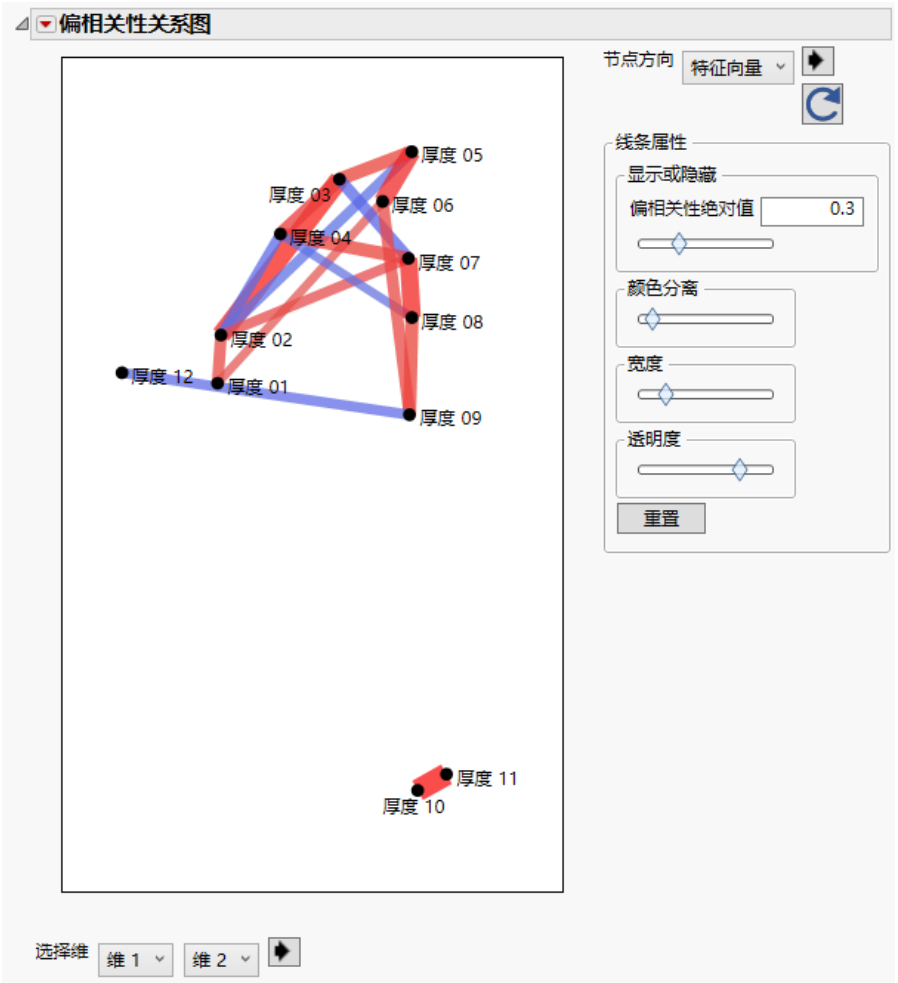
“偏相关性关系图”可直观演示偏相关性。图 3.9 中的关系图为每个厚度变量都显示节点，变量的坐标对应于第一个和第二个统一尺度的特征向量。“厚度 10”和 “厚度 11”在前两个维上彼此紧挨，但与其余厚度变量分开。连接 “厚度 10”和 “厚度 11”对应的节点的粗红线指示这两个变量之间也存在较强的正偏相关。

由于其余节点聚在一起且难以解释，因此可以通过指定“偏相关性绝对值”的值来调整所显示的行。这样可仅显示偏相关性大于 “偏相关性绝对值”所指定的值的那些行。

8. 在 “偏相关性绝对值”旁边的框中输入 1。

这将从关系图中删除所有行。要轻松查看哪些变量具有最强的偏相关性，请使用“偏相关性绝对值”下方的滑块并将菱形缓慢向左拖动。图 3.10 中的关系图所显示的线条连接的是具有绝对值大于 0.3 的偏相关性的各个变量。

图 3.10 偏相关性大于 0.3



“多元”平台的统计详细信息

本节包含 “多元”平台的统计详细信息。

- “方差估计方法的统计详细信息”
- “Pearson 乘积矩相关系数的统计详细信息”
- “关联的非参数测度的统计详细信息”
- “逆相关性矩阵的统计详细信息”
- “距离测度的统计详细信息”
- “Cronbach α 的统计详细信息”

方差估计方法的统计详细信息

本节包含 “多元”平台中使用的方差估计方法的统计详细信息。

REML

当数据包含缺失值时，与 ML（最大似然）估计方法相比，REML（限制最大似然）估计值的偏倚更小。REML 方法基于误差对比将边缘似然最大化。REML 方法经常用于估计方差和协方差。“主成分”平台中的 REML 方法与针对重复测量数据（具有非结构化协方差矩阵）使用的混合模型的 REML 估计相同。有关混合模型的 REML 估计的信息，请参见 SAS Institute Inc.(2020e) 中的“MIXED 过程”一章。

稳健

该方法实际上通过极大地降低所有离群值的权重来忽略它们。使用以下权重执行一系列迭代再加权数据拟合：

若 $Q < K$，则 $w_i = 1.0$；其他情况下 $w_i = K/Q$

在此， K 是一个等于卡方分布的 0.75 分位数的常数，该分布的自由度等于数据表中的列数。 Q 定义如下：

$$Q = (y_i - \mu)^T (S)^{-1} (y_i - \mu)$$

在该等式中，y_i = 第 i 个观测的响应，μ = 均值向量的当前估计值，S = 协方差矩阵的当前估计值，T = 转置矩阵运算。最后一步是减小方差矩阵的偏倚。

这是一个折衷方法：当数据中的离群值不多时，您可以得到较高的方差估计值；但当数据确实包含离群值时，您可以得到准确得多的方差估计值。

Pearson 乘积矩相关系数的统计详细信息

在 “多元”平台中， Pearson 乘积矩相关系数测量两个变量之间线性关系的强度。对于响应变量 X 和 Y，该系数表示为 r，计算公式如下所示：

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2} \sqrt{\sum (y - \bar{y})^2}}$$

若两个变量之间存在精确线性关系，则相关系数为 1 或 –1，具体取决于变量之间是正相关还是负相关。若不存在线性关系，则相关系数趋向于 0。

关联的非参数测度的统计详细信息

“多元”平台提供三个关联的非参数测度：Spearman、Kendall 或 Hoeffding 相关性。要计算以上任何相关系数，首先要对数据进行秩排序。随后根据数据值的秩执行计算。若存在结值，则使用平均秩。

注意：当指定了“权重”变量时，缺失值和零值权重会从非参数相关性计算中排除。所有其他权重值均视为 1。

Spearman ρ (rho) 系数

使用上述 Pearson 相关性公式基于数据的秩计算 Spearman ρ 相关系数。

Kendall τ_b 系数

Kendall τ_b 系数基于一致对和不一致对的数目。若两个变量所对应的一对行对于哪一个变量较大这一点上保持一致，则认为这一对行**一致**。否则这一对行不一致或为同分对。

公式

$$\tau_b = \frac{\sum_{i<j} \operatorname{sgn}(x_i - x_j)\operatorname{sgn}(y_i - y_j)}{\sqrt{(T_0 - T_1)(T_0 - T_2)}}$$

计算 Kendall τ_b，其中：

$$T_0 = (n(n-1))/2$$

$$T_1 = \sum((t_i)(t_i-1))/2$$

$$T_2 = \sum((u_i)(u_i-1))/2$$

请注意以下事项：

- 若 z>0，则 sgn(z) 等于 1；若 z=0，则 sgn(z) 等于 0；若 z<0，则 sgn(z) 等于 -1。
- t_i (u_i) 是第 i 组 x (y) 结值中的 x (y) 结值的数目。
- n 是观测数。
- Kendall τ_b 的取值范围为 –1 到 1。若指定了权重变量，则忽略该统计量。

按照以下方式执行计算：

- 根据第一个变量的值对观测进行秩排序。
- 然后根据第二个变量的值对观测重新进行秩排序。
- 第一个变量的交换次数用于计算 Kendall τ_b。

Hoeffding D 统计量

Hoeffding *D* (1948) 的公式如下

$$D = 30\left(\frac{(n-2)(n-3)D_1 + D_2 - 2(n-2)D_3}{n(n-1)(n-2)(n-3)(n-4)}\right)$$

其中：

$$D_1 = \Sigma_i(Q_i-1)(Q_i-2)$$
$$D_2 = \Sigma_i(R_i-1)(R_i-2)(S_i-1)(S_i-2)$$
$$D_3 = \Sigma_i(R_i-2)(S_i-2)(Q_i-1)$$

请注意以下事项：

- R_i 和 S_i 分别是 x 值和 y 值的秩。
- Q_i（有时称为二元秩）是 1 加上 x 值和 y 值均小于第 i 个点的点数。
- 在其 x 值或 y 值上有结值的点（而不是这两个值上同时具有结值的点）为 Q_i 贡献了 1/2（若其他值小于第 i 个点的相应值）。同时在 x 和 y 上有结值的点为 Q_i 贡献了 1/4。

若观测中没有结值，*D* 统计量的值介于 –0.5 和 1 之间，其中 1 指示完全依赖。若指定了权重变量，则忽略该统计量。

逆相关性矩阵的统计详细信息

在“多元”平台中，逆相关性矩阵提供了有用的多元信息。逆相关性矩阵的对角线元素有时称为方差膨胀因子 (VIF)，它是一个函数，表示为一个变量在多大程度上接近由其他变量构成的线性函数。具体而言，若相关性矩阵表示为 **R** 并且逆相关性矩阵表示为 $\mathbf{R}^{-1}$，则对角线元素表示为 r_{ii}，它的计算公式如下：

$$r_{ii} = \mathrm{VIF}_i = \frac{1}{1-\mathrm{R}_i^2}$$

其中，R_i^2 是用第 i 个解释变量对其他解释变量作回归的模型的变异系数。因此，较大的 r_{ii} 指示第 i 个变量与任意数量的其他变量高度相关。

距离测度的统计详细信息

本节包含 “多元”平台的 “离群值分析”图中使用的距离测度的统计详细信息。

Mahalanobis 距离测度

Mahalanobis 距离将数据的相关性结构和不同的尺度考虑在内。对于每个值， Mahalanobis 距离都表示为 M_i，它的计算公式如下：

$$M_i = \sqrt{(Y_i - \bar{Y})'S^{-1}(Y_i - \bar{Y})}$$

其中：

Y_i 是第 i 行的数据

$\bar{Y}$ 是均值行

S 是数据的估计协方差矩阵

“Mahalanobis 距离”图上绘制的上控制限参考线 (Mason and Young, 2002) 计算如下：

$$UCL_{Mahalanobis} = \sqrt{\frac{(n-1)^2}{n}\beta_{\left[1-\alpha;\frac{p}{2};\frac{n-p-1}{2}\right]}}$$

其中：

n = 观测数

p = 变量 （列）数

$\beta_{\left[1-\alpha;\frac{p}{2};\frac{n-p-1}{2}\right]}$ = Beta $\left(\frac{p}{2}, \frac{n-p-1}{2}\right)$ 分布的第 (1–α) 分位数

若某个变量是其他变量的精确线性组合，则相关性矩阵是奇异矩阵，并且该变量的行和列将被清零。生成的广义逆仍可用于距离计算。

Jackknife 距离测度

使用均值、标准差和相关性矩阵 （不含观测本身）的估计值计算 Jackknife 距离。对于每个值，Jackknife 距离计算如下：

$$J_i = \sqrt{\frac{(n-2)n^2}{(n-1)^3} \times \frac{M_i^2}{1 - \frac{nM_i^2}{(n-1)^2}}}$$

其中：

n = 观测数

p = 变量 （列）数

M_i = 第 i 个观测的 Mahalanobis 距离

“Jackknife 距离”图上绘制的上控制限参考线 (Penny, 1996) 计算如下：

$$UCL_{Jackknife} = \sqrt{\frac{(n-2)n^2}{(n-1)^3} \times \frac{UCL_{Mahalanobis}^2}{1 - \frac{n \bullet UCL_{Mahalanobis}^2}{(n-1)^2}}}$$

T^2 距离测度

T^2 距离是 Mahalanobis 距离的平方，所以 $T_i^2 = M_i^2$。

T^2 距离的上控制限为：

$$UCL_{T^2} = \frac{(n-1)^2}{n} \beta_{\left[1-\alpha;\frac{p}{2};\frac{n-p-1}{2}\right]} = (UCL_{Mahalanobis})^2$$

其中

n = 观测数

p = 变量 （列）数

$\beta_{\left[1-\alpha;\frac{p}{2};\frac{n-p-1}{2}\right]}$ = Beta $\left(\frac{p}{2}, \frac{n-p-1}{2}\right)$ 分布的第 (1–α) 分位数

多元距离可用于发现多维数据中的离群值。不过，若变量在多元意义上高度相关，那么某个点在多元空间中会被视为离群值，但沿着任意维子集的方向却无异常之处。换言之，若各值相关，沿着一两个轴查看某个点时，该点可能并不值得注意，但因为违反了相关性仍可能为离群值。

Cronbach α 的统计详细信息

“多元”红色小三角菜单的“项目信度”选项中使用的统计量为 Cronbach α。Cronbach α 值定义如下：

$$\alpha = \frac{kc}{v + (k-1)c}$$

其中

k = 尺度范围中的项目数

c = 项目之间的平均协方差

v = 项目之间的平均方差

若项目经过标准化具有常数方差，公式将变为

$$\alpha = \frac{k(r)}{1 + (k-1)r}$$

其中

r = 项目之间的相关性平均值

总体 α 系数越大，您越能够自信地认为您的项目对尺度或检验的信度有贡献。若有许多高度相关的项目，该系数可能接近 1.0。

第 4 章

主成分

对数据降维

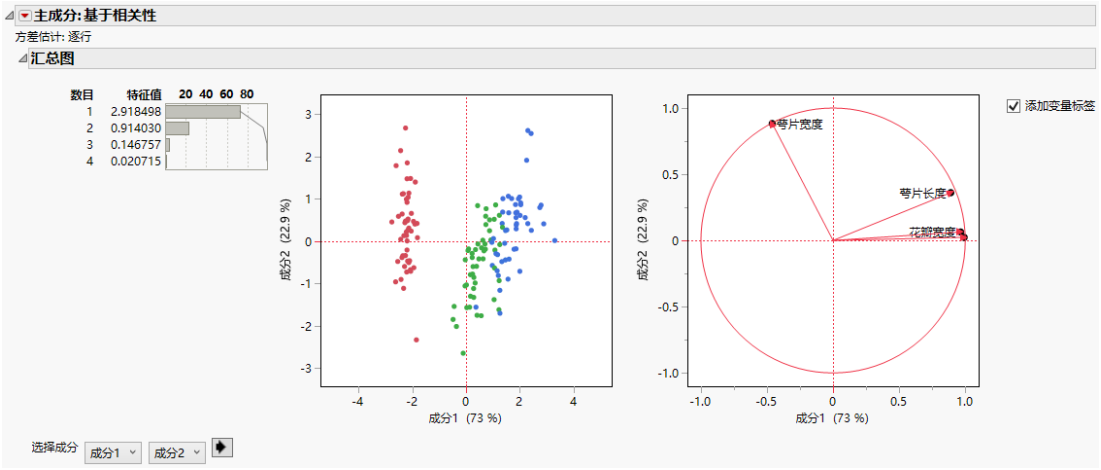
主成分分析的目的是要从一组测量变量中得到少数几个相互独立的线性组合（**主成分**），使用它们来捕获原始变量中尽可能多的变异性。主成分分析是一种降维方法，也是一种探索性数据分析工具。主成分分析也可用于构造预测模型，如**主成分分析回归**（亦称 PCA 回归或 PCR）所述。

对于包含大量变量的数据，“主成分”平台提供称为“宽”方法的估计方法。“宽”方法使您能够在较短的计算时间内计算主成分。随后可将这些主成分用在 PCA 回归中。

对于包含大量零的数据，也称为**稀疏数据**，“主成分”平台提供“稀疏”估计方法。与“宽”方法类似，“稀疏”方法能够以较短的计算时间计算主成分。与“宽”方法的不同之处在于，“稀疏”方法计算用户定义的固定数量的主成分，而不是完整集合。

“主成分”平台还支持因子分析。JMP 提供若干类型的正交和斜交因子分析样式的旋转，用来帮助解释提取的成分。有关因子分析的信息，请参见“因子分析”。

图 4.1 主成分示例



目录

“主成分”平台概述

主成分分析采用一组变量的少数几个相互独立的线性组合（**主成分**）的形式，对这组变量中的变异建模。

若您想要查看各点在许多相关变量中的排列，则可以使用主成分分析显示高维数据的最显著方向。使用主成分分析可减少一组数据的维度。主成分是使用尽量少的变量尽可能完全标绘数据结构的一种方式。

对于 p 个变量，这是 p 个主成分形成的方式：

- 第一主成分是进行标准化之后的原始变量的线性组合，具有最大可能方差。
- 随后的每个主成分都是具有最大可能方差并且与之前定义的所有成分都无关的变量的线性组合。

通过将相关性矩阵（协方差矩阵或平方和与叉积矩阵）的特征向量与变量进行线性组合来计算每个主成分。特征值表示每个成分的方差。

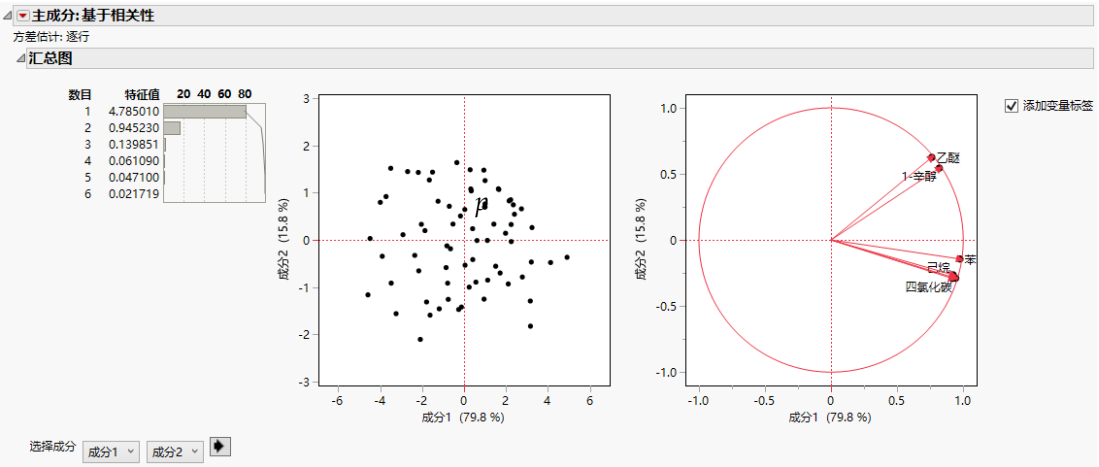
“主成分”平台使您能够针对相关性矩阵、协方差矩阵或未统一尺度且未中心化的数据执行分析。您也可以在“主成分”平台内执行“因子分析”。请参见“因子分析”。

主成分分析的示例

在本例中，您针对六个因子执行主成分分析。

1. 选择**帮助 > 样本数据文件夹**，然后打开 Solubility.jmp。
2. 选择**分析 > 多元方法 > 主成分**。
3. 选择所有连续列并点击 **Y, 列**。
4. 由于您仅分析六个因子，所以选择 “方法系列”旁边的**窄数据**。
5. 保留默认的 “方差估计”，然后点击**确定**。

图 4.2 “主成分 : 基于相关性”报表



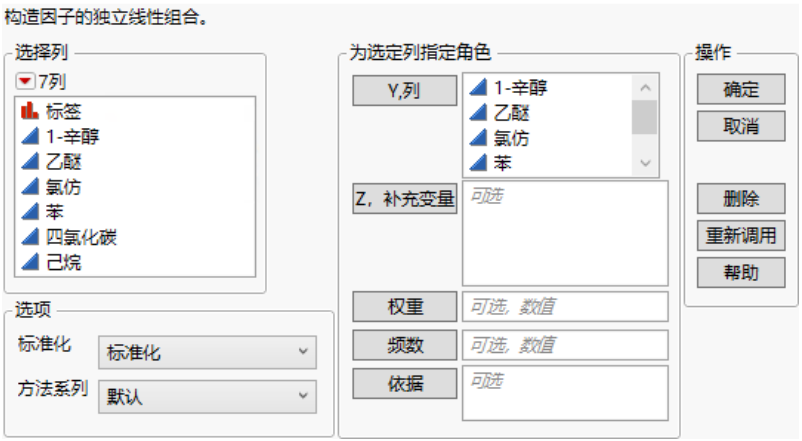
该报表提供特征值和一个条形图，图中显示了每个主成分所解释的变异百分比。在本例中，第一个主成分解释数据中大约 80% 的变异。前两个主成分一起，共同解释了数据中几乎全部的变异 (95.5%)。同时还有一个“得分图”和一个“载荷图”。请参见“‘主成分’报表”。

启动“主成分”平台

通过选择**分析 > 多元方法 > 主成分**来启动“主成分”平台。“多元”平台和“三维散点图”平台也提供主成分分析。

“主成分分析的示例”中所述的示例使用 Solubility.jmp 样本数据表中的所有连续变量。

图 4.3 “主成分”启动窗口



有关“选择列”红色小三角菜单中选项的详细信息，请参见《使用 JMP》。

Y，列 要进行成分分析的变量。

Z，补充变量 要显示的补充变量。补充变量不包含在主成分计算中，包括它们不会影响结果。连续型补充变量可以投影到载荷图中，用于增强解释。

权重 标识一列，该列的数值为分析中的每一行都分配一个权重。

注意：“宽”方差估计方法和“稀疏”方差估计方法忽略“权重”角色。

频数 标识一列，该列的数值为分析中的每一行都分配一个频数。

注意：“宽”方差估计方法和“稀疏”方差估计方法忽略“频数”角色。

依据 为“依据”列指定的每个值创建“主成分”报表，以便您可以为每个组执行单独的分析。

标准化 指定每列是否中心化和标准化。这决定了用于计算主成分的矩阵。

标准化 分别对每列执行中心化和标准化。基于相关性矩阵计算主成分。

未统一尺度 分别对每列执行中心化。基于协方差矩阵计算主成分。

未统一尺度且未中心化 基于未统一尺度且未中心化的矩阵计算主成分。

方法系列 指定数据类型。

默认 若列数小于 500 或小于行数，则将“窄数据”指定为“方法系列”。若列数大于 500 并且大于行数，则 JMP 警示窗口会推荐宽估计方法。点击**宽方法（快速）**以使用宽数据估计方法，或点击**默认方法（慢速）**以使用窄数据估计方法。

窄数据 使用协方差矩阵、相关性矩阵或未统一尺度且未中心化的矩阵获取主成分。

宽数据 使用奇异值分解获取主成分。

方差估计 （仅当将“窄数据”指定为“方法系列”时才可用。）指定计算相关性的方法。其中有些方法解决了缺失数据的处理。

默认 默认选项使用“逐行”、“配对”或“REML”方法。“JMP 警示”还建议在适当的时候切换至“宽”方法。

- 逐行估计用于不含缺失值的数据表。
- 配对估计用于包含缺失值并且多于 10 列、多于 5,000 行或列数多于行数的数据表。
- 在其他情况下使用 REML 估计。

REML 限制最大似然 (REML) 估计使用所有数据，即使存在缺失值。由于存在偏倚修正因子，若您的数据集很大且包含许多缺失值，该方法会很慢。因此，REML 最适用于较小的数据集。若数据中不含缺失单元格，则 REML 和 ML 估计值等价于样本协方差矩阵。若存在缺失单元格，与 ML 估计相比，REML 的方差和协方差估计值的偏倚更小。有关统计详细信息，请参见“REML”。

ML 最大似然 (ML) 估计使用所有数据，即使存在缺失值。由于 ML 的估计值生成速度更快，该方法最适用于包含缺失数据的大数据表。

稳健 稳健估计使用所有数据，即使存在缺失值。该方法降低了极值的权重，因此最适用于可能具有离群值的数据表。有关统计详细信息，请参见“稳健”。

逐行 逐行估计为每对列计算 Pearson 相关性系数。有关统计详细信息，请参见“Pearson 乘积矩相关系数的统计详细信息”。逐行估计不使用包含缺失值的观测。该方法可用于排除包含缺失数据的所有观测。

配对 配对估计使用所有数据，即使存在缺失值。该方差估计方法使用这两列中不含缺失值的所有观测为每对列计算 Pearson 相关性系数。有关统计详细信息，请参见“Pearson 乘积矩相关系数的统计详细信息”。配对估计最适用于包含缺失值并且列数多于行数、多于 10 列或多于 5,000 行的数据表。

- 若您选择 “REML”、“ML”或 “稳健”，而您的数据表中的列数多于行数并且具有缺失值， JMP 会将 “方差估计”切换为 “配对”。
- 若您选择“稳健”而您的数据表中的列数多于行数并且不含缺失值，则 JMP 会将“方差估计”切换为 “逐行”。
- 若数据表超过 500 列且列数多于行数，则无论最初选择哪种方法， JMP 都会将 “方差估计”切换为 “宽”。

注意：对于数据表超过 500 列并且列数多于行数的情况，“JMP 警示”窗口会建议使用宽估计方法。这是因为在列数过多时使用其他方法，计算时间会相当长。点击**宽方法（快速）**切换为宽估计方法，或点击**默认方法 （慢速）**使用您最初选定的方法。

成分数 （仅当将 “宽数据”指定为 “方法系列”时才可用。）指定要估计的成分数。通常，成分数远小于数据的维。

指定 使用“截断 SVD”估计方法估计指定数量的成分。“截断 SVD”估计使用所有数据，即使存在缺失值。该估计方法使用基于部分奇异值分解的算法，它只计算第一个指定数量的奇异值和奇异值向量。该算法避免计算协方差矩阵以及不必要的主成分，因此计算效率较高，当数据稀疏（即包含很多零时）或当数据中具有大量列时，该算法非常有用。有关统计详细信息，请参见“截断 SVD”。

注意：这在 JMP 17 之前被称为“稀疏”估计方法。

全部 使用“完全 SVD”估计方法估计所有成分。“完全 SVD”估计不使用含缺失值的观测，因此会排除包含缺失单元格的行。该估计方法使用基于完整奇异值分解的算法。该算法避免计算协方差矩阵，因此计算效率较高，适用于数据中列非常多的情况。有关统计详细信息，请参见“完全 SVD”。

注意：这在 JMP 17 之前被称为“宽”估计方法。

缺失值插补 （仅当将 “宽数据”指定为 “方法系列”时才可用。）通过矩阵完成来插补缺失值。

特殊方法 （仅当将 “宽数据”指定为 “方法系列”并且有指定数量的成分需要估计时才可用。）提供用于计算指定数量的成分的其他方法。

快速近似 使用“随机化奇异值分解”估计指定数量的成分。请参见“随机化 SVD”。

稳健 PCA 使用一系列奇异值分解和阈值步骤来分解数据矩阵，从而估计指定数量的成分。该方法也用在“探索离群值”平台中。有关“稳健 PCA”方法的详细信息，请参见《预测和专业建模》。

缺失数据

在“主成分”平台中，处理缺失数据的方式取决于方差估计方法。您还可以采用以下方式在平台之外估计缺失值：

- 使用**多元方法 > 多元**下面的“补缺缺失数据”选项。请参见“补缺缺失数据”。
- 使用**分析 > 筛选 > 探索缺失值**下面的“多元正态补缺”或“多元 SVD 补缺”实用工具。请参见《预测和专业建模》。

“主成分”报表

初始报表类型取决于您在启动窗口中选择的方差估计方法。对于多数方法，都会显示“主成分: 基于相关性”报表。对于“宽”和“稀疏”方法，则显示特定的报表。

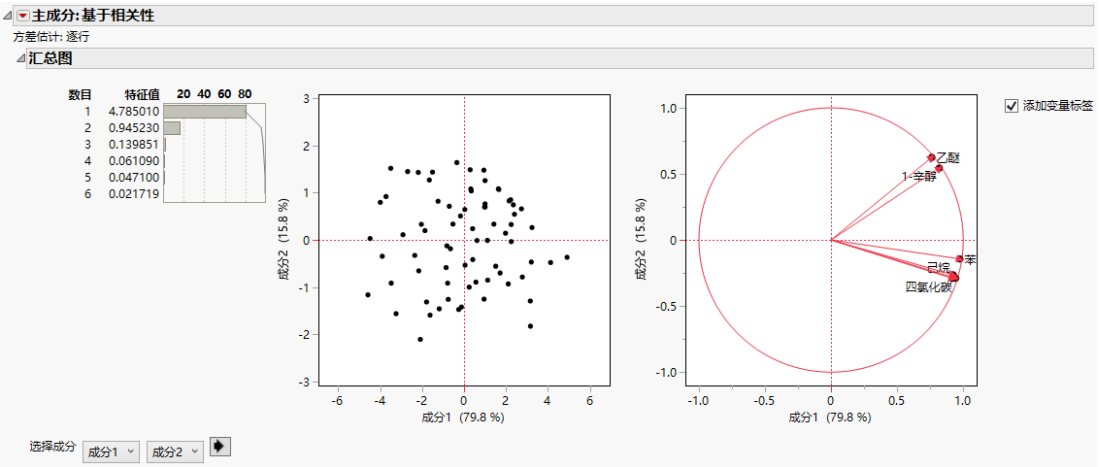
“主成分: 基于相关性”报表使用主成分汇总了指定 *Y* 变量的变异（图 4.4）。通过从红色小三角菜单中选择“主成分”选项，您可以切换到基于协方差矩阵或未统一尺度且未中心化数据的分析。

根据您的选择，主成分可以由以下任一矩阵的特征值分解推导得出：

- 相关性矩阵
- 协方差矩阵
- 未统一尺度和未中心化数据的平方和与叉积矩阵

报表中的详细信息显示主成分如何吸收数据变异。主成分点从变量的特征向量线性组合推导得出。

图 4.4 “主成分 : 基于相关性”报表



该报表提供特征值和一个条形图，图中显示了每个主成分所解释的变异百分比。同时还有一个“得分图”和一个“载荷图”。特征值指示基于每个成分贡献的方差量提取的总成分数。

“得分图”绘制每个成分相对于其他成分的计算值，按照均值和标准差调整了每个值。

“载荷图”绘制变量与成分之间的未旋转载荷矩阵。值越接近 1，成分对变量的影响越大。

默认情况下，报表为前两个主成分显示“得分图”和“载荷图”。使用“选择成分”旁边的列表指定在“得分图”和“载荷图”上绘制的主成分。

“主成分”报表选项

“主成分”红色小三角菜单包含以下选项：

注意：某些选项对于“宽”方差估计方法或“稀疏”方差估计方法不可用。

主成分 （对 “宽”方差估计方法或 “稀疏”方差估计方法不可用。 ）支持您基于相关性、协方差或未统一尺度且未中心化创建主成分。

相关性 （对 “宽”方差估计方法或 “稀疏”方差估计方法不可用。）变量之间的相关性矩阵。

注意：对角线上的值为 1.0。

协方差矩阵 （对 “宽”方差估计方法或 “稀疏”方差估计方法不可用。 ）显示或隐藏变量的协方差。

特征值 按顺序从大到小列出与各个主成分对应的特征值。特征值表示多元样本中总变异的一部分。

特征值的尺度取决于您选择用于提取主成分的矩阵：

– 对于“基于相关性”选项，特征值已统一尺度为总和等于变量数。

– 对于“基于协方差”选项，特征值不统一尺度。

– 对于“未统一尺度且未中心化”选项，特征值需要除以总观测数。

若从红色小三角菜单中选择 **Bartlett 检验**选项，将为每个特征值提供假设检验 （图 4.6） (Jackson, 2003)。

图 4.5 特征值

特征值

数目	特征值	百分比	20 40 60 80	累积百分比
1	4.785010	79.750		79.750
2	0.945230	15.754		95.504
3	0.139851	2.331		97.835
4	0.061090	1.018		98.853
5	0.047100	0.785		99.638
6	0.021719	0.362		100.000

特征向量 按顺序从左到右显示或隐藏每个主成分的特征向量的表。使用这些系数形成原始变量的线性组合，可生成主成分变量。根据标准规范，特征向量的模为 1。

注意：显示的特征向量数等于相关性矩阵的秩或在启动窗口中所指定的成分数（若选择“稀疏”方法）。

Bartlett 检验 （对“宽”方差估计方法或“稀疏”方差估计方法不可用。）显示或隐藏齐性检验的结果 （追加至 “特征值”表）。该检验通过计算检验的卡方、自由度 (DF) 和 *p* 值 （概率 > 卡方）来确定特征值是否具有相同的方差。请参见 Bartlett (1937, 1954)。

图 4.6 Bartlett 检验

特征值

数目	特征值	百分比	20 40 60 80	累积百分比	卡方	自由度	概率>卡方
1	4.785010	79.750		79.750	701.245	11.243	<.0001*
2	0.945230	15.754		95.504	317.186	13.125	<.0001*
3	0.139851	2.331		97.835	58.444	9.270	<.0001*
4	0.061090	1.018		98.853	17.589	5.280	0.0044*
5	0.047100	0.785		99.638	9.723	1.899	0.0069*
6	0.021719	0.362		100.000	0.000	.	.

载荷矩阵 显示或隐藏每个成分的载荷表。表值的透明度指示绝对载荷值与 0 之间的距离。越靠近 0 的绝对载荷值比远离 0 的绝对载荷值更加透明。该选项还显示每个成分的载荷图。有一个按钮支持您在水平条和垂直条之间旋转。“载荷矩阵图”红色小三角菜单包含以下选项：

图选择 指定成分的载荷在图中的显示方式。“概述”选项显示指定数量的成分的载荷。“单值”选项显示一个选定成分的载荷。

条样式 将图中的条样式设置为并排或堆叠填充。

若指定补充变量，则会为每个补充连续变量和补充分类变量的每个水平显示一个附加坐标表。这些值绘制到连续补充变量的载荷图中。

载荷和坐标的尺度取决于您选择用于提取主成分的矩阵：

- 对于“基于相关性”选项，载荷的第 *i* 个列是第 *i* 个特征向量乘以第 *i* 个特征值的平方根。第 *i,j* 个载荷是第 *i* 个变量与第 *j* 个主成分之间的相关性。
- 对于“基于协方差”选项，载荷的第 *i* 个列中的第 *j* 个条目是第 *i* 个特征向量乘以第 *i* 个特征值的平方根再除以第 *j* 个变量的标准差。第 *i,j* 个载荷是第 *i* 个变量与第 *j* 个主成分之间的相关性。
- 对于“未统一尺度且未中心化”选项，载荷的第 *i* 个列中的第 *j* 个条目是第 *i* 个特征向量乘以第 *i* 个特征值的平方根再除以第 *j* 个变量的标准误差。第 *j* 个变量的标准误差是平方和与叉积矩阵的第 *j* 个对角线元素除以行数 (**X′X**/n)。

注意：分析未统一尺度且未中心化的数据时，第 *i,j* 个载荷**不**是第 *i* 个变量与第 *j* 个主成分之间的相关性。

格式化的载荷矩阵 显示或隐藏每个成分的载荷表。该表按第一个主成分的载荷降序排序。因此，变量按照第一个成分的载荷降序列出。

图 4.7 格式化的载荷矩阵

格式化的载荷矩阵

	主成分1	主成分2	主成分3	主成分4	主成分5	主成分6
苯	0.974761	-0.143460	-0.081914	-0.023369	-0.108186	-0.101488
四氯化碳	0.942849	-0.289099	0.069134	-0.059654	-0.099757	0.095746
己烷	0.923483	-0.263641	0.256572	0.026770	0.099672	-0.034523
氯仿	0.917422	-0.290351	-0.242516	0.075629	0.093454	0.027695
1-辛醇	0.819019	0.544318	-0.041397	-0.162739	0.068711	0.002762
乙醚	0.761978	0.625283	0.044774	0.155130	-0.045337	0.016883

隐藏小于该值的绝对载荷值: 0.3
文本变暗 0.4

隐藏小于该值的绝对载荷值 该值确定哪些载荷在“格式化的载荷矩阵”报表中不可用。您可以使用文本框或滑块将绝对值落在选定值之下的载荷灰显。

文本变暗 “格式化的载荷矩阵”报表中的灰显值的透明度。您可以使用文本框或滑块设置灰显载荷的透明度。透明度的范围从 0 到 1，值越小越透明。例如，将透明度设置为 0 时完全将不可用载荷从矩阵中移除，设置为 1 时载荷仍可看到。

变量的平方余弦 显示或隐藏包含变量的平方余弦的表。若指定补充变量，则会显示补充变量平方余弦的附加表。对于每个变量，各主成分的平方余弦值的总和等于 1。平方余弦使您能够查看主成分表示变量的好坏程度。您还可以确定表示特定变量需要多少个主成分。该选项还显示这些成分的平方余弦图。有一个按钮支持您在水平条和垂直条之间旋转。“变量的平方余弦图”红色小三角菜单包含以下选项：

图选择 指定成分的平方余弦在图中的显示方式。“概述”选项显示指定数量的成分的平方余弦。“单值”选项显示一个选定成分的平方余弦。

条样式 将图中的条样式设置为并排或堆叠填充。

注意：若使用 “稀疏”方差估计方法并且选择的成分数小于 3，则在图上只显示指定的成分数。

变量的部分贡献 显示或隐藏包含变量的部分贡献的表。部分贡献使您能够查看每个变量对每个主成分的贡献百分比。该选项还显示这些成分的部分贡献图。有一个按钮支持您在水平条和垂直条之间旋转。“变量的部分贡献图”红色小三角菜单包含以下选项：

图选择 指定成分的部分贡献在图中的显示方式。“概述”选项显示指定数量的成分的部分贡献。“单值”选项显示一个选定成分的部分贡献。

条样式 将图中的条样式设置为并排或堆叠填充。

注意：若使用 “稀疏”方差估计方法并且选择的成分数小于 3，则在图上只显示指定的成分数。

汇总图 显示或隐藏在默认报表中生成的汇总信息。该汇总信息包括特征值图、得分图和载荷图。默认情况下，报表为前两个主成分显示得分图和载荷图。报表中提供选项，可用于指定要绘制的主成分。请参见 “‘主成分’报表”。

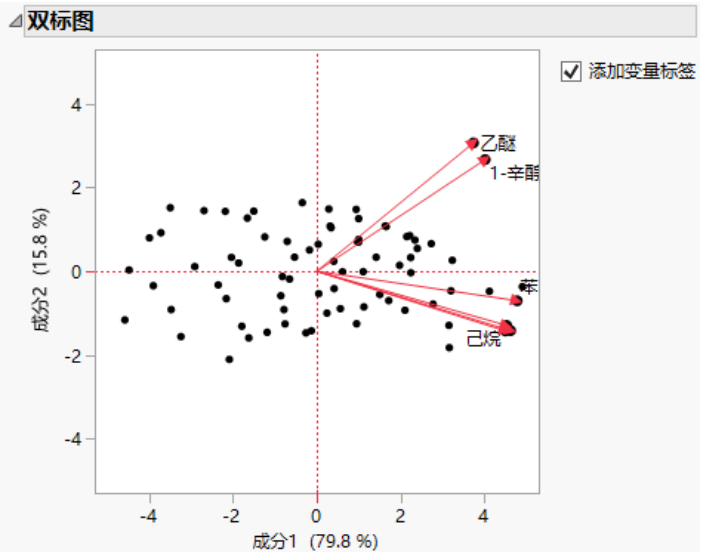
注意：若您的数据包含缺失值，则插补得分将绘制在得分图上。

提示：选择载荷图中的箭头尖端可选择数据表中对应的列。按 Ctrl 键并点击某个箭头尖端可取消选择相应的列。

双标图 显示或隐藏一个图，用来叠加指定成分数的得分图和载荷图。

注意：若您的数据包含缺失值，则插补得分将绘制在 “双标图”上。

图 4.8 双标图

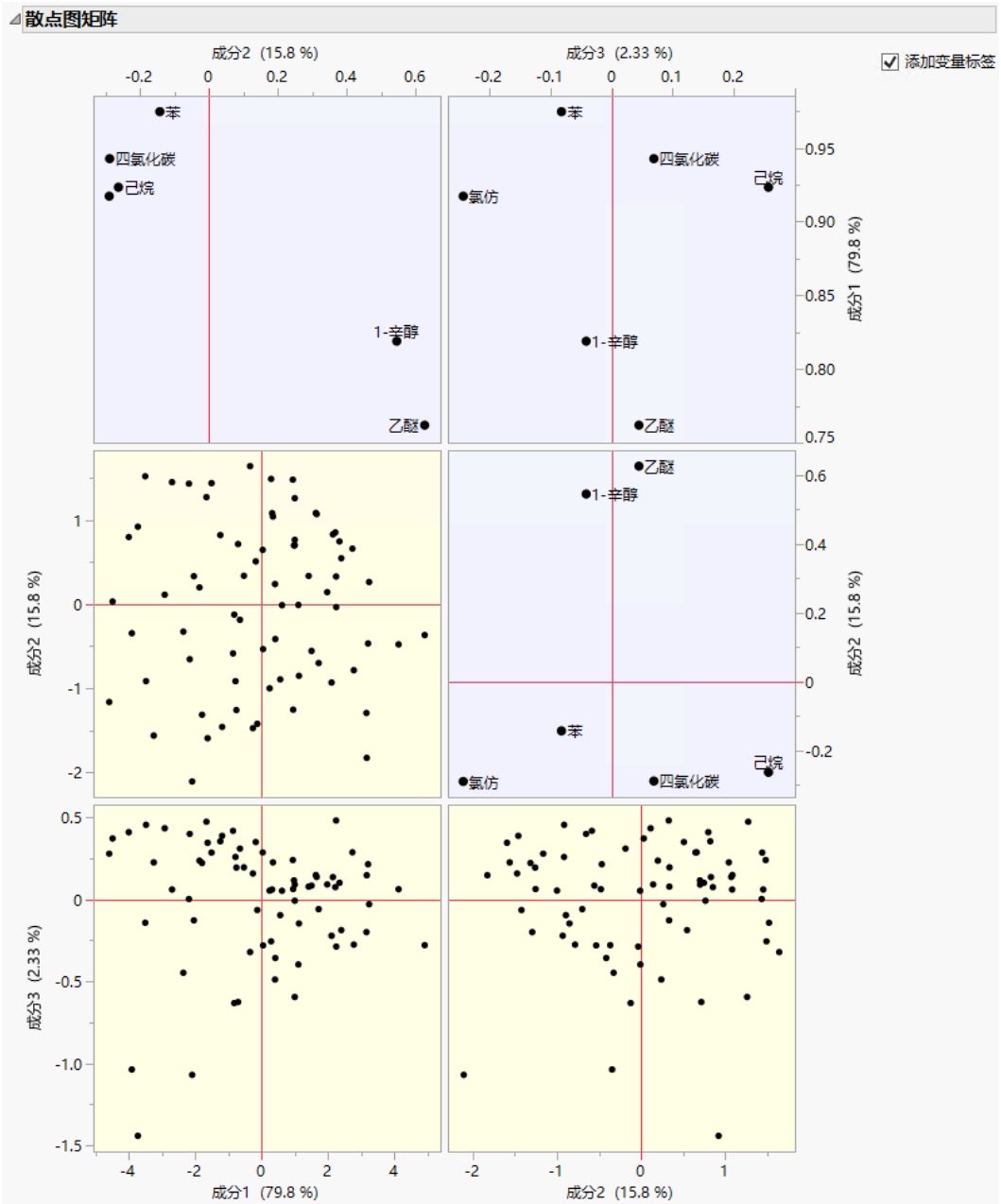


注意：得分图标记为圆点，载荷图标记为菱形。

散点图矩阵 显示或隐藏指定主成分数的得分图和载荷图的矩阵。散点图矩阵在一个空间内排列得分图和载荷图。得分图具有黄色阴影背景。载荷图具有蓝色阴影背景。

注意：若您的数据包含缺失值，则插补得分将绘制在 “散点图矩阵”中。

图 4.9 散点图矩阵



注意：显示在“散点图矩阵”中的载荷图矩阵是您选择“载荷图”选项时获得的载荷图矩阵的转置矩阵。

陡坡图 显示或隐藏每个成分的特征值图。该陡坡图帮助直观演示数据空间的维度。

得分图 显示或隐藏指定成分数的成对主成分的得分散点图矩阵。该图显示在图 4.4 中（最左侧图）。

载荷图 显示或隐藏指定成分数的因子载荷的二维表示矩阵。若变量数未超过 30 个，则载荷图会对变量添加标签。若变量数超过 30 个，则默认不显示标签。该信息显示在图 4.4 中（最右侧图）。

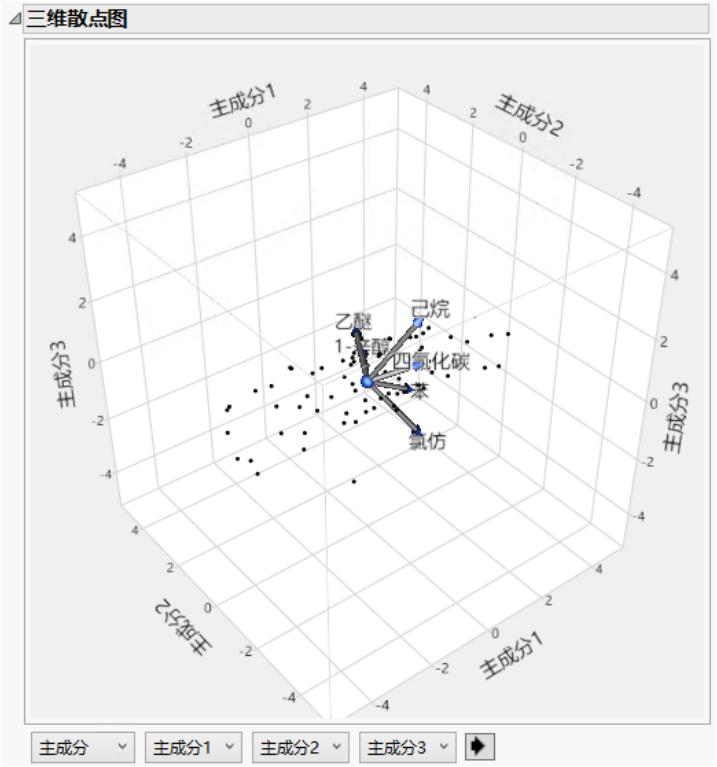
提示：选择载荷图中的箭头尖端可选择数据表中对应的列。按 Ctrl 键并点击某个箭头尖端可取消选择相应的列。

完成补缺的得分图 （对“宽”方差估计方法或“稀疏”方差估计方法不可用。）补缺所有缺失值并创建得分图。仅当包含缺失值时，该选项才可用。

三维得分图 （对“宽”方差估计方法或“稀疏”方差估计方法不可用。）显示或隐藏任何三个主成分得分的三维散点图。首次调用该命令时，将显示前三个主成分。

注意：若您的数据包含缺失值，则插补得分将绘制在“三维得分图”上。

图 4.10 三维散点图



图来源 图中数据点的来源。可用选项为“主成分”、“旋转主成分”和“数据列”。

轴控件 每个轴的内容。若选择“主成分”选项或“旋转成分”选项，则“轴控件”的选项为“主成分”。若选择“数据列”选项，则选项为分析中的变量。

循环按钮 在所有轴内容可能性中循环。

变量在图中显示为射线。这些射线，称为**双标图射线**，将变量近似计算为轴上主成分的函数。若只有两三个变量，射线就正好代表变量。射线对应于主成分载荷。

得分椭圆 在每对主成分的汇总得分图上显示或隐藏椭圆。椭圆既可以构造为基于 alpha 水平的置信椭圆，也可以构造为基于观测距中心的距离的控制限椭圆。默认情况下，这些椭圆为 95% 置信椭圆。

得分椭圆覆盖率 显示一个支持您更改得分椭圆的构造方式的子菜单。按照置信水平或按照用 *k* sigma 表示的距中心的距离来指定得分椭圆。置信水平 *p* 与 *k* sigma 之间的关系为 $p = 1 - \exp(-k^2/2)$。

显示选项

箭头线 允许您在可显示箭头的所有图上显示或隐藏箭头。若变量数不超过 1000 个，则显示箭头。若变量数超过 1000 个，则默认不显示箭头。

显示补充变量 （仅当您指定补充变量时可用。）在双标图、得分图和载荷图中，显示或隐藏连续补充变量的箭头线或分类补充变量的标签标记。

离群值分析 显示或隐藏“离群值分析”报表，该报表支持您通过 T^2 和贡献统计量检测数据中的离群值。请参见 “离群值分析”。

因子分析 （对“宽”方差估计方法或“稀疏”方差估计方法不可用。）用主成分法进行因子旋转得到的因子分析，或者其他方法得到的因子分析。请参见 “因子分析”。

JMP PRO **聚类变量** （对“宽”方差估计方法或“稀疏”方差估计方法不可用。）通过将变量划分为不重叠的聚类，对这些变量执行聚类分析。变量聚类提供将类似变量分组到代表组中的方法。每个聚类随后可由单个成分或变量表示。成分是聚类中所有变量的线性组合。或者，聚类也可由标识为聚类中最典型成员的变量来表示。请参见 “聚类变量”。

注意：“聚类变量”对所有计算都使用相关性矩阵，即便您选择“基于协方差”或“基于未统一尺度且未中心化”选项也是如此。

模型驱动的多元控制图 保存指定数量的主成分的公式，并启动“模型驱动的多元控制图”(MDMCC) 启动窗口。在 MDMCC 启动窗口中，主成分公式作为过程列分配。在点击“确定”之前，可以添加或删除过程、添加时间 ID 或设置历史数据的结束位置。请参见《质量和过程方法》。

预测值的刻画器 使用每个变量的预测公式启动“刻画器”启动窗口。预测公式针对每个变量，并使用刻画器启动之前所指定的指定数量的主成分。您可以在点击“确定”之前在启动窗口中添加噪声因子或其他预测公式。请参见《刻画器指南》。

保存列 提供保存选项子菜单。

保存主成分 将您指定的主成分数与用于计算每个成分的公式一同保存到数据表中。该公式无法计算包含缺失值的行。

主成分的计算取决于您选择用于提取主成分的矩阵：

- 对于“基于相关性”选项，第 *i* 个主成分是将第 *i* 个特征向量的条目用作系数的中心化和统一尺度观测的线性组合。
- 对于“基于协方差”选项，第 *i* 个主成分是将第 *i* 个特征向量的条目用作系数的中心化观测的线性组合。
- 对于“未统一尺度且未中心化”选项，第 *i* 个主成分是将第 *i* 个特征向量的条目用作系数的原始观测的线性组合。

注意：若指定的成分数超过相关性矩阵的秩，则保存的成分数设置为相关性矩阵的秩。

保存主成分值 将您指定的主成分数保存到数据表中。这些列不是公式列，其中只包含成分的值。

提示：对宽数据使用该选项。

保存低秩主成分 （仅当在宽数据或窄数据的启动窗口中指定“稳健 PCA”估计方法时才可用。）保存低秩矩阵中的主成分得分，其中已清除离群值和噪声。

保存预测值 将预测变量（给定了指定数量的主成分）保存到数据表中的新列。

将预测值另存为成分公式 将预测变量保存为公式，这些公式是指定数量的成分的线性组合。公式中使用的主成分也保存到数据表中。

保存标准化 DModX 将到主成分模型的观测距离 (DModX)（给定了指定数量的主成分）保存到数据表的新列中。较大的 DModX 值表示数据中存在轻中度的离群值。请参见《质量和过程方法》。

保存单值平方余弦 将单值平方余弦（给定了指定数量的主成分）保存到数据表中的新列。

保存单值部分贡献 将单值部分贡献（给定了指定数量的主成分）保存到数据表中的新列。

保存旋转成分 （对“宽”方差估计方法或“稀疏”方差估计方法不可用。）将旋转成分与用于计算成分的公式一同保存到数据表中。该选项仅在使用“因子分析”选项后可用。该公式无法计算包含缺失值的行。

保存完成补缺的主成分 （对“宽”方差估计方法或“稀疏”方差估计方法不可用。）补缺缺失值，并将主成分保存到数据表中。该列包含用于执行补缺并计算主成分的公式。仅当包含缺失值时，该选项才可用。

保存完成补缺的旋转成分 （对“宽”方差估计方法或“稀疏”方差估计方法不可用。）补缺缺失值，并将旋转成分保存到数据表中。该列包含用于执行补缺并计算旋转成分的公式。使用“因子分析”选项后，若存在缺失值，该选项才可用。

保存插补公式 （仅当数据表包含缺失值时才可用。）对于包含缺失值的列，将包含用于估计缺失值的公式的新列保存到数据表中。新列称为**补缺_<列名>**。

发布成分公式 创建指定数量的主成分公式并在“公式存储库”平台中将它们保存为公式列脚本。若未打开“公式存储库”报表，该选项将创建“公式存储库”报表。请参见《预测和专业建模》。

发布标准化 DModX 公式 将基于指定数量的主成分的“标准化 DModX”公式另存为“公式存储库”平台中的公式列脚本。若未打开“公式存储库”报表，该选项将创建“公式存储库”报表。请参见《预测和专业建模》。

请参见《使用 JMP》获取有关下列选项的信息：

本地数据过滤器 显示或隐藏支持您过滤特定报表中使用的数据的本地数据过滤器。

重新运行 包含使您可以重复或重新启动分析的选项。在支持该功能的平台中，“自动重新计算”选项立即在相应报表窗口中反映您对数据表所做的更改。

平台首选项 包含的选项支持您查看当前平台首选项或更新平台首选项以匹配当前 JMP 报表中的设置。

保存脚本 包含的选项支持您保存可将报表重现到若干目标的脚本。

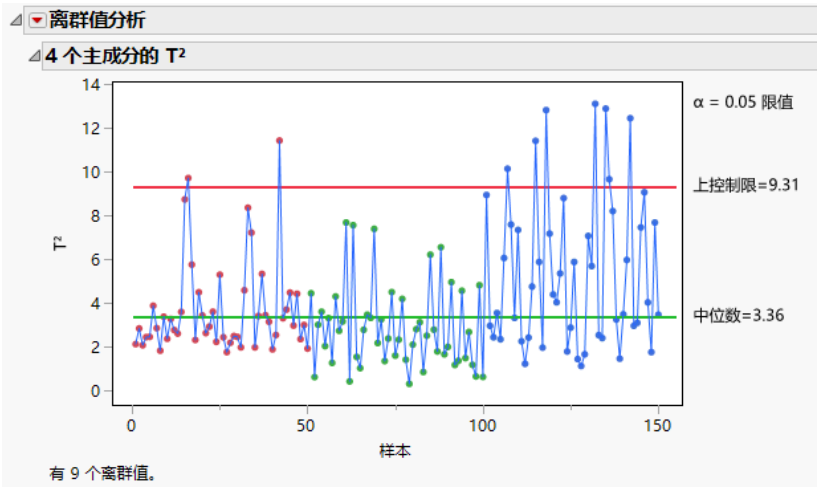
保存“依据”组脚本 包含使您可以保存脚本的选项，可将为“依据”变量的所有水平重新生成平台报表的脚本保存到多个不同的位置。仅当在启动窗口中指定“依据”变量时才可用。

离群值分析

在“主成分”平台中，默认情况下，“离群值分析”报表显示“<*A*> 个主成分的 T^2”图。该图显示每个观测的 T^2 值以及位于中位数和上控制限 (UCL) 处的水平线。有关如何计算 T^2 值、中位数和上控制限的详细信息，请参见“离群值分析计算的统计详细信息”。

用于计算上控制限的 α 水平显示在图旁边。检测到的离群值数目显示在图下方。该数字是 T^2 值大于上控制限的观测数。

图 4.11 “离群值分析”报表



提示：悬停在 T^2 图中的某个点上方以查看该观测的 T^2 贡献比例图。点击 T^2 贡献比例图将其添加到报表窗口。

“离群值分析”报表选项

“离群值分析”红色小三角菜单包含以下选项：

T^2 图 显示或隐藏 T^2 图。默认开启。

贡献热图 为所有观测显示或隐藏 T^2 贡献值的热图。

贡献比例热图 显示或隐藏所有观测 （表示为各行的 T^2 的比例）的 T^2 贡献值的热图。通过计算贡献的平方再除以对单个行的所有贡献的平方和，可以得到该单个行的比例。

选定样本的贡献图 （仅当在 T^2 图中选定一个或多个点时才可用。）显示一个报表，其中为每个选定的样本都显示一个 T^2 贡献图。 T^2 贡献图显示每个变量对样本的 T^2 统计量的贡献。有关如何计算贡献的详细信息，请参见《质量和过程方法》。使用贡献图可调查离群值。具有最大正 / 负贡献的变量是那些对样本贡献最大 （具有较大 T^2 值）的变量。请参见 “‘贡献图”报表选项”了解有关红色小三角菜单选项的信息。

选定样本的贡献比例图 （仅当在 T^2 图中选定一个或多个点时才可用。）显示一个报表，其中为每个选定的样本都显示一个 T^2 贡献比例图。 T^2 贡献比例图显示选定观测的贡献值，这些值表示为各行的 T^2 的比例。这是 “贡献比例热图”中所示信息的一个不同的演示。请参见 “‘贡献图”报表选项”了解有关红色小三角菜单选项的信息。

标准化 DModX 图 （仅当成分数少于变量数时才可用。）显示或隐藏标准化 DModX 值的图。DModX 值可用于检测数据中的中度离群值。

成分数 支持您指定在 T^2 和 T^2 贡献统计量中使用的主成分数。更改成分数时， T^2 图、热图和标准化 DModX 图自动更新。

设置 α 水平 支持您指定 α 水平。

保存 T^2 将 T^2 值保存到数据表中的新列。

保存贡献 将 T^2 贡献保存到数据表中的新列。每个 Y 变量都有一列。

保存标准化 DModX （仅当成分数少于变量数时才可用。）在数据表的新列中保存标准化 DModX 值。

“贡献图”报表选项

选定样本的 T^2 贡献图、选定样本的 T^2 贡献比例图和选定样本的 T^2 均值贡献比例图包含以下红色小三角菜单选项。

直条标签 显示用于为贡献图中的直条添加标签的选项子菜单。标签选项包括“无标签”、“将值用作标签”和“将列用作标签”。

删除图 从报表中删除贡献图。

选定项的控制图 （仅当在贡献图中选择了一个或多个条形段时才可用。）打开“控制图生成器”窗口，其中包含每个选定过程和组的控制图结果。

提示：也可以通过悬停在贡献图中的条形段上方来查看控制图。点击控制图打开“控制图生成器”窗口。

“主成分”平台的统计详细信息

本节包含“主成分”平台的统计详细信息。

- “方差估计方法的统计详细信息”
- “宽方法的统计详细信息”
- “离群值分析计算的统计详细信息”

方差估计方法的统计详细信息

本节包含“主成分”平台中使用的方差估计方法的统计详细信息。

REML

当数据包含缺失值时，与 ML（最大似然）估计方法相比，REML（限制最大似然）估计值的偏倚更小。REML 方法基于误差对比将边缘似然最大化。REML 方法经常用于估计方差和协方差。“主成分”平台中的 REML 方法与针对重复测量数据（具有非结构化协方差矩阵）使用的混合模型的 REML 估计相同。请参见 SAS PROC MIXED 文档，了解有关混合模型的 REML 估计。

宽方法的统计详细信息

本节包含在“主成分”启动窗口中将“宽数据”指定为“方法系列”时使用的估计方法的统计详细信息。这些方法基于奇异值分解 (SVD)。这避免了计算协方差矩阵，从而使这些方法成为计算高效的算法。

完全 SVD

“完全 SVD”方法使用基于矩阵的完全奇异值分解的算法。考虑以下符号：

- n = 行数
- p = 变量数
- $\mathbf{X}$ = 数据值的 $n \times p$ 矩阵

非零特征值数以及得到的主成分数均等于 $\mathbf{X}$ 的相关性矩阵的秩。非零特征值数不得超过 n 和 p 中较小的那个值。

在实施 “完全 SVD”方法之前，数据已标准化。要将某个值标准化，需减去其均值，再除以其标准差。用 $\mathbf{X}_s$ 来表示标准化数据值的 n x p 矩阵。之后，标准化数据的协方差矩阵成为 $\mathbf{X}$ 的相关性矩阵，该矩阵定义如下：

$$Cov = \mathbf{X}_s'\mathbf{X}_s/(n-1)$$

使用奇异值分解，$\mathbf{X}_s$ 可表示为 $\mathbf{U}Diag(\Lambda)\mathbf{V}'$。这种表示法用于获取 $\mathbf{X}_s'\mathbf{X}_s$ 的特征向量和特征值。主成分或得分通过 $\mathbf{X}_s\mathbf{V}$ 计算得出。更多背景信息，请参见“ “宽线性”方法和奇异值分解”。

JMP PRO 截断 SVD

“截断 SVD”方法使用基于奇异值分解但不执行完全分解的算法。而该算法在奇异值分解时仅计算第一个指定数量的奇异值和奇异向量。因此，只返回第一个指定数量的特征值和主成分。有关算法的详细信息，请参见 Baglama and Reichel (2005)。

随机化 SVD

“随机化 SVD”方法使用基于奇异值分解但不执行完全分解的算法。该算法是一个涉及低秩矩阵近似的两步过程。通常，低秩矩阵近似的目标是要近似计算一个 $m \times k$ 矩阵 $\mathbf{A}$，方法是查找 $m \times k$ 矩阵 $\mathbf{B}$ 和 $k \times p$ 矩阵 $\mathbf{C}$，以使 k 远远小于 p 且 $\mathbf{A} \approx \mathbf{BC}$。

考虑“完全 SVD”中所述的相同符号，其中的目标是要分解 $\mathbf{X}_s$。

在“随机化 SVD”算法的第一步中，发现了矩阵 $\mathbf{Q}$，使得以下条件成立：

- $\mathbf{Q}$ 具有 l 个正交列，其中 $k \le l \le p$
- $\mathbf{X}_s \approx \mathbf{QQ}'\mathbf{X}_s$

关于高效计算 $\mathbf{Q}$ 使其具有尽量少的列的详细信息，请参见 Halko, Martinsson, and Tropp (2011)。

算法的第二步使用 $\mathbf{Q}$ 计算 $\mathbf{X}_s$ 的奇异值分解。令 $\mathbf{B} = \mathbf{Q}'\mathbf{X}_s$。然后，以下条件成立：

$$\mathbf{B} = \mathbf{Q}'\mathbf{X}_s \Rightarrow \mathbf{QB} = \mathbf{QQ}'\mathbf{X}_s \Rightarrow \mathbf{QB} = \mathbf{X}_s$$

然后，计算 $\mathbf{B}$ 的 SVD（远远小于 $\mathbf{X}_s$）：

$$\mathbf{B} \approx \mathbf{U}^* Diag(\Lambda)\mathbf{V}'$$

$$\mathbf{QB} \approx (\mathbf{QU}^*) Diag(\Lambda)\mathbf{V}'$$

$$\mathbf{A} \approx \mathbf{U}Diag(\Lambda)\mathbf{V}', \text{ 其中 } \mathbf{U} = \mathbf{QU}^*$$

有关完整的“随机化 SVD”算法的详细信息，请参见 Halko, Martinsson, and Tropp (2011)。

离群值分析计算的统计详细信息

在 “主成分”平台中，“离群值分析”报表中的计算使用以下符号：

n = 观测数

A = 主成分数

X_{ci} = 第 i 个观测的标准化数据

T^2 统计量

按以下方式计算第 i 个观测的 T^2 统计量：

$$T_i^2 = X_{ci} P_A L^{-1} P_A^{\mathrm{T}} X_{ci}^{\mathrm{T}}$$

其中，P_A 是包含第一个 A 特征向量的矩阵，L 是包含第一个 A 特征值的对角线矩阵。

T^2 图的中位数和上控制限计算如下：

$$CL_{T^2,\,q} = \frac{(n-1)^2}{n}\beta\left[q, \frac{A}{2}, \frac{n-A-1}{2}\right]$$

其中

$$\beta\left[q, \frac{A}{2}, \frac{n-A-1}{2}\right] = \text{Beta}\left(\frac{A}{2}, \frac{n-A-1}{2}\right) \text{ 分布的第 } q \text{ 分位数}$$

要计算中位数，请使用 $q = 0.5$。要计算上控制限，请使用 $q = (1 - \alpha)$。

有关贡献统计量的信息，请参见《质量和过程方法》。

第 5 章

判别分析

基于连续变量预测分类

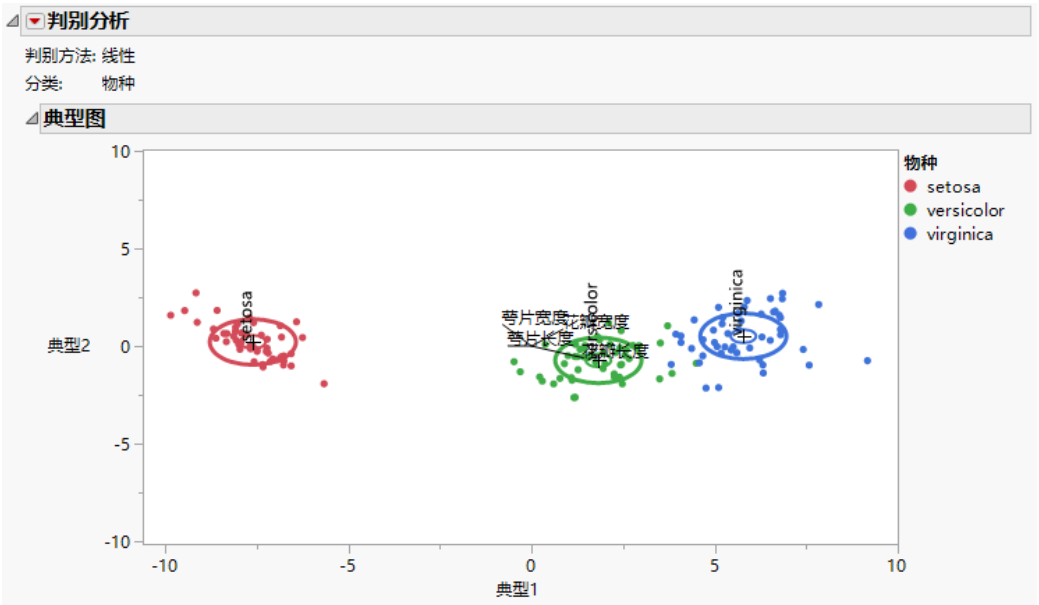
判别分析基于几个连续变量的观测值预测一个组或一个类别的成员关系。具体来说，判别分析基于已知的连续响应 (Y) 来预测一个分类 (X) 变量（分类）。判别分析的数据包含具有已知组成员关系的观测样本以及它们对应连续变量的值。

例如，您可能尝试基于期望概率将贷款申请人分为三个信贷类别 (X)：低利率贷款、长期贷款或无贷款。您可能使用一些连续变量（如当前工资、当前工作年数、年龄和债务负担）(Y) 来预测个人的最大利润贷款类别。您可以使用判别分析生成一个预测模型，从而将个人归为某个贷款类别。

“判别”平台的功能包括：

- 一个逐步选择选项来帮助选择能很好进行判别的变量。
- 拟合方法选项：“线性”、“二次”、“正则”和“宽线性”。
- 一个典型图和误分类汇总。
- 判别得分以及到各组的平方距离。
- 用于将预测距离和概率保存到数据表的选项。

图 5.1 典型图



目录

“判别”平台概述

判别分析尝试将由连续变量值描述的一系列观测值进行分组。由分类变量 *X* 定义的组成员关系由连续变量来预测。这些变量称为**协变量**并用 *Y* 表示。

判别分析不同于 Logistic 回归。在 Logistic 回归中，分类变量是随机的并由连续变量来预测。在判别分析中，分类是固定的而协变量 (*Y*) 是通过随机变量实现的。但是，在这两种方法中，分类值均由连续变量来预测。

“判别”平台提供拟合模型的四种方法。所有方法均使用 Mahalanobis 距离估计每个观测到每个组的多元均值（**重心**）的距离。您可以指定组成员关系的先验概率，在距离计算中要采用它们。观测会被归类到距离最近的组中。

拟合方法包括以下几种：

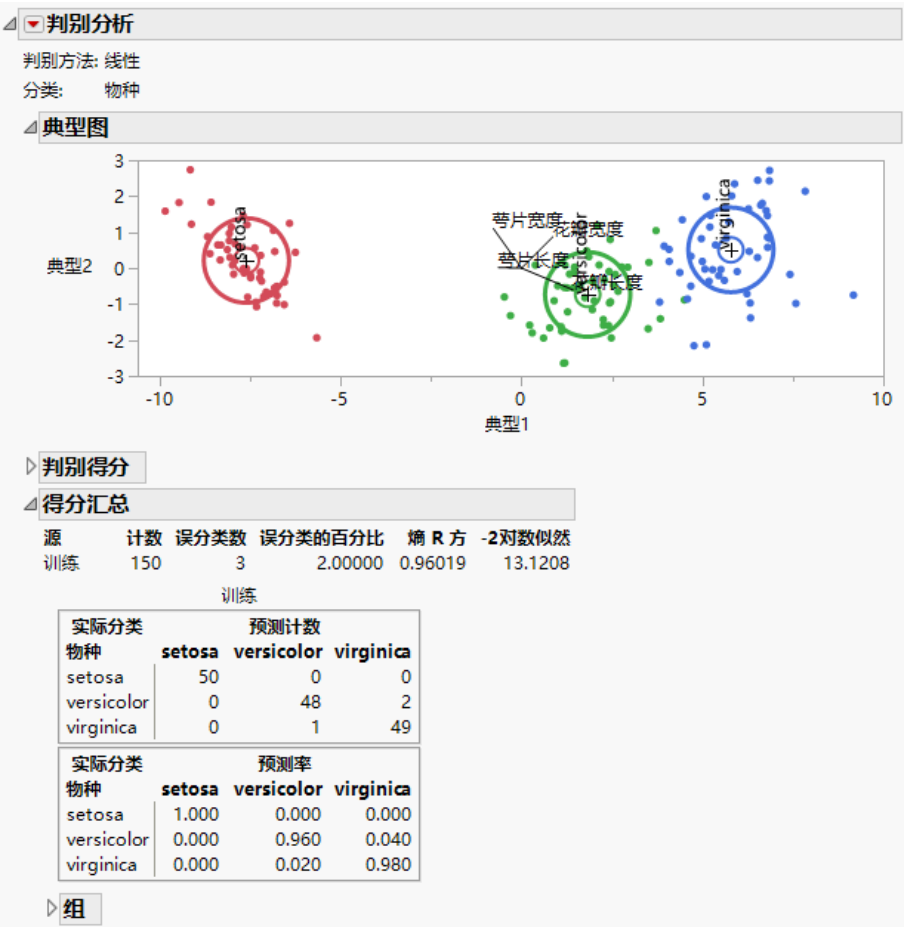
- 线性 — 假定组内协方差矩阵是相等的。假定由 *X* 定义的组的协变量均值不同。
- 二次 — 假定组内协方差矩阵不同。这要求估计比线性方法更多的参数。若组样本大小很小，则可能得到不稳定的估计值。
- 正则 — 在组内协方差矩阵不同时提供两种方法来增加估计值的稳定性。若组样本大小很小，这是很有用的选项。
- 宽线性 — 在拟合有很多协变量的模型时很有用，此时其他方法可能计算困难。它假定所有协方差矩阵是相等的。

判别分析的示例

在本例中，目标是要使用从花中获取的四个测量值来准确地识别鸢尾花的种类。

1. 选择**帮助** > **样本数据文件夹**，然后打开 Iris.jmp。
2. 选择**分析** > **多元方法** > **判别**。
3. 选择**萼片长度**、**萼片宽度**、**花瓣长度**和**花瓣宽度**，然后点击 **Y，协变量**。
4. 选择**物种**并点击 **X，类别**。
5. 点击**确定**。

图 5.2 “判别分析”报表窗口

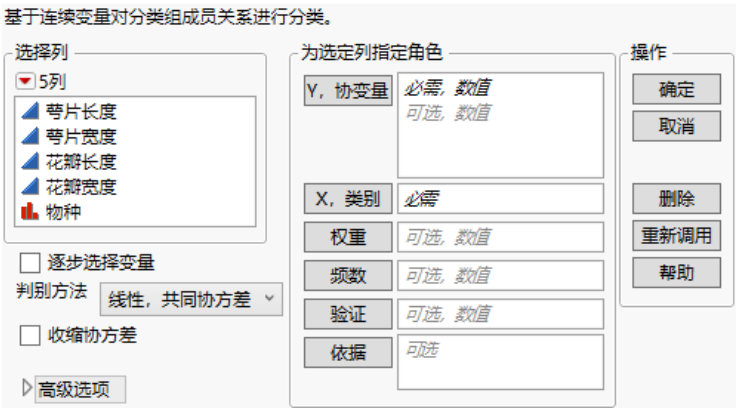


因为**物种**有三个分类，因此有两个典型变量。在“典型图”中，针对两个典型坐标绘制每个观测。该图显示这两个坐标分隔了三个物种。因为没有验证集，“得分汇总”报表只显示“训练集”的面板。没有验证集时，将整个数据集视为训练集。在 150 个观测中，只有三个被误分类。

启动“判别”平台

通过选择**分析** > **多元方法** > **判别**来启动“判别”平台。

图 5.3 Iris.jmp 的“判别”启动窗口



有关“选择列”红色小三角菜单中选项的详细信息，请参见《使用 JMP》。

注意：“验证”按钮仅显示在 JMP Pro 中。在 JMP 中，您可以使用排除行来定义验证集。请参见“JMP 和 JMP Pro 中的验证”。

Y，协变量 包含用于将观测分为各个类别的连续变量的列。

X，类别 包含将观测归入其中的类别或组的列。

权重 用于将一个权重值分配给要分析的每个行的列。

频数 用于将一个频数值分配给要分析的每个行的列。一般来说，频数列的作用是扩展数据表，以便具有整数频数 k 的任何行扩展为 k 行。您可以指定小数频数。

JMP PRO **验证** 用于定义验证集的数值列。该列应包含至多三个非重复值。

- 若有两个值，较小的值定义训练集，较大的值定义验证集。
- 若有三个值，这些值按大小递增的顺序定义训练集、验证集和测试集。
- 若验证列有三个以上的水平，则包含最小三个值的行将定义验证集。其他所有行都从分析中排除。

“判别”平台使用验证列来训练和评估模型，除非使用“逐步选择变量”。若在启动时选定“逐步选择变量”选项，则“判别”平台使用验证列来训练和微调模型，或者训练、微调和评估模型。有关验证的详细信息，请参见《预测和专业建模》。

提示：若未使用“逐步选择变量”，则验证列应只包含两个非重复值。

若在“选择列”列表中没有选择任何列的情况下点击“验证”按钮，您可以向数据表添加一个验证列。有关“生成验证列”实用工具的详细信息，请参见《预测和专业建模》。

依据 为指定列的每个水平执行单独的分析。

逐步选择变量 （将“宽线性”选作“判别方法”时不可用。）使用协方差分析和 *p* 值逐步选择变量。请参见“逐步选择变量”。

若指定了验证集，则还会显示验证集的统计量。验证集统计量用于确定使用“执行”按钮时要执行的步骤数。

判别方法 提供四种执行判别分析的方法。请参见“判别方法”。

收缩协方差 收缩合并的组内协方差矩阵和组内协方差矩阵的非对角线元素。这样可以提高预测的稳定性并且减小预测的方差。请参见“收缩协方差”。

高级选项 包含以下选项：

未中心化典型 禁止典型得分中心化以便与较旧版本的 JMP 兼容。

使用伪逆元素 当协方差矩阵是奇异矩阵时，在分析中使用 Moore-Penrose 伪逆元素。所得的分数涉及所有协变量。若未选中它，则分析会删除以下协变量：这些协变量是在 Y，**协变量**列表中位于它们之前的协变量的线性组合。

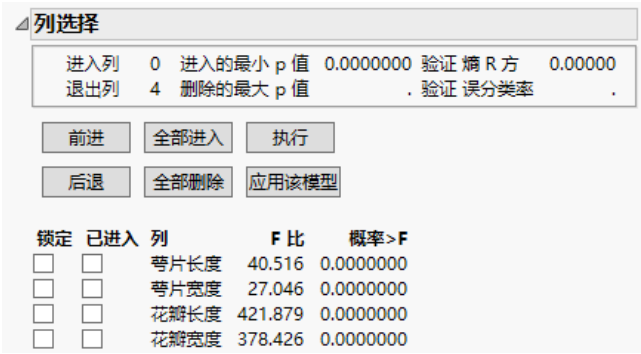
按排除行交叉验证 指定排除行将形成计算其拟合统计量的验证集。

逐步选择变量

若您在启动窗口中选择了“逐步选择变量”选项，“判别分析”报表会包含“列选择”面板。您可以执行逐步分析，使用按钮选择变量或使用“锁定”和“已进入”复选框手动选择它们。根据您的选择会更新 *F* 比和 *p* 值。有关如何更新这些值的详细信息，请参见“更新“F 比”和“概率>F””。

若指定任何类型的验证集，则会显示“执行”按钮。当您点击“执行”时，JMP 使用验证集统计量来确定要执行的步骤数。

图 5.4 Iris.jmp 具有验证集的 “列选择” 面板



更新 “F 比” 和 “概率 >F”

当您在模型中输入变量或删除变量时，会根据对具有以下结构的协方差模型的分析更新 “F 比” 和 “概率 >F” 值：

- 考虑之中的协变量为响应变量。
- 已进入模型的协变量为预测变量。
- 组变量为预测变量。

“逐步” 报表中给出的 “F 比” 和 “概率>F” 值是组变量的协方差分析检验的 *F* 比和 *p* 值。组变量的协方差分析检验是它相对于考虑之中的协变量的判别能力的指示符。

统计量

进入列 当前已选择进入判别模型的列数。

退出列 当前可以进入判别模型的列数。

进入的最小 p 值 所有可以进入模型的协变量的 *p* 值中最小的那个 *p* 值。

删除的最大 p 值 在当前已选择进入模型的所有协变量的 *p* 值中最大的那个 *p* 值。

验证熵 R 方 验证集的熵 R 方。值越大表示拟合效果越好。“熵 R 方” 的值为 1 表示分类预测完美。由于判别模型的预测概率的不确定性很典型，因此 “熵 R 方” 值往往很小。

请参见 “熵 R 方”。仅当使用验证集时才可用。

注意：“验证熵 R 方” 可能为负数。

验证误分类率 验证集的误分类率。值越小表示分类效果越好。仅当使用验证集时才可用。

按钮

前进 使尚未进入的协变量中最显著的协变量进入。若使用了验证集，则“概率 >F”值基于训练集。

后退 从已进入但是未锁定的协变量中删除最不显著的协变量。若使用了验证集，则“概率 >F”值基于训练集。

全部进入 通过选中在“已进入”列中未锁定的所有协变量来使所有协变量进入。

全部删除 通过在“已进入”列中取消选择未锁定的所有协变量来删除它们。

应用该模型 基于在“已进入”列中选中的协变量生成判别分析报表。关闭“选择列”分级显示项，并更新“判别分析”窗口来基于选定的判别方法显示分析结果。

提示：在点击**应用该模型**后，您选择的列将显示在“得分汇总”报表的顶部。

执行 在前进步中使协变量进入，直到“验证熵 R 方”开始减小。当执行两个前进步而未改进“验证熵 R 方”时终止进入。仅当在 JMP 中具有排除行或在 JMP Pro 中具有验证列时可用。

列

锁定 强制协变量保持当前状态而不管使用按钮执行的任何步进操作。

请注意以下事项：

- 若您使一个协变量进入然后为它选择**锁定**，则它仍留在模型中而不管使用控件按钮所做的选择。已锁定的协变量的**已进入**框显示变暗的选中标记，以指示它在模型中。
- 若您为未进入的协变量选择了**锁定**，则它不进入模型而不管使用控件按钮所做的选择。

已进入 指示哪些列当前在模型中。您可以手动选择进入或退出模型的列。变暗的选中标记指示已进入模型的锁定的协变量。

列 关注的协变量。

F 比 使用协方差分析模型获得的组变量检验的 *F* 比。请参见“更新“F 比”和“概率 >F””。

概率 > F 使用协方差分析模型获得的组变量检验的 *p* 值。请参见“更新“F 比”和“概率 >F””。

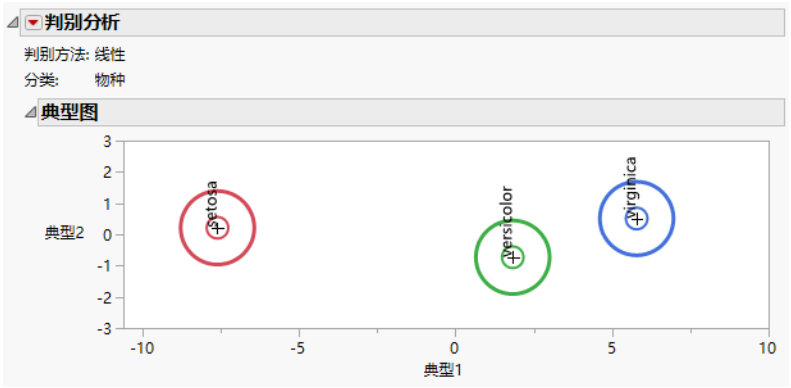
判别方法

在“判别”平台中，提供了若干方法来执行判别分析：“线性”、“二次”、“正则”和“宽线性”。前三个方法的基础模型不同。当协变量数很大时，“宽线性”方法是拟合线性模型的有效方式。

注意：您使 500 个以上的协变量进入时，“JMP 警示”建议您切换到“宽线性”方法。这是因为在列数过多时使用其他方法，计算时间会相当长。点击**宽线性，许多列**可切换到“宽线性”方法。点击**继续**可使用您最初选定的方法。

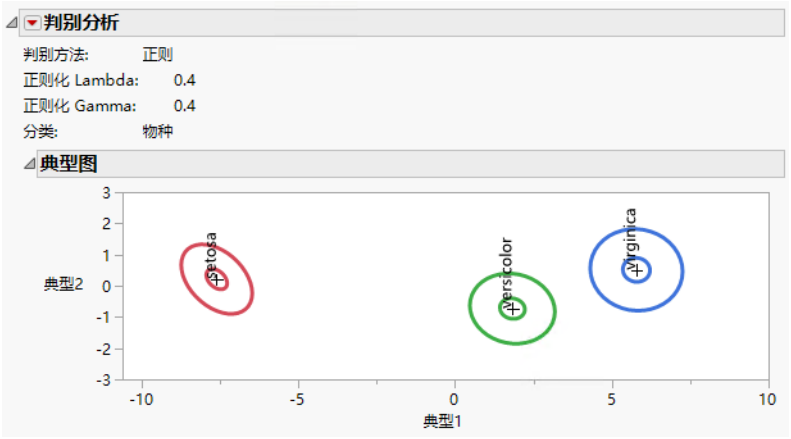
图 5.5 “线性”、“二次” 和 “正则” 判别分析

线性



二次

正则
(λ=0.4, γ=0.4)



在图 5.5 中举例说明了 “线性”、“二次” 和 “正则” 方法。在此处简要介绍一下这些方法。请参见 “保存的公式的统计详细信息”。

线性，共同协方差 执行线性判别分析。该方法假定组内协方差矩阵是相等的。请参见 “线性判别方法”。

二次，不同协方差 执行二次判别分析。该方法假定组内协方差矩阵不同。该方法需要估计比线性方法更多的参数。若组样本大小很小，则可能得到不稳定的估计值。请参见“二次判别方法”。

若协变量在 X 变量的各水平之间保持不变，则它在组内协方差矩阵中的相关元素的协方差为零。为了使矩阵可以求逆，零协方差被替换为相应的合并组内协方差。完成后，一条注释会显示在报表窗口中，标识有问题的协变量和 X 的水平。

提示：二次方法的缺点在小数据集中显露出来。它很难构造可逆且稳定的协方差矩阵。正则法在允许组间差异的基础上改善了以上问题。

正则，折衷方法 当组内协方差矩阵不同时提供使估计值稳定的两种方法。当组样本大小很小时该选项很有用。请参见“正则，折衷方法”和“正则判别方法”。

宽线性，许多列 基于很多协变量拟合模型时很有用，此时使用其他方法计算会很困难。该方法假定所有组内协方差矩阵是相等的。该方法使用奇异值分解方法来计算合并的组内协方差矩阵的逆矩阵。请参见“宽线性算法的统计详细信息”。

注意：使用“宽线性”选项时，通常为其他判别方法显示的几个功能不可用。这是因为该算法不显式计算很大的合并组内协方差矩阵。

正则，折衷方法

正则判别分析由两个非负参数确定。

- 第一个参数（Lambda，收缩到共同协方差）指定如何混合单个协方差矩阵和组协方差矩阵。对于该参数，1 对应于线性判别分析，0 对应于二次判别分析。
- 第二个参数（Gamma，收缩到对角线）是一个乘数，指定对非对角元素（各变量上的协方差）应用多少缩小量。若您选择 1，则强制协方差矩阵为对角矩阵。

为这两个参数都赋值 0 与请求执行二次判别分析的效用相同。类似地，为 Lambda 赋值 1 并为 Gamma 赋值 0 表示请求线性判别分析。使用表 5.1 帮助您决定正则化。有关线性、二次和正则判别分析的示例，请参见图 5.5。

表 5.1 正则判别分析

使用较小的 Lambda	使用较大的 Lambda	使用较小的 Gamma	使用较大的 Gamma
协方差矩阵不同	协方差矩阵相同	变量相关	变量不相关
很多行	很少行		
很少变量	很多变量		

收缩协方差

在“判别”启动窗口中，您可以选择“收缩协方差”选项。当一些组具有少量的观测时建议使用该选项。判别分析需要对协方差矩阵求逆。收缩非对角元素可以提高稳定性并减小预测方差。“收缩协方差”选项通过一个因子对非对角线元素进行收缩，该因子使用 Schafer and Strimmer (2005) 中所述的方法来确定。

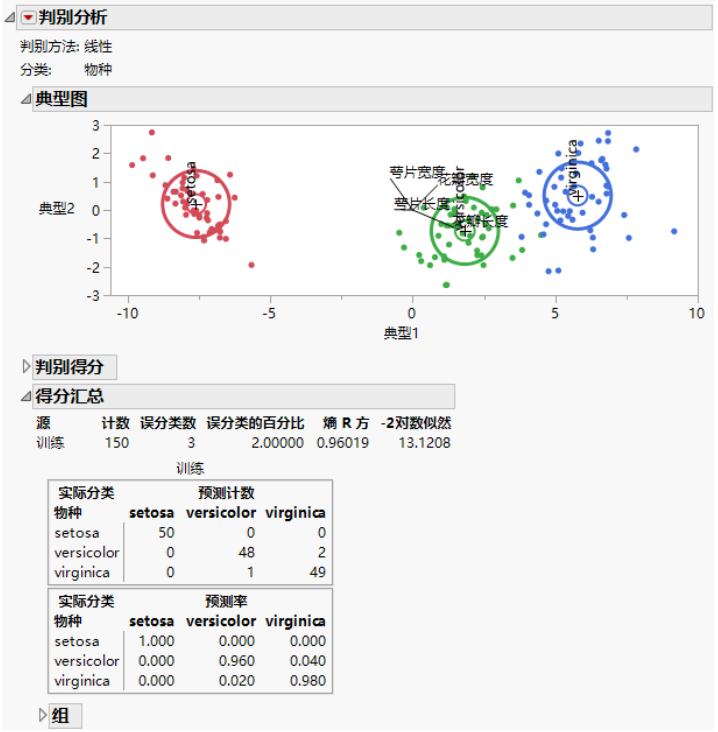
若您在启动窗口中对“线性”判别方法使用“收缩协方差”选项，这样会提供协方差矩阵收缩，它等效于具有适当 Lambda 和 Gamma 值的“正则”判别方法提供的收缩。当您选择“收缩协方差”选项并运行分析时，“收缩”报表将给出“总体收缩”值和“总 Lambda”值。要使用“正则”方法获得相同的分析结果，请在“正则化参数”窗口中，输入 1 作为 Lambda 并使用“收缩”报表中的“总 Lambda”作为 Gamma。

“判别分析”报表

“判别分析”报表基于您选择的判别方法提供判别结果。“判别方法”和“分类”变量显示在报表顶部。若您选择了“正则”方法，还显示相关参数。

您可以通过从“判别分析”红色小三角菜单中选择选项来更改“判别方法”。将更新报表中的结果以反映选择的方法。

图 5.6 “判别分析”报表的示例



默认 “判别分析”报表包含以下部分：

- 当您选择“宽线性”判别方法时，将显示“主成分”报表。请参见“主成分”。
- “典型图”显示分组效果最佳的两个维中的点和多元均值。请参见“典型图和典型结构”。
- “判别得分”报表提供有关如何对每个观测进行分类的详细信息。请参见“判别得分”。
- “得分汇总”报表提供观测分类效果的概况。请参见“判别得分”。

主成分

仅当在启动窗口中将 “宽线性”选作 “判别方法”时，才会显示该部分。考虑以下符号：

- 用 **Y** 表示 $n \times p$ 的协变量矩阵，其中 n 是观测数，p 是协变量数。
- 对于 **Y** 中的每个观测，减去协变量均值，并将差值除以协变量的合并标准差。用 $\mathbf{Y}_S$ 表示所得的矩阵。

该报表给出以下值：

数目 提取的特征值数。提取特征值，直到“累积百分比”至少为 99.99%，这表示解释了 99.99% 的变异。

特征值 按降序排列的 $\mathbf{Y}_S$ 的协方差矩阵的特征值，即 $(\mathbf{Y}_S'\mathbf{Y}_S)/(n - p)$。

累积百分比 特征值累积和占所有特征值之和的百分比。这些特征值的和等于 $\mathbf{Y}_S'\mathbf{Y}_S$ 的秩。

奇异值 按降序排列的 $\mathbf{Y}_S$ 的奇异值。

典型图和典型结构

在“判别分析”报表中，“典型图”是一个双标图，它描述变量的典型相关性结构。

典型结构

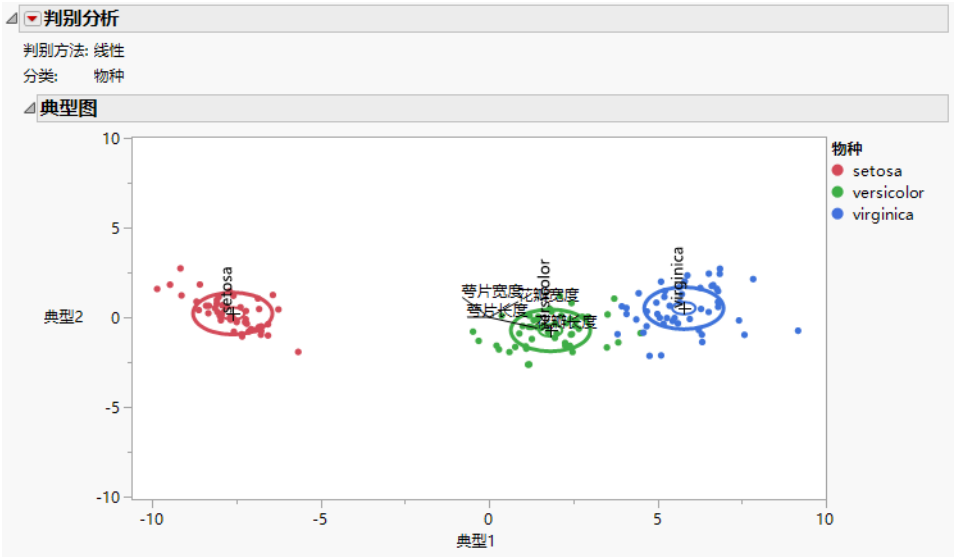
“X，类别”列的每个水平定义一个指示符变量。典型相关性在表示类别的指示符变量组和协变量之间执行。会推导得出协变量的线性组合（称为**典型变量**）。这些典型变量尝试汇总类别间的变异。

第一个典型变量是协变量的线性组合，它最大化类别指示符变量和协变量之间的多重相关性。第二个典型变量是与第一个典型变量不相关的协变量的线性组合，它最大化协变量与类别的多重相关性。若“X，类别”列具有 k 个水平，则获取 $k - 1$ 个典型变量。

典型图

图 5.7 显示了数据表 Iris.jmp 的线性判别分析的典型图。这些点已按**物种**着色。

图 5.7 Iris.jmp 的典型图



双标图轴是前两个典型变量。这些变量定义两个维度来提供组之间的最大分隔。每个典型变量是协变量的线性组合。（请参见 “典型结构”。）双标图显示如何用典型变量表示每个观测以及每个协变量对典型变量的贡献大小。

- 观测值和每个组的多元均值表示为双标图上的点。它们用前两个典型变量表示。
  - 对应于每个多元均值的点用加号 (“+”) 标记表示。
  - 为每个均值标绘 95% 置信水平椭圆。若两个组显著不同，则置信椭圆倾向于不相交。
  - 为每个组绘制表示 50% 等高线的椭圆。这在前两个典型变量的空间中绘制一个区域来包含大约 50% 的观测（假定正态性）。
- 图中显示的一组射线表示协变量。
  - 对于每个典型变量，线性组合中协变量的系数可以解释为**权重**。
  - 为了帮助在权重之间进行比较，协变量进行了标准化，使得每个协变量的均值为 0 且标准差为 1。标准化协变量的系数称为**典型权重**。协变量的典型权重越大，它与典型变量的关联越强。
  - 双标图中每条射线的长度和方向指示相应的协变量与前两个典型变量的关联度。射线长度是典型权重的倍数。
  - 这些射线从点 (0,0) 发出，该点表示用典型变量表示的数据的总均值。
  - 通过从“判别分析”红色小三角菜单中选择**典型选项 > 显示典型详细信息**可获取权重系数值。在“典型详细信息”报表底部，点击“标准化得分系数”。请参见“标准化得分系数”。

修改典型图

有更多选项供您修改双标图：

- 通过从“判别分析”红色小三角菜单中选择**典型选项 > 显示均值置信限椭圆**来显示或隐藏 95% 置信椭圆。
- 通过从“判别分析”红色小三角菜单中选择**典型选项 > 显示双标图射线**来显示或隐藏射线。
- 将双标图射线中心拖到图中其他位置。通过从“判别分析”红色小三角菜单中选择**典型选项 > 双标图射线位置**来指定其位置和尺度。除非需要调整以使射线可见，“典型图”中显示的默认“射线尺度”为 1.5。
- 通过从“判别分析”红色小三角菜单中选择**典型选项 > 显示正态 50% 等高线**来显示或隐藏 50% 等高线。
- 通过从“判别分析”红色小三角菜单中选择**典型选项 > 点着色**来对点进行颜色编码以匹配椭圆。

分为三个或更多类别

对于 Iris.jmp 数据，有三个**物种**，因此只有两个典型变量。图 5.7 中的图显示了使用这两个典型变量很好地分隔开三个组。

图中的射线指示以下信息：

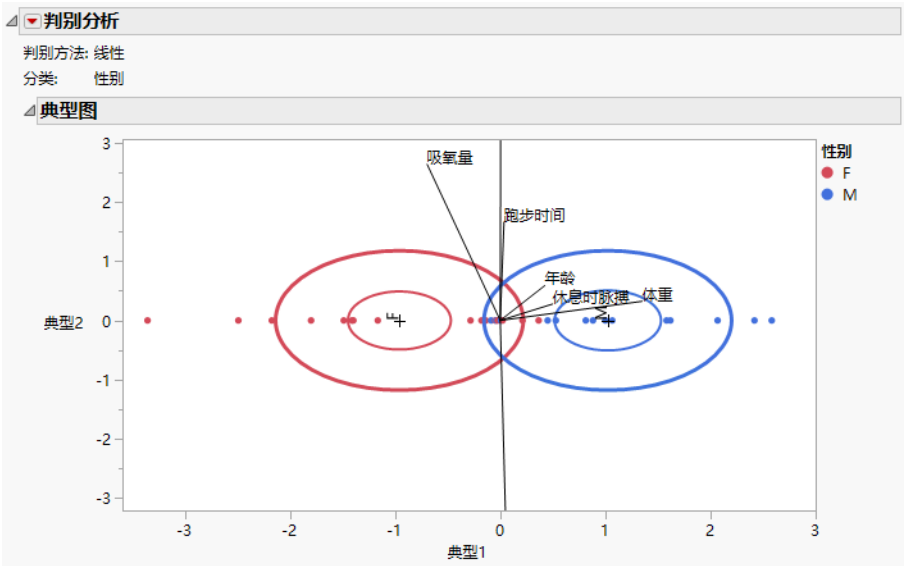
- **花瓣长度**与“典型 1”正相关，与“典型 2”负相关。与“典型 2”相比，定义“典型 1”时它具有更大的权重。
- **花瓣宽度**与“典型 1”和“典型 2”均正相关。在定义两个典型变量时，它具有大概相同的权重。
- **萼片宽度**与“典型 1”负相关，与“典型 2”正相关。与“典型 1”相比，定义“典型 2”时它具有更大的权重。
- **萼片长度**在定义“典型1”时负加权，并且与定义“典型2”具有很弱的关联。

分为两个类别

分类变量只有两个水平时，针对单个典型变量（在图中用“典型1”表示）绘制点。每个协变量的典型权重仅与“典型1”有关。射线只显示垂直成分来分隔它们。将这些射线投影到“典型1”轴以比较它们与单个典型变量的相对关联度。

图 5.8 显示了样本数据表 Fitness.jmp 的“典型图”。使用七个连续变量将个体分类为 M（男性）或 F（女性）。因为分类变量只有两个类别，因此只有一个典型变量。

图 5.8 Fitness.jmp 的 “典型图”



“典型图”中的点已经按**性别**着色。请注意，这两个组用“典型1”的值很好地分隔开。

尽管对应于七个协变量的射线有垂直成分，在这种情况下您必须仅根据它们在 “典型 1”轴上的投影来解释射线。您注意到：

- **最大脉搏**、**跑步时间**和**跑步时脉搏**与“典型1”的关联度很低。
- **体重**、**休息时脉搏**和**年龄**与“典型 1”正相关。**体重**的关联度最高。协变量**休息时脉搏**和**年龄**具有类似但是更小的关联度。
- **吸氧量**与“典型1”负相关。

判别得分

在 “判别分析”报表中，“判别得分”部分提供每个观测的预测分类和支持信息。

行 数据表中的观测行。

实际分类 数据表中给出的观测分类。

SqDist(实际分类) 数据表中给出的观测分类的已保存公式 SqDist[< 水平 >] 的值。请参见“得分选项”。

> **注意：**由于公式中有偏移项， SqDist(实际分类) 可能为负。

Prob(实际分类) 观测的实际分类的估计概率。

-Log(Prob) Prob(实际分类) 的对数的负数。这个负对数似然值很大则表示使用实际类别的成员关系预测观测的效果不好。

-Log(Prob) 的图显示在 -Log(Prob) 值右侧。大的直条指示预测效果不好。星号 (*) 指示误分类的观测。

若您使用验证集或测试集，则验证集中的观测使用 “v” 标记，测试集中的观测使用 “t” 标记。

预测分类 观测的预测分类。预测分类是具有成员关系的最高预测概率的类别。

Prob(预测分类) 观测的预测分类的估计概率。

其他 列出其他预测概率超过 0.1 的类别 （若存在）。

图 5.9 显示了使用 “线性” 判别方法得到的 Iris.jmp 样本数据表的 “判别得分” 报表。选择了**得分选项** > **仅显示需要的行**选项，因而只显示误分类的行或预测概率介于 0.05 到 0.95 之间的行。

图 5.9 仅显示需要的行

判别得分

行	实际分类	SqDist(实际分类)	Prob(实际分类)	-Log(Prob)		预测值	Prob(预测分类)	其他
71	versicolor	8.66970	0.2532	1.373	*	virginica	0.7468	
73	versicolor	4.87619	0.8155	0.204		versicolor	0.8155	virginica 0.18
78	versicolor	4.66698	0.6892	0.372		versicolor	0.6892	virginica 0.31
84	versicolor	8.43926	0.1434	1.942	*	virginica	0.8566	
120	virginica	8.19641	0.7792	0.249		virginica	0.7792	versicolor 0.22
124	virginica	3.57858	0.9029	0.102		virginica	0.9029	
127	virginica	3.90184	0.8116	0.209		virginica	0.8116	versicolor 0.19
128	virginica	3.31470	0.8658	0.144		virginica	0.8658	versicolor 0.13
130	virginica	9.08495	0.8963	0.109		virginica	0.8963	versicolor 0.10
134	virginica	7.23593	0.2706	1.307	*	versicolor	0.7294	
135	virginica	15.83301	0.9340	0.068		virginica	0.9340	
139	virginica	4.09385	0.8075	0.214		virginica	0.8075	versicolor 0.19

“*” 表示误分类

得分汇总

在 “判别分析” 报表中，“得分汇总” 部分提供判别得分的概述。图 5.10 中的表显示实际分类和预测分类。若所有观测均正确分类，则非对角计数为零。

图 5.10 Iris.jmp 的 “得分汇总”

得分汇总

源	计数	误分类数	误分类的百分比	熵 R 方	-2对数似然
训练	65	1	1.53846	0.94487	7.81888
验证	49	3	6.12245	0.86644	
测试	36	0	0.00000	0.98410	

训练

实际分类	预测计数		
物种	setosa	versicolor	virginica
setosa	18	0	0
versicolor	0	22	1
virginica	0	0	24

实际分类	预测率		
物种	setosa	versicolor	virginica
setosa	1.000	0.000	0.000
versicolor	0.000	0.957	0.043
virginica	0.000	0.000	1.000

验证

实际分类	预测计数		
物种	setosa	versicolor	virginica
setosa	16	0	0
versicolor	0	15	2
virginica	0	1	15

实际分类	预测率		
物种	setosa	versicolor	virginica
setosa	1.000	0.000	0.000
versicolor	0.000	0.882	0.118
virginica	0.000	0.063	0.938

测试

实际分类	预测计数		
物种	setosa	versicolor	virginica
setosa	16	0	0
versicolor	0	10	0
virginica	0	0	10

实际分类	预测率		
物种	setosa	versicolor	virginica
setosa	1.000	0.000	0.000
versicolor	0.000	1.000	0.000
virginica	0.000	0.000	1.000

“得分汇总”报表提供以下信息：

列 若您使用了 “逐步选择变量”构造模型，则列出进入模型的列。

源 若未使用验证，则所有观测构成训练集。若使用了验证，则分别显示一行表示训练集和验证集，或分别显示一行表示训练集、验证集和测试集。

误分类数 提供指定的集中被错误分类的观测数。

误分类的百分比 提供指定的集中被错误分类的观测百分比。

熵 R 方 一个拟合测度。值越大表示拟合效果越好。“熵 R 方”的值为 1 表示分类预测完美。由于判别模型的预测概率的不确定性很典型，因此 “熵 R 方”值往往很小。

请参见 “熵 R 方”。

注意：“熵 R 方”可能为负值。

-2 对数似然 训练集中观测的负对数似然的两倍 （基于模型）。值越小表示拟合效果越好。仅对训练集提供。请参见 《拟合线性模型》。

混杂矩阵 显示分类变量 X 的每个水平的 “预测计数 - 实际分类”矩阵。若使用的是带验证功能的 JMP Pro，将为每组观测提供一个矩阵。若您在 JMP 中使用了排除行，则将排除行视为验证集且给出单独的验证矩阵。请参见 “JMP 和 JMP Pro 中的验证”。

熵 R 方

“熵 R 方”是一个拟合测度。为训练集计算该值，若使用了验证，则为验证集和测试集计算该值。

训练集的“熵 R 方”

对于训练集，按以下方式计算 “熵 R 方”：

- 使用训练集拟合判别模型。
- 获取基于模型的预测概率。
- 使用这些预测概率，为训练集中的观测计算似然值。称之为似然_完全$_{训练}$。
- 使用训练集拟合简化模型（无预测变量）。
- 使用简化模型中 X 变量的各水平的预测概率计算训练集中观测的似然值。将该量称为似然_简化$_{训练}$。
- 训练集的“熵 R 方”为：

$$熵\ R\ 方_{训练} = 1 - \frac{\log(似然\_完全_{训练})}{\log(似然\_简化_{训练})}$$

验证集和测试集的“熵 R 方”

对于验证集，按以下方式计算 “熵 R 方”：

- 仅使用训练集拟合判别模型。
- 为所有观测获取基于训练集模型的预测概率。
- 使用这些预测概率，为验证集中的观测计算似然值。称之为似然_完全$_{验证}$。
- 仅使用训练集拟合简化模型（无预测变量）。
- 使用简化模型中 X 变量的各水平的预测概率计算验证集中观测的似然值。将该量称为似然_简化$_{验证}$。
- “验证熵 R 方”为：

$$验证熵\ R\ 方 = 1 - \frac{\log(似然\_完全_{训练})}{\log(似然\_简化_{训练})}$$

测试集的“熵 R 方”计算方式与验证集的“熵 R 方”计算方式相似。

“判别分析”选项

“判别分析”红色小三角菜单包含以下选项：

逐步选择变量 （对于 “宽线性”方法不可用。）显示或隐藏 “列选择”控制面板。该控制面板包含的选项支持您使用协方差分析和 *p* 值逐步选择变量。请参见 “逐步选择变量”。

判别方法 指定判别方法。选择“线性”、“二次”、“正则”和“宽线性”。请参见“判别方法”。

判别得分 显示或隐藏每行的判别得分表。

得分选项 提供用于对观测评分的选项。

仅显示需要的行 在“判别得分”报表中，只显示误分类的行和预测概率介于 0.05 到 0.95 之间的行。

显示分类计数 显示或隐藏“得分汇总”报表中的混淆矩阵和混淆率矩阵。混淆矩阵是实际响应和预测响应的双向分类。混淆率矩阵等同于混淆矩阵，只不过其中的数字要除以行合计。默认情况下，“得分汇总”报表显示分类变量 X 的每个水平的混淆矩阵和混淆率矩阵。若使用的是带验证功能的 JMP Pro，将为每组观测提供一个矩阵。若您在 JMP 中使用了排除行，则将排除行视为验证集且给出单独的验证矩阵。请参见“JMP 和 JMP Pro 中的验证”。

显示各组距离 显示或隐藏一个报表，其中包含每个观测到每个组均值的 Mahalanobis 平方距离。

显示各组概率 显示或隐藏一个报表，其中包含某个观测属于分类变量 X 所定义的每个组的概率。

ROC 曲线 在“得分汇总”报表中显示或隐藏“受试者操作特征 (ROC)”曲线。有关 ROC 曲线的详细信息，请参见《预测和专业建模》。

选择误分类的行 在数据表以及按行显示列表的报表窗口中选择误分类的行。

选择不确定的行 在数据表以及按行显示列表的报表窗口中选择具有不确定分类的行。不确定的行是指该行属于**任意**组的成员的概率既不接近于 0 也不接近于 1。

选择该选项时，将打开一个窗口，您可以在此指定反映不确定性的预测概率的范围。默认情况下，概率与 0 或 1 存在超过 0.1 的差值的任何行定义为不确定的行。因此，默认选择其概率介于 0.1 到 0.9 之间的行。

保存公式 将距离、概率和预测的成员关系公式保存到数据表。请参见“保存的公式的统计详细信息”。

- 距离公式为 SqDist0 和 SqDist[<水平>]，其中 <水平> 表示 X 的水平。距离公式生成与 Mahalanobis 距离计算有关的中间值。
- 概率公式为 Prob[<水平>]，其中 <水平> 表示 X 的水平。每个概率列给出在 X 的该水平上观测所属成员关系的后验概率。将“响应概率”列属性保存到每个概率列。有关“响应概率”列属性的详细信息，请参见《使用 JMP》。
- 预测的成员关系公式为 Pred <X>，它包含“最可能的水平”分类规则。
- “宽线性”方法还保存**判别数据矩阵**列，该列包含协变量的向量和**判别主成分**公式。请参见“宽线性判别方法”。

注意：对于“宽线性”之外的其他方法，当您保存公式时，“RowEdit Prob”脚本会保存到数据表。该脚本选择数据表中不确定的行。该脚本将其概率与 0 或 1 存在超过 0.1 的差值的行定义为不确定的行。它还打开一个“行编辑器”窗口，您可以在其中查看不确定的行。若您拟合新模型（“宽线性”除外）并选择“保存公式”，则现有“RowEdit Prob”脚本会被替换为适用于新拟合的脚本。

生成评分脚本 （仅在 JMP 标准版中可用。）创建一个脚本，它构造使用“保存公式”选项保存的公式列。您可以保存并使用该脚本（可能涉及其他数据表），以创建计算成员关系概率以及预测组成员关系的公式列。

JMP PRO 发布概率公式 （仅限于 JMP Pro。）创建概率公式并在“公式存储库”平台中将它们保存为公式列脚本。若未打开“公式存储库”报表，该选项将创建“公式存储库”报表。请参见《预测和专业建模》。

典型图 显示或隐藏典型图。请参见 “典型图和典型结构”。

典型选项 提供影响 “典型图”和 “典型三维图”的选项。

显示点 在“典型图”和“典型三维图”中显示或隐藏点。

显示均值置信限椭圆 显示或隐藏“典型图”和“典型三维图”中的每个组的均值的 95% 置信椭圆（假定正态性）。

显示正态 50% 等高线 显示或隐藏每个组的 50% 预测椭圆或椭球体。在“典型图”中，每个椭圆在前两个典型变量的空间中绘制一个估计每组 50% 的观测会落入的区域（假定正态性）。在“典型三维图”中，每个椭球在前三个典型变量的空间中绘制一个估计 50% 的观测会落在的区域（假定多元正态性）。

显示双标图射线 在“典型图”和“典型三维图”中显示或隐藏双标图射线。带标签的射线显示协变量在典型空间中的方向。该方向表示每个协变量与每个典型变量的关联度。

双标图射线位置 允许您在“典型图”和“典型三维图”中指定双标图射线的位置和射线尺度。

– 默认情况下，这些射线从点 (0,0) 发出，该点表示用典型变量表示的数据的总均值。在“典型图”中，您可以拖动射线或使用该选项指定坐标。

– “典型图”中的默认“射线尺度”为 1.5，除非需要调整以使射线可见。“射线尺度”的指定与“标准化得分系数”有关。

点着色 基于 *X* 变量的水平为“典型图”和“典型三维图”中的点着色。将颜色标记添加到数据表中的行。该选项等效于选择**行 > 按列设定颜色或标记**并选择 *X* 变量。它也等效于右击图形并选择**行图例**，然后按分类列着色。

显示典型详细信息 显示或隐藏“典型详细信息”报表。请参见“显示典型详细信息”。

显示典型结构 显示或隐藏“典型结构”报表。请参见“显示典型结构”。对于“宽线性”判别方法不可用。

保存典型得分 在数据表中创建包含每个观测的典型得分公式的列。第 *k* 个典型得分的列命名为 Canon[<k>]。

提示：在脚本中，将脚本命令 `Save to New Data Table` 发送到 “判别”对象会将以下内容保存到新数据表：典型变量的组均值、标准化得分系数的射线尺度为 1.5 的双标图射线以及典型得分。对于 “宽线性”判别方法不可用。

典型三维图 显示三维典型图。仅当分类变量 X 有四个或更多水平时该选项才可用。请参见 ““典型三维图”的示例”。

指定先验值 允许您指定 *X* 变量每个水平的先验概率。以下选项可用于指定先验值：

相等概率 将相等先验概率分配给所有组。这是默认值。

发生比例 将先验概率分配给与观测数据中的频数成比例的组。

其他 允许您指定定制先验概率。

考虑新水平 指定某些点可能不适合任何已知组，并且应考虑来自未计算分数的新组。请参见“考虑新水平”。

显示组内协方差 显示或隐藏以下报表：

- 给出合并的组内协方差和相关性矩阵的“协方差矩阵”报表。
- 对于“二次”和“正则”方法，显示组内相关性矩阵的“每组的相关性”报表。

  对于每个组，还显示组内协方差矩阵的行列式的对数。
- 对于“二次”判别方法，将“组协方差”分级显示项添加到显示组内协方差矩阵的“协方差矩阵”报表。

“显示组内协方差”对于 “宽线性”判别方法不可用。

显示组均值 显示或隐藏提供每个协变量均值的 “组均值”报表。显示 *X* 变量的每个水平的均值和总均值。

保存判别矩阵 将名为**判别结果**的脚本保存到数据表。该脚本是在 JSL 中使用的以下对象的列表：

- 协变量 (Y) 的列表
- 分类变量 X
- X 的水平的列表
- X 不同水平下的协变量均值矩阵
- 合并的组内协方差矩阵

“保存判别矩阵”对于 “宽线性”判别方法不可用。请参见“保存判别矩阵”。

散点图矩阵 （对于“宽线性”判别方法不可用。）在单独窗口中打开“散点图矩阵”平台，该窗口包含协变量的下三角散点图矩阵。散点图包含数据表中的所有观测，即使使用了验证。对于分类变量 X 的每个水平显示具有 90% 覆盖率的椭圆。对于 “线性”判别方法，这些椭圆基于合并的组内协方差矩阵。

《基本绘图》对 “散点图矩阵”红色小三角菜单中的选项进行了说明。

刻画器 显示或隐藏交互式刻画器报表，其中的分类概率合并为单个刻画器行。因子值的更改体现在估计的分类概率中。有关红色小三角菜单中选项的详细信息，请参见《刻画器指南》。

请参见《使用 JMP》获取有关下列选项的信息：

本地数据过滤器 显示或隐藏支持您过滤特定报表中使用的数据的本地数据过滤器。

重新运行 包含使您可以重复或重新启动分析的选项。在支持该功能的平台中，“自动重新计算”选项立即在相应报表窗口中反映您对数据表所做的更改。

平台首选项 包含的选项支持您查看当前平台首选项或更新平台首选项以匹配当前 JMP 报表中的设置。

保存脚本 包含的选项支持您保存可将报表重现到若干目标的脚本。

保存 “依据”组脚本 包含使您可以保存脚本的选项，可将为“依据”变量的所有水平重新生成平台报表的脚本保存到多个不同的位置。仅当在启动窗口中指定“依据”变量时才可用。

显示典型详细信息

在“判别”平台中，“典型详细信息”报表显示说明协变量和分组变量 *X* 之间的关系的检验。相关矩阵显示在报表底部。

图 5.11 Iris.jmp 的 “典型详细信息”

典型详细信息

从总合并组内协方差矩阵计算得到的典型详细信息。

特征值	百分比	累积百分比	典型相关性	似然比	近似的 F 值	分子自由度	分母自由度	概率>F
32.1919292	99.1213	99.1213	0.98482089	0.02343863	199.1453	8	288	<.0001*
0.28539104	0.8787	100.0000	0.47119702	0.77797337	13.7939	3	145	<.0001*

检验	值	近似的 F 值	分子自由度	分母自由度	概率>F
Wilks Lambda	0.0234386	199.1453	8	288	<.0001*
Pillai 迹	1.1918988	53.4665	8	290	<.0001*
Hotelling-Lawley	32.47732	582.1970	8	203.4	<.0001*
Roy 最大根	32.191929	1166.9574	4	145	<.0001*

▷组内矩阵
▷组间矩阵
▷得分系数
▷标准化得分系数

注意：计算报表中的结果所用的矩阵是合并的组内协方差矩阵（称之为组内矩阵）。该矩阵是所有判别方法的“典型详细信息”报表的基础。“典型详细信息”报表中的统计量和检验对于所有判别方法都是相同的。

统计量和检验

“典型详细信息”报表列出特征值并给出零特征值的似然比检验。对于典型相关性为零这个原假设提供四个检验。

特征值 组间矩阵和组内矩阵的逆矩阵之积的特征值。它们按从大到小的顺序列出。特征值的大小反映相关的判别函数所解释的变异量。

百分比 给定的特征值在特征值总和中所占的比例。

累积百分比 累积的比例和。

典型相关性 协变量和分类变量 X 所定义的组之间的典型相关性。假定您定义数值指标变量来表示 X 所定义的组。然后使用协变量作为一组变量并使用表示 X 中的各组的指标变量作为另一组变量来执行典型相关性分析。“典型相关性”值是从该分析得到的典型相关性值。

似然比 一个检验的似然比统计量，该检验确定相应典型相关性和所有更小的相关性的总体值是否为零。对于给定典型相关性和所有更小的典型相关性，该比值等于 (1 - 典型相关性 ^{2}) 值的乘积 。

检验 列出针对在各组上协变量的均值相等这个原假设的四个标准检验：Wilk Lambda、 Pillai 迹、 Hotelling-Lawley 和 Roy 最大根。请参见“多元检验的统计详细信息”和“近似 F 检验的统计详细信息”。

近似的 F 值 与相应检验相关的 F 值。对于某些检验，F 值是近似值或是一个上限。请参见“近似 F 检验的统计详细信息”。

分子自由度 相应检验的分子自由度。

分母自由度 相应检验的分母自由度。

概率 >F 相应检验的 p 值。

矩阵

与典型结构有关的四个矩阵显示在报表的底部。要查看矩阵，点击其名称旁边的展开图标。要隐藏矩阵，点击该矩阵的名称。

组内矩阵 合并的组内协方差矩阵。

组间矩阵 组间协方差矩阵 $\mathbf{S}_B$。请参见 “组间协方差矩阵的统计详细信息”。

得分系数 用于根据原始数据计算典型得分的系数。这些是用于选项**典型选项 > 保存典型得分**的系数。有关如何计算这些系数的详细信息，请参见 SAS Institute Inc.(2020b) 中的“CANDISC 过程”一章。

标准化得分系数 用于根据标准化数据计算典型得分的系数。经常称为**典型权重**。有关如何计算这些系数的详细信息，请参见 SAS Institute Inc.(2020b) 中的 “CANDISC 过程”一章。

显示典型结构

在“判别”平台中，“典型结构”报表给出三个矩阵，它们提供典型变量和协变量之间的相关性。另一个矩阵显示组变量的各个水平上的均值。要查看矩阵，点击其名称旁边的展开图标。要隐藏矩阵，点击该矩阵的名称。

图 5.12 Iris.jmp 的 “典型结构”（显示组间典型结构）


典型结构

▷合计典型结构

组间典型结构

	萼片长度	萼片宽度	花瓣长度	花瓣宽度
典型 “1”	0.9914683	-0.825658	0.99975	0.9940442
典型 “2”	0.1303484	0.5641714	0.0223578	0.1089775

▷合并的组内典型结构

▷典型变量的组均值

合计典型结构 典型变量和协变量之间的相关性。通常称为**载荷**。

组间典型结构 典型变量的组均值和协变量的组均值之间的相关性。

合并的组内典型结构 典型变量和协变量之间的偏相关性，已针对组变量调整。

典型变量的组均值 提供每个典型变量在组变量的各个水平上的均值。

考虑新水平

若您怀疑一些观测对于分类变量的指定水平来说是离群值，可以使用“判别分析”红色小三角菜单中的“考虑新水平”选项。选择该选项时，将显示一个菜单，要求您指定新水平的先验概率。

对于用新组来拟合效果会更好的观测，会分配给称为“其他”的新水平。“其他”组中成员关系的概率假定这些观测服从整个观测集的分布，该分布假定**无组结构**。这导致得到与协方差结构相关的相应宽的正态等高线。距离计算按指定的先验概率进行了调整。

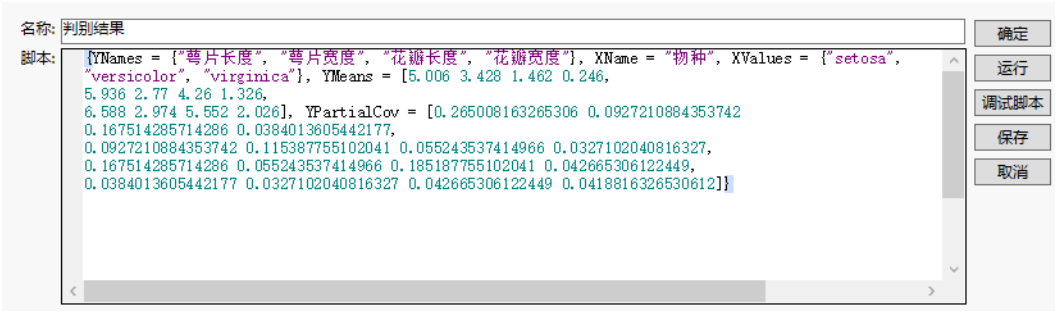
保存判别矩阵

在“判别分析”红色小三角菜单中，“保存判别矩阵”选项创建一个全局列表 (`DiscrimResults`)，可供在 JMP 脚本语言中使用。该列表包含以下为训练集计算的信息：

- Y 名称，协变量 (Y) 的列表
- X 名称，分类变量
- X 值，X 的水平列表
- Y 均值，协变量均值 X 的水平的矩阵
- Y 偏相关性，组内协方差矩阵

考虑使用 Iris.jmp 样本数据表中的判别脚本获得的分析。若您从“判别分析”红色小三角菜单中选择保存判别矩阵，则会将脚本判别结果保存到数据表。

图 5.13 Iris.jmp 的 “判别结果” 表脚本



注意：在脚本中，您可以将脚本命令 “Get Discrim Matrices” 发送到 “判别” 平台对象。这将获得与 “保存判别矩阵” 相同的值，但是不在数据表中储存它们。

JMP 和 JMP Pro 中的验证

在 “判别” 平台中，指定要使用的验证集取决于 JMP 的版本。在标准 JMP 中，您可以通过排除构成验证集的行来指定验证集。选择要用作验证集的行，然后选择**行 > 排除/撤销排除**。没有被排除的行会被视为训练集。

在 JMP Pro 中，您可以在 “判别” 启动窗口中指定验证列。验证列必须具有数值数据类型，应包含至少两个非重复值。

请注意以下事项：

- 若列包含两个值，则较小的值定义训练集，较大的值定义验证集。
- 若列包含三个值，则这些值按大小递增的顺序定义训练集、验证集和测试集。
- 若列包含四个或更多非重复值，则只使用最小的三个值和相关观测按该顺序定义训练集、验证集和测试集。

指定验证集时，“判别” 平台执行以下步骤：

- 使用训练数据拟合模型。
- “逐步选择变量” 选项给出模型的 “验证熵 R 方” 和 “验证误分类率” 统计量。请参见 “统计量” 和 “验证集和测试集的 “熵 R 方” ”。
- “判别得分” 报表显示一个指示符，它用于标识验证集和测试集中的行。
- “得分汇总” 报表显示训练集、验证集和测试集的 “预测分类-实际分类”。

判别分析的更多示例

本节包含使用 “判别”平台的示例。

“典型三维图”的示例

使用 “判别”平台创建某变量的若干水平的 “典型三维图”。

隐藏和排除包含缺失值的行

1. 选择**帮助** > **样本数据文件夹**，然后打开 Owl Diet.jmp。
2. 选择**行** > **行选择** > **选择符合条件的行**。
3. 选择**物种**并点击**添加条件**。
   这将选择缺失**物种**的所有行。您需要隐藏和排除这些行。
4. 选择**行** > **隐藏和排除**。

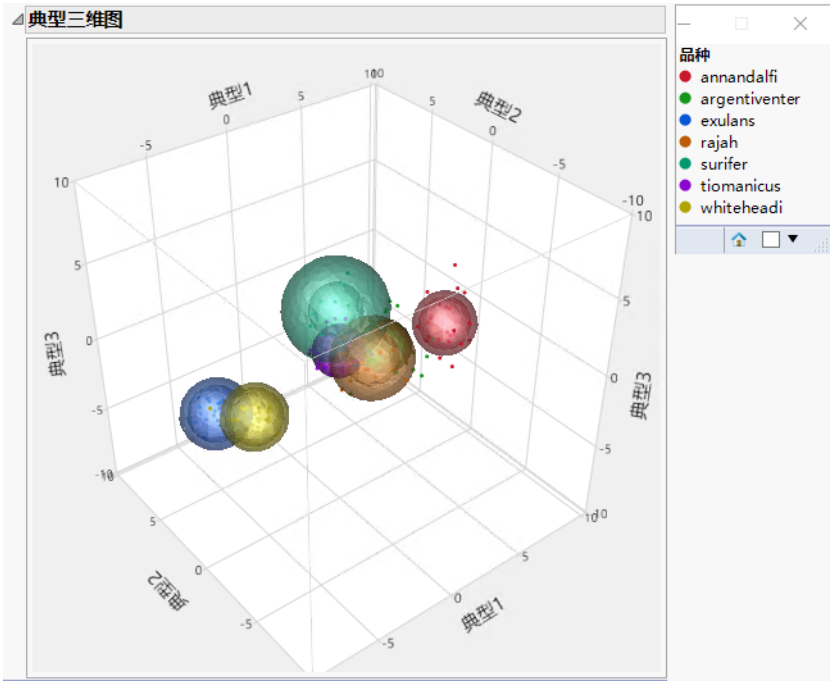
按物种对行着色

5. 选择**行** > **按列设定颜色或标记**。
6. 选择**品种**。
7. 从 “颜色”菜单中，选择 **JMP 深色**。
8. 选中**生成具有图例的窗口**。
9. 点击**确定**。
   随即显示一个小 “图例”窗口。数据表中的行按**品种**分配了颜色。

执行判别分析

10. 选择**分析** > **多元方法** > **判别**。
11. 选择**头骨长度**、**牙齿排数**、**上颚孔**和**下颚长度**，然后点击 **Y，协变量**。
12. 选择**物种**并点击 **X，类别**。
13. 点击**确定**。
14. 点击 “判别分析”红色小三角并点击**典型三维图**。

图 5.14 具有 “图例” 窗口的 “典型三维图”



您可以点击 “图例” 中的类别以便在 “典型三维图” 中突出显示相应的点。还可以在三维图中点击并拖动以便旋转它。

逐步选择变量的示例

在本例中，您使用 “判别” 平台中的逐步选择变量确定在最终模型中包括哪些变量。

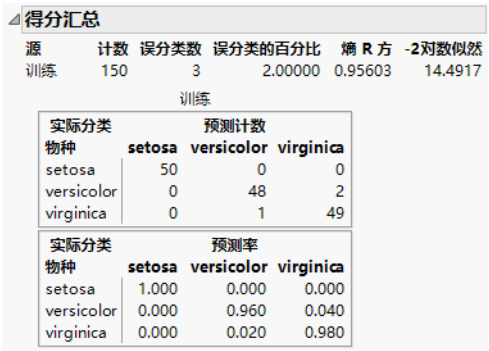
1. 选择**帮助** > **样本数据文件夹**，然后打开 Iris.jmp。
2. 选择**分析** > **多元方法** > **判别**。
3. 选择**萼片长度**、**萼片宽度**、**花瓣长度**和**花瓣宽度**，然后点击 **Y，协变量**。
4. 选择**物种**并点击 **X，类别**。
5. 选择**逐步选择变量**。
6. 点击**确定**。
7. 点击**前进**三次。

图 5.15 Iris.jmp 的逐步模型

三个协变量进入模型。“进入的最小 p 值”显示在顶部面板。它的值为 0.0103288，指示剩余协变量**萼片长度**在**物种**的判别分析模型中也可能是很有用的。

8. 点击**应用该模型**。

图 5.16 显示所选协变量的 “得分汇总”报表



根据进入的协变量和您选择的判别方法，窗口更新以显示拟合的报表。请注意，您为模型选择的协变量在 “得分汇总”报表顶部列出。

“判别”平台的统计详细信息

本节包含在判别分析中使用的统计详细信息。

- “宽线性算法的统计详细信息”
- “保存的公式的统计详细信息”
- “多元检验的统计详细信息”
- “近似 F 检验的统计详细信息”
- “组间协方差矩阵的统计详细信息”

宽线性算法的统计详细信息

按以下方式执行 “宽线性”判别分析：

- 通过减去组均值后除以合并的标准差对数据进行标准化。
- 使用奇异值分解从一组奇异向量获得主成分变换矩阵。
- 保留的成分数表示奇异值平方和至少为 0.9999。
- 对变换后的数据执行线性判别分析，其中数据未进行组均值变换。这是快速计算，因为合并的组内协方差矩阵是对角矩阵。

保存的公式的统计详细信息

本节给出通过“判别分析”红色小三角菜单中的“得分选项”>“保存公式”选项保存的推导公式。这些公式依赖于判别方法。

对于由分类变量 X 定义的每个组，假定协变量的观测服从 p 维多元正态分布，其中 p 是协变量数。表 5.2 中给出了公式中使用的符号。

表 5.2 “保存公式”选项给出的公式符号

p	协变量数
T	组（X 的水平）总数
$t = 1,..., T$	用于区分由 X 定义的各组的下标
n_t	第 t 组中的观测数
$n = n_1 + n_2 + ... + n_T$	总观测数
$\mathbf{y}$	某个观测的协变量的 $1 \times p$ 向量
$\mathbf{y}_{it} = (y_{i1t}, y_{i2t}, \cdots, y_{ipt})$	组 t 中第 i 个观测，它包含 p 个协变量的向量
$\bar{\mathbf{y}}_t$	针对组 t 中的观测，协变量 $\mathbf{y}$ 的均值的 $1 \times p$ 向量
$\mathbf{y}_{bar}$	所有观测的协变量均值的 $1 \times p$ 向量
$\mathbf{S}_t = \frac{1}{n_t - 1} \sum_{i=1}^{n_t} (\mathbf{y}_{it} - \bar{\mathbf{y}}_t)(\mathbf{y}_{it} - \bar{\mathbf{y}}_t)'$	估计的第 t 组的组内协方差矩阵 $(p \times p)$

表 5.2 “保存公式”选项给出的公式符号 (续)

$\mathbf{S}_p = \frac{1}{n-T}\sum_{t=1}^{T}(n_t-1)\mathbf{S}_t$	估计的 (p x p) 合并的组内协方差矩阵
q_t	组 t 的成员关系的先验概率
$p(t\|\mathbf{y})$	$\mathbf{y}$ 属于组 t 的后验概率
$\|\mathbf{A}\|$	矩阵 A 的行列式

线性判别方法

在线性判别分析中，假定所有组内协方差矩阵是相等的。共同协方差矩阵通过 $\mathbf{S}_p$ 来估计。请参见表 5.2 了解相关符号。

观测 $\mathbf{y}$ 到组 t 的 Mahalanobis 距离按以下方式定义：

$$d_t^2 = (\mathbf{y}-\bar{\mathbf{y}}_t)'\mathbf{S}_p^{-1}(\mathbf{y}-\bar{\mathbf{y}}_t)$$

观测 $\mathbf{y}$ 属于第 t 组的似然函数按以下方式估计：

$$\begin{aligned} l_t(\mathbf{y}) &= (2\pi)^{-T/2}\left|\mathbf{S}_p\right|^{-1/2}\exp(-(\mathbf{y}-\bar{\mathbf{y}}_t)'\mathbf{S}_p^{-1}(\mathbf{y}-\bar{\mathbf{y}}_t)/2) \\ &= (2\pi)^{-T/2}\left|\mathbf{S}_p\right|^{-1/2}\exp(-d_t^2/2) \end{aligned}$$

请注意，必须为合并的协方差矩阵估计的参数数目是 $p(p+1)/2$，必须为均值估计的参数数目是 Tp。必须估计的参数总数是 $p(p+1)/2 + Tp$。

按以下方式定义组 t 中的成员关系的后验概率：

$$p(t|\mathbf{y}) = \frac{q_t l_t(\mathbf{y})}{\sum_{u=1}^{T} q_u l_u(\mathbf{y})} = \frac{1}{1+\sum_{u\neq t}\exp(-[(d_u^2-2\log(q_u))-(d_t^2-2\log(q_t))]/2)}$$

观测 $\mathbf{y}$ 被分配给具有最大后验概率的组。

按以下方式定义 “线性”判别方法所保存的公式：

SqDist[0]	$\mathbf{y}'\mathbf{S}_p^{-1}\mathbf{y}$

SqDist[<组 *t*>]	$d_t^2 - 2\log(q_t)$
Prob[<组 *t*>]	$p(t\|\mathbf{y})$
Pred <*X*>	*t*，对于它 $p(t\|\mathbf{y})$ 为最大值， $t = 1, \ldots, T$

二次判别方法

在二次判别分析中，不假定组内协方差矩阵是相等的。组 *t* 的组内协方差矩阵由 $\mathbf{S}_t$ 估计。这意味着必须为组内协方差矩阵估计的参数数目是 *Tp(p+1)/2*，必须为均值估计的参数数目是 *Tp*。必须估计的参数总数是 *Tp(p+3)/2*。

组样本大小相对于 *p* 很小时，组内协方差矩阵的估计值倾向于很不稳定。判别得分受组内协方差矩阵的逆矩阵的最小特征值影响很大。请参见 Friedman (1989)。因此，若您的组样本大小相对于 *p* 来说很小，您可能要考虑“正则判别方法”中所述的“正则”方法。

请参见表 5.2 了解相关符号。观测 $\mathbf{y}$ 到组 *t* 的 Mahalanobis 距离按以下方式定义：

$$d_t^2 = (\mathbf{y} - \bar{\mathbf{y}}_t)' \mathbf{S}_t^{-1} (\mathbf{y} - \bar{\mathbf{y}}_t)$$

观测 $\mathbf{y}$ 属于第 *t* 组的似然函数按以下方式估计：

$$\begin{aligned} l_t(\mathbf{y}) &= (2\pi)^{-T/2} |\mathbf{S}_t|^{-1/2} \exp(-(\mathbf{y} - \bar{\mathbf{y}}_t)' \mathbf{S}_t^{-1} (\mathbf{y} - \bar{\mathbf{y}}_t)/2) \\ &= (2\pi)^{-T/2} |\mathbf{S}_t|^{-1/2} \exp(-d_t^2/2) \end{aligned}$$

组 *t* 的成员关系的后验概率为：

$$\begin{aligned} p(t|\mathbf{y}) &= (q_t l_t(\mathbf{y})) / \left(\sum_{u=1}^{T} q_u l_u(\mathbf{x}) \right) \\ &= \frac{1}{1 + \sum_{u \neq t} \exp(-[(d_u^2 + \log|\mathbf{S}_u| - 2\log(q_u)) - (d_t^2 + \log|\mathbf{S}_t| - 2\log(q_t))]/2)} \end{aligned}$$

观测 $\mathbf{y}$ 被分配给具有最大后验概率的组。

按以下方式定义 “二次”判别方法所保存的公式：

SqDist[<组 *t*>]	$d_t^2 + \log\|\mathbf{S}_t\| - 2\log(q_t)$

Prob[<组 *t*>]	$p(t\|\mathbf{y})$
Pred <*X*>	*t*，对于它 $p(t\|\mathbf{y})$ 为最大值， $t = 1, ..., T$

注意：SqDist[<组 *t*>] 可为负。

正则判别方法

“正则”判别分析允许两个参数：λ 和 γ。

- 参数 λ 权衡分配给合并的协方差矩阵和组内协方差矩阵（不假定相等）的权重。
- 参数 γ 确定向对角矩阵的收缩量。

该方法允许您利用正则化的两个特征，提高了二次判别分析估计值的稳定性。请参见 Friedman (1989)。请参见表 5.2 了解相关符号。

对于正则方法，组 *t* 的协方差矩阵为：

$$\Sigma_t = (1-\gamma)(\lambda\mathbf{S}_p + (1-\lambda)\mathbf{S}_t) + \gamma Diag((\lambda\mathbf{S}_p + (1-\lambda)\mathbf{S}_t))$$

观测 **y** 到组 *t* 的 Mahalanobis 距离按以下方式定义：

$$d_t^2 = (\mathbf{y}-\bar{\mathbf{y}}_t)'\Sigma_t^{-1}(\mathbf{y}-\bar{\mathbf{y}}_t)$$

观测 **y** 属于第 *t* 组的似然函数按以下方式估计：

$$\begin{aligned} l_t(\mathbf{y}) &= (2\pi)^{-T/2}\left|\Sigma_t\right|^{-1/2}\exp(-(\mathbf{y}-\bar{\mathbf{y}}_t)'\Sigma_t^{-1}(\mathbf{y}-\bar{\mathbf{y}}_t)/2) \\ &= (2\pi)^{-T/2}\left|\Sigma_t\right|^{-1/2}\exp(-d_t^2/2) \end{aligned}$$

按以下方式计算组 *t* 的成员关系的后验概率：

$$\begin{aligned} p(t|\mathbf{y}) &= (q_t l_t(\mathbf{y}))/\left(\sum_{u=1}^{T} q_u l_u(\mathbf{x})\right) \\ &= \frac{1}{1+\sum_{u\neq t}\exp(-[(d_u^2+\log\left|\Sigma_u\right|-2\log(q_u))-(d_t^2+\log\left|\Sigma_t\right|-2\log(q_t))]/2)} \end{aligned}$$

观测 **y** 被分配给具有最大后验概率的组。

“正则”判别方法所保存的公式定义如下：

SqDist[<组 *t*>]	$d_t^2 + \log\left\|\Sigma_t\right\| - 2\log(q_t)$
Prob[<组 *t*>]	$p(t\|\mathbf{y})$
Pred <*X*>	t，对于它 $p(t\|\mathbf{y})$ 为最大值， $t = 1, ..., T$

注意：SqDist[<组 *t*>] 可为负。

宽线性判别方法

当您有很多协变量特别是协变量数超过观测数 ($p > n$) 时，“宽线性”方法很有用。该方法的核心是高效计算合并的组内协方差矩阵 $\mathbf{S}_p$ 的逆矩阵或它的转置矩阵（若 $p > n$）。它使用奇异值分解方法来避免为大的协方差矩阵反转和分配空间。

“宽线性”方法假定组内协方差矩阵相等，若观测数等于或超过协变量数，则该方法等效于“线性”方法。

宽线性计算

请参见表 5.2 了解相关符号。在 “宽线性”计算中使用以下步骤：

1. 计算组内样本均值的 T x p 矩阵 $\mathbf{M}$。$\mathbf{M}$ 的第 (t,j) 个元素 m_{tj} 是第 j 个协变量上的组 t 成员的样本均值。
2. 对于每个协变量 j，计算各组的合并标准差。称之为 s_{jj}。
3. 用 $\mathbf{S}_{diag}$ 表示具有对角元素 s_{jj} 的对角矩阵。
4. 对每个协变量的值进行中心化和统一尺度：
   - 减去观测所属组的均值。
   - 将差值除以合并标准差。

   使用符号，对于组 t 中的观测 i，第 j 个协变量的组中心化和统一尺度值为：

   $$y_{ij}^* = \frac{y_{ij} - m_{t(i)j}}{s_{jj}}$$

   符号 *t(i)* 表示观测 *i* 所属的组 *t*。
5. 用 $\mathbf{Y}_s$ 表示 y_{ij}^* 值的矩阵。
6. 用 $\mathbf{R}$ 表示组中心化和统一尺度的协变量的合并组内协方差矩阵。按以下方式计算矩阵 $\mathbf{R}$：

   $$\mathbf{R} = (\mathbf{Y}_s'\mathbf{Y}_s)/(n - T)$$

7. 将奇异值分解应用到 $\mathbf{Y}_S$：

$$\mathbf{Y}_s = \mathbf{UDV}'$$

其中 **U** 和 **V** 是正交的， **D** 是对角线上具有正元素 （奇异值）的对角矩阵。请参见 “奇异值分解”。

则 **R** 可以表示为：

$$\mathbf{R} = (\mathbf{Y}_s'\mathbf{Y}_s)/(n-T) = (\mathbf{VD}^2\mathbf{V}')/(n-T)$$

8. 若 **R** 是满秩的，按以下方式得到 $\mathbf{R}^{-1/2}$：

$$\mathbf{R}^{-1/2} = (\mathbf{VD}^{-1}\mathbf{V}')/\sqrt{n-T}$$

其中 $\mathbf{D}^{-1}$ 是对角矩阵，其对角元素为 **D** 的对角元素的逆。

若 **R** 不是满秩的，则按以下方式定义 **R** 的伪逆矩阵：

$$\mathbf{R}^- = (\mathbf{VD}^{-2}\mathbf{V}')/(n-T)$$

然后按以下方式定义 **R** 的平方根倒数：

$$(\mathbf{R}^-)^{1/2} = (\mathbf{VD}^{-1}\mathbf{V}')/\sqrt{n-T}$$

9. 若 **R** 是满秩的，结果就是 $\mathbf{R}^- = \mathbf{R}^{-1}$。因此，为了保持完整性，我们使用伪逆矩阵继续讨论。

按以下方式定义 $p \times p$ 的矩阵 $\mathbf{T}_S$：

$$\mathbf{T}_s = (\mathbf{S}_{diag}^{-1}\mathbf{VD}^{-1})/(\sqrt{n-T})$$

则：

$$(\mathbf{T}_s\mathbf{T}_s') = \left(\mathbf{S}_{diag}^{-1}\mathbf{V}(\mathbf{D}^{-1})^2\mathbf{V}'\mathbf{S}_{diag}^{-1}\right)/(n-T) = \mathbf{S}_{diag}^{-1}\mathbf{R}^-\mathbf{S}_{diag}^{-1} = \mathbf{S}_p^-$$

其中， $\mathbf{S}_p^-$ 是使用 SVD 计算的原始数据的合并组内协方差矩阵的广义逆矩阵。

Mahalanobis 距离

Mahalanobis 距离、似然函数和后验概率的公式与“线性判别方法”中所述的那些公式相同。但是，$\mathbf{S}_p$ 的逆矩阵由使用奇异值分解计算的广义逆矩阵替代。

当您保存公式时，Mahalanobis 距离以分解的形式给出。对于观测 **y**，到组 *t* 的平方距离如下所示，其中最后一个等式中的 *SqDist*[0] 和判别主成分在 “保存的公式” 中定义：

$$
\begin{aligned}
d_t^2 &= (\mathbf{y}-\bar{\mathbf{y}}_t)'\bar{\mathbf{S}}_p(\mathbf{y}-\bar{\mathbf{y}}_t) \\
&= (\mathbf{y}-\bar{\mathbf{y}}_t)'\mathbf{T}_s\mathbf{T}_s'(\mathbf{y}-\bar{\mathbf{y}}_t) \\
&= ((\mathbf{y}-\bar{\mathbf{y}})-(\bar{\mathbf{y}}_t-\bar{\mathbf{y}}))'\mathbf{T}_s\mathbf{T}_s'((\mathbf{y}-\bar{\mathbf{y}})-(\bar{\mathbf{y}}_t-\bar{\mathbf{y}})) \\
&= (\mathbf{T}_s'(\mathbf{y}-\bar{\mathbf{y}}))'(\mathbf{T}_s'(\mathbf{y}-\bar{\mathbf{y}}))-2(\mathbf{T}_s'(\bar{\mathbf{y}}_t-\bar{\mathbf{y}}))'(\mathbf{T}_s'(\mathbf{y}-\bar{\mathbf{y}}))+(\mathbf{T}_s'(\bar{\mathbf{y}}_t-\bar{\mathbf{y}}))'(\mathbf{T}_s'(\bar{\mathbf{y}}_t-\bar{\mathbf{y}})) \\
&= SqDist[0]-2(\mathbf{T}_s'(\bar{\mathbf{y}}_t-\bar{\mathbf{y}}))'\text{判别主成分}+(\mathbf{T}_s'(\bar{\mathbf{y}}_t-\bar{\mathbf{y}}))'(\mathbf{T}_s'(\bar{\mathbf{y}}_t-\bar{\mathbf{y}}))
\end{aligned}
$$

保存的公式

这些是 “宽线性”判别方法所保存的公式：

判别数据矩阵	观测值在协变量上的向量
判别主成分	使用主成分得分矩阵变换的数据，它呈现组内不相关的数据。通过 $\mathbf{T}_s'(\mathbf{y}-\bar{\mathbf{y}})$ 给出，其中 $\bar{\mathbf{y}}$ 是包含总均值的 $1 \times p$ 向量。
SqDist[0]	$(\mathbf{y}-\bar{\mathbf{y}})'\mathbf{T}'_s\mathbf{T}_s'(\mathbf{y}-\bar{\mathbf{y}})$
SqDist[<组 *t*>]	观测值到组重心的 Mahalanobis 距离。请参见 “Mahalanobis 距离” 。
Prob[<组 *t*>]	$p(t\|\mathbf{y})$，在 “线性判别方法” 中给出
Pred <*X*>	t，对于它 $p(t\|\mathbf{y})$ 为最大值，$t = 1, \ldots, T$

多元检验的统计详细信息

在下文中，**E** 是残差叉积矩阵，**H** 是模型叉积矩阵。**E** 的对角线元素是每个变量的残差平方和。**H** 的对角线元素是每个变量的模型平方和。在判别分析文献中，**E** 通常称为 **W**，其中 **W** 表示**组内**。

多元结果表中的检验统计量是 $\mathbf{E}^{-1}\mathbf{H}$ 的特征值 λ 的函数。以下列表说明了每个检验统计量的计算。

注意：在指定响应设计后，将初始 **E** 和 **H** 矩阵用 **M**′ 自左乘，然后用 **M** 自右乘。

- Wilks Lambda

$$\Lambda = \frac{\det(\mathbf{E})}{\det(\mathbf{H}+\mathbf{E})} = \prod_{i=1}^{n}\left(\frac{1}{1+\lambda_i}\right)$$

- Pillai 迹

$$V = \text{迹}[\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}] = \sum_{i=1}^{n}\frac{\lambda_i}{1+\lambda_i}$$

- Hotelling-Lawley 迹

$$U = \text{迹}(\mathbf{E}^{-1}\mathbf{H}) = \sum_{i=1}^{n}\lambda_i$$

- Roy 最大根

$\Theta = \lambda_1$，$\mathbf{E}^{-1}\mathbf{H}$ 的最大特征值。

E 和 **H** 定义如下：

$$\mathbf{E} = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'(\mathbf{X}'\mathbf{X})\mathbf{b}$$

$$\mathbf{H} = (\mathbf{L}\mathbf{b})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}')^{-1}(\mathbf{L}\mathbf{b})$$

其中 **b** 是模型系数的估计向量，$\mathbf{A}^{-}$ 表示矩阵 **A** 的广义逆。

整体模型 **L** 是与以下单位矩阵拼接而成的零列（用于截距）：该矩阵的行数和列数与模型中的参数数目相等。效应的 **L** 矩阵是整体模型 **L** 矩阵的行的子集。

近似 F 检验的统计详细信息

为了计算 F 值和自由度，用 p 表示 $\mathbf{H}+\mathbf{E}$ 的秩。用 q 表示 $\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'$ 的秩，其中 **L** 矩阵标识与要检验的效应关联的 $\mathbf{X}'\mathbf{X}$ 的元素。用 v 表示误差自由度，用 s 表示 p 和 q 中的最小项。还假定 $m = 0.5(|p-q|-1)$ 和 $n = 0.5(v-p-1)$。

表 5.3 列出了从相应的检验统计量如何计算每个近似 F 值。

表 5.3 近似 F 统计量

检验	近似 F	分子自由度	分母自由度
Wilks Lambda	$F = \left(\frac{1-\Lambda^{1/t}}{\Lambda^{1/t}}\right)\left(\frac{rt-2u}{pq}\right)$	pq	$rt-2u$

表 5.3 近似 F 统计量

检验	近似 F	分子自由度	分母自由度
Pillai 迹	$F = \left(\frac{V}{s-V}\right)\left(\frac{2n+s+1}{2m+s+1}\right)$	$s(2m+s+1)$	$s(2n+s+1)$
Hotelling-Lawley 迹	$F = \frac{2(sn+1)U}{s^2(2m+s+1)}$	$s(2m+s+1)$	$2(sn+1)$
Roy 最大根	$F = \frac{\Theta(v-\max(p,q)+q)}{\max(p,q)}$	$\max(p,q)$	$v-\max(p,q)+q$

组间协方差矩阵的统计详细信息

使用表 5.2 中的符号，“判别”平台中的组间协方差矩阵定义如下：

$$\mathbf{S}_B = \frac{1}{T-1}\sum_{t=1}^{T} T\left(\frac{n_t}{n}\right)(\bar{\mathbf{y}}_t - \mathbf{y}_{bar})(\bar{\mathbf{y}}_t - \mathbf{y}_{bar})'$$

第 6 章

偏最小二乘模型

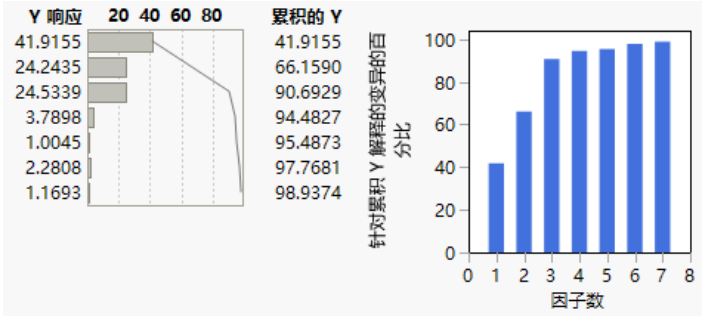
使用 Y 和 X 之间的相关性构建模型

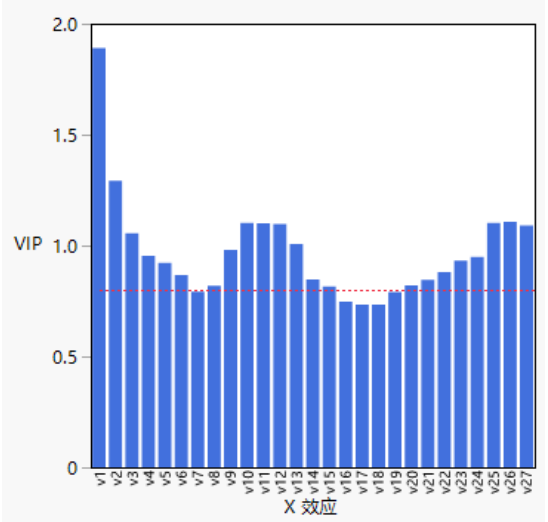
“偏最小二乘”(PLS) 平台基于因子（即解释变量 (*X*) 的线性组合）来拟合线性模型。这些因子是通过将 *X* 与一个或多个响应 (*Y*) 之间的协方差最大化得到的。“偏最小二乘”利用 *X* 和 *Y* 之间的相关性揭示底层的潜在结构。

JMP PRO JMP Pro 提供更多功能，从而支持您构造“PLS 判别分析 (PLS-DA)”，包括各种模型效应、利用几种验证方法、补缺缺失数据以及获取各种统计量分布的 Bootstrap 估计值。

偏最小二乘适用于以下情况：X 变量数比观测数多；X 变量之间高度相关；X 变量非常多；只有几个 Y 变量但有很多 X 变量，此时使用普通最小二乘法将无法得到满意的结果。

图 6.1 “偏最小二乘”报表的一部分





目录

“偏最小二乘”平台概述

与普通最小二乘法相比，当预测变量数比观测数多时，可以使用 PLS。PLS 广泛用于对高维数据建模的领域中，如光谱学、化学计量学、基因组学、心理学、教育、经济学、政治学和环境学。

当解释变量数比观测数多或解释变量高度相关时，使用 PLS 方法拟合模型很有用。您可以使用 PLS 来同时对几个响应变量拟合一个模型。请参见 Garthwaite (1994), Wold (1994), Wold et al. (2001)、Eriksson et al. (2006) 和 Cox and Gaudard (2013)。

该平台提供两种模型拟合算法：非线性迭代偏最小二乘 (NIPALS) 和“PLS 的统计启发修改” (SIMPLS)。有关 NIPALS 的详细信息，请参见 Wold (1980)。有关 SIMPLS 的详细信息，请参见 De Jong (1993)。有关两种方法的说明，请参见 Boulesteix and Strimmer (2007)。SIMPLS 算法开发的目的在于解决特定的最优问题。对于单个响应，这两种方法会给出相同的模型。对于多重响应，它们略有不同。

在 JMP 中，只能通过“分析”>“多元方法”>“偏最小二乘”来访问 PLS 平台。在 JMP Pro 中，您还可以通过“分析”>“拟合模型”来访问“偏最小二乘”特质。

JMP PRO 在 JMP Pro 中，您可以：

- 使用“拟合模型”中的“偏最小二乘”特质，通过拟合具有名义型建模类型的响应来执行 PLS-DA（PLS 判别分析）。
- 使用“拟合模型”中的“偏最小二乘”特质拟合多项式、交互作用和分类效应。
- 在几种验证和交叉验证方法中进行选择。
- 补缺缺失数据。
- 获取各种统计量分布的 Bootstrap 估计值。右击关注的报表。有关 Bootstrap 估计值的详细信息，请参见《基本分析》。

“偏最小二乘”使用 van der Voet T^2 检验和交叉验证来帮助您选择要提取的最佳因子数。

- 在 JMP 中，该平台使用“留一法”交叉验证。您也可以选择不使用验证。
- JMP PRO 在 JMP Pro 中，您可以选择“K 重”、“留一法”或“随机保留”交叉验证，也可以指定一个验证列。您也可以选择不使用验证。

“偏最小二乘”示例

使用“偏最小二乘”平台构建一个模型，用于预测波罗的海海水样本中存在的三种不同污染化合物的数量。关注的三种化合物分别为：

- 木质素磺酸盐，它是纸浆工业产生的污染
- 黑腐酸，它是天然森林的产物
- 洗涤剂产生的增白剂

预测变量表示为在某个波长范围内 (v1–v27) 测量的光谱发射强度。本例来自光谱仪校准，这是偏最小二乘很有效的领域。

为了校准模型，使用了已知成分的样本。校准数据包含已知木质素磺酸盐、黑腐酸和洗涤剂浓度的 16 个样本。记录了 27 个等距波长下的发射强度。

1. 选择**帮助** > **样本数据文件夹**，然后打开 Baltic.jmp。

   注意: Baltic.jmp 数据表中的数据在 Umetrics (1995) 中报告。原始来源是 Lindberg, Persson and Wold (1983)。

2. 选择**分析** > **多元方法** > **偏最小二乘**。
3. 将 ls、 ha 和 dt 分配给 **Y，响应**角色。
4. 将 Intensities （它包含 27 个强度变量 v1-v27）分配给 **X，因子**角色。
5. 点击**确定**。

   随即显示 “偏最小二乘模型启动”控制面板。
6. 选择**留一法**作为验证方法。
7. 点击**执行**。

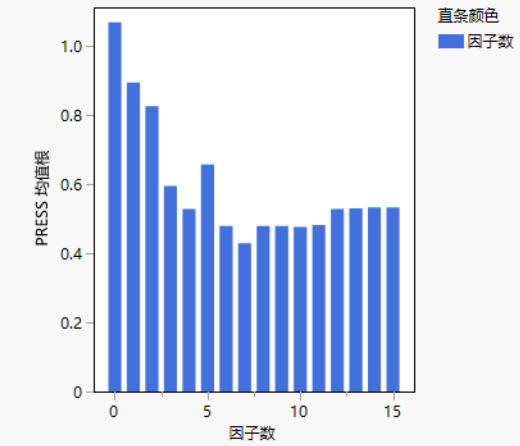
   因为 van der Voet 检验是随机化检验，您的 “概率 > van der Voet T^2”值可能略有不同。

图 6.2 “偏最小二乘”报表

留一法交叉验证 (方法为“NIPALS”，使用 快速 SVD)

因子数	PRESS 均值根	van der Voet T²	概率 > van der Voet T²	Q²	累积 Q²	R²X	累积 R²X	R²Y	累积 R²Y
0	1.066667	10.070473	<.0001*	-0.137778	-0.137778	0.000000	0.000000	0.000000	0.000000
1	0.892851	11.141942	<.0001*	-0.098077	-0.098077	0.974480	0.974480	0.421421	0.421421
2	0.824551	8.605908	<.0001*	0.773385	0.751159	0.022034	0.996514	0.247724	0.669146
3	0.593922	8.034522	0.0080*	0.774570	0.943904	0.001772	0.998286	0.241683	0.910829
4	0.527742	7.879423	0.0020*	0.656229	0.980716	0.001153	0.999438	0.037043	0.947872
5	0.656013	6.883024	0.0250*	0.356257	0.987586	0.000386	0.999825	0.012502	0.960374
6	0.478479	3.446269	0.3360	0.640049	0.995532	0.000113	0.999938	0.019350	0.979724
7	0.428141	0.000000	1.0000	0.809320	0.999148	0.000017	0.999955	0.010643	0.990367
8	0.478562	7.031041	0.0130*	0.818516	0.999845	0.000011	0.999967	0.004666	0.995032
9	0.478739	3.066627	0.4040	0.810850	0.999971	0.000012	0.999979	0.001452	0.996485
10	0.476024	3.847943	0.2480	0.845631	0.999995	0.000007	0.999986	0.001417	0.997902
11	0.481440	4.822874	0.1310	0.820847	0.999999	0.000004	0.999990	0.001248	0.999150
12	0.527436	3.968884	0.2030	0.810490	1.000000	0.000004	0.999994	0.000637	0.999787
13	0.529249	4.028960	0.2100	0.812180	1.000000	0.000003	0.999997	0.000158	0.999945
14	0.531553	4.408278	0.1610	0.816103	1.000000	0.000003	1.000000	0.000055	1.000000
15	0.531553	4.408278	0.1640	0.816103	1.000000	0.000000	1.000000	0.000000	1.000000

PRESS 均值根图

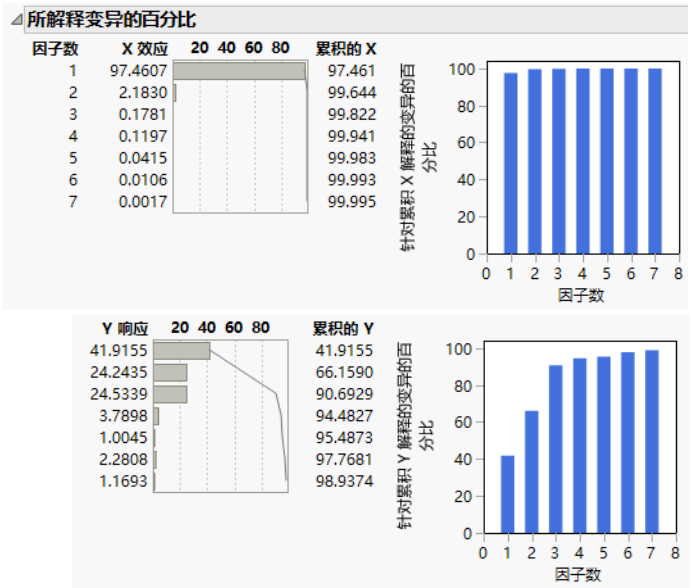


注意: 最小 PRESS 均值根是 0.42814，且最低因子数是 7。

“PRESS (预测残差平方和) 均值根图”显示当因子数为 7 时 PRESS 均值根的值最小。这在“PRESS 均值根图”下的注释中有说明。生成名为**带 7 个因子使用快速 SVD 的“NIPALS”拟合**的报表。该报表的一部分显示在图 6.3 中。

van der Voet T^2 统计量检验具有不同因子数的模型是否与具有最小 PRESS 值的模型显著不同。常见做法是提取 van der Voet 显著性水平超过 0.10 的最小因子数 (SAS Institute Inc, 2020f; Tobias 1995)。若您要在此处应用该方法，可以通过在**模型启动**面板中输入 6 作为**因子数**来拟合新模型。

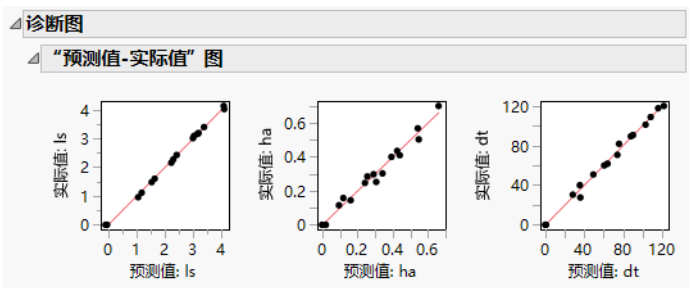
图 6.3 提取的七个因子



8. 点击 “带 7 个因子使用快速 SVD 的 “NIPALS” 拟合” 红色小三角并选择**诊断图**。

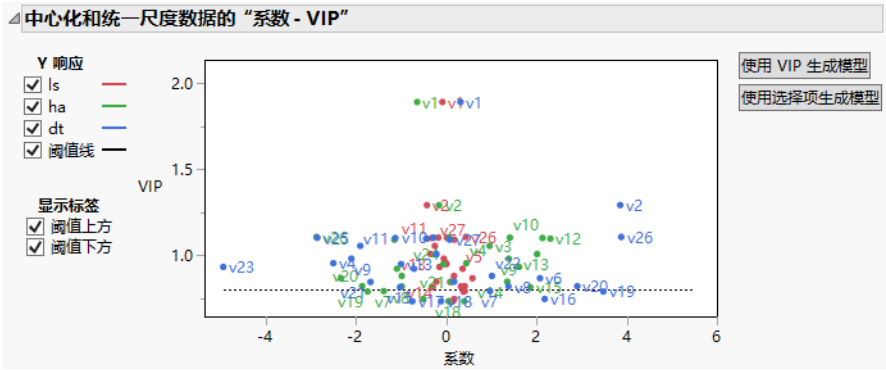
   这得到一个报表用于显示 “预测值 - 实际值” 图以及三个报表用于显示各种残差图。“‘预测值 - 实际值’ 图” 显示预测的化合物含量与实际含量相一致的程度。

图 6.4 诊断图



9. 选择 “带 7 个因子使用快速 SVD 的 “NIPALS” 拟合” 红色小三角并选择**系数 -VIP 图**。

图 6.5 系数 -VIP 图



“系数 -VIP 图”帮助识别对拟合多个响应有影响的变量。例如， v23、 v2 和 v26 都有超过 0.8 的 VIP 值和相对较大的系数。

启动“偏最小二乘”平台

有两种启动 “偏最小二乘”平台的方法：

- 选择**分析** > **多元方法** > **偏最小二乘**。
- JMP PRO 选择**分析** > **拟合模型**，然后从“特质”菜单中选择**偏最小二乘**。该方法允许您执行以下操作：
  - 输入分类变量作为 *Y* 或 *X*。通过输入分类 *Y* 来执行 PLS-DA。
  - 将交互作用项和多项式项添加到您的模型。
  - 使用“标准化 X”选项以使用中心化和统一尺度的列构造高阶项。
  - 保存您的模型规格脚本。

“拟合模型”启动窗口上的一些功能不适用于 “偏最小二乘”特质：

- “权重”、“嵌套”、“特性”、“变换”和“无截距”。

  提示：您可以通过在“选择列”框中右击某个变量并选择一个“变换”选项来变换该变量。

- 下面的宏：混料响应曲面和 Scheffé 三次项。

图 6.6 JMP Pro “偏最小二乘”启动窗口 （选择了 EM 作为补缺方法）

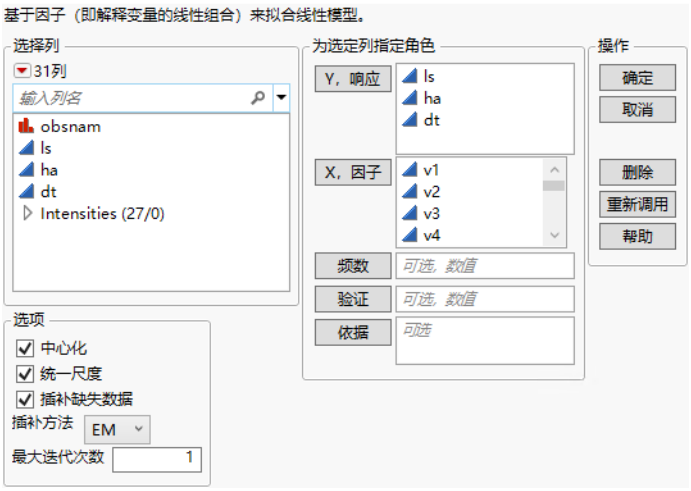
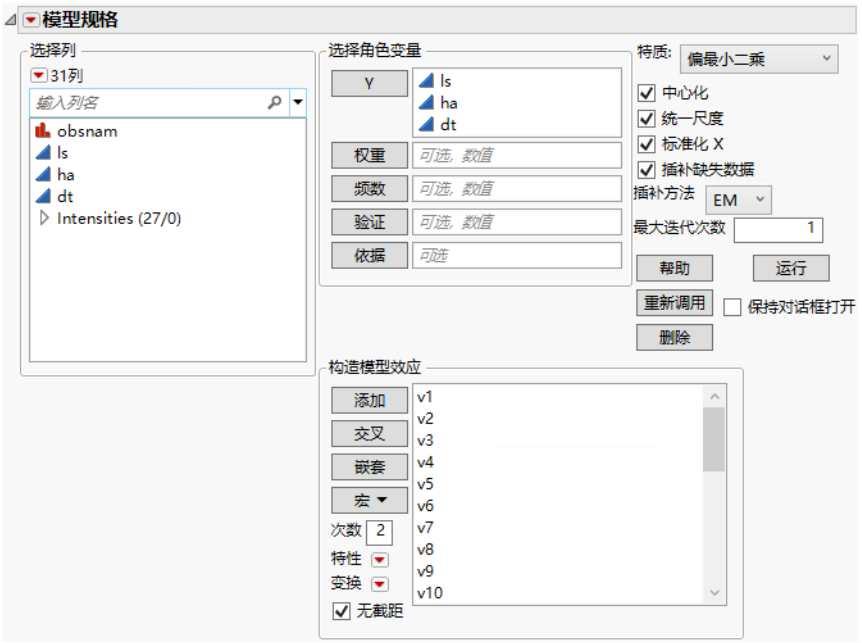


图 6.7 JMP Pro 拟合模型 “偏最小二乘”启动窗口



有关“拟合模型”启动选项的详细信息，请参见《拟合线性模型》。

有关“选择列”红色小三角菜单中选项的详细信息，请参见《使用 JMP》。

“偏最小二乘”启动窗口包含以下选项：

Y，响应 输入数值响应列。若您输入了多个列，则它们联合建模。

JMP PRO 在 JMP Pro 中，您可以在“拟合模型”启动窗口中输入名义型响应列来执行 PLS-DA。请参见 “PLS 判别分析的统计详细信息”。

X，因子 输入预测变量列。“偏最小二乘”启动窗口只允许数值型预测变量。

JMP PRO 在 JMP Pro 中，您可以在 “拟合模型”启动窗口中输入名义型和有序型模型效应。有序型效应被视为名义型处理。

频数 若您的数据进行了汇总，则输入其值包含各行计数的列。

JMP PRO **验证** 用于定义验证集的数值列。验证列只能包含连续的整数值。

- 若验证列有两个水平，则较小的值定义训练集，较大的值定义验证集。
- 若验证列有三个水平，则这些值按大小递增的顺序定义训练集、验证集和测试集。
- 若验证列有三个以上的水平，则使用“K 重交叉验证”。有关其他验证选项的信息，请参见“验证方法”。

PLS 平台使用验证列来训练和微调模型，或者训练、微调和评估模型。有关验证的详细信息，请参见 《预测和专业建模》。

注意：若在 “选择列”列表中没有选择任何列的情况下点击 “验证”按钮，您可以向数据表添加一个验证列。有关“生成验证列”实用工具的详细信息，请参见《预测和专业建模》。

依据 输入一列，用于创建为变量的每个水平包含单独分析的报表。若分配了多个 “依据”变量，则为 “依据”变量水平的每个可能组合生成单独分析。

中心化 通过从每个列减去均值将所有 Y 变量和模型效应中心化。请参见 “中心化和统一尺度”。

统一尺度 通过将每个列除以其标准差对所有 Y 变量和模型效应统一尺度。请参见 “中心化和统一尺度”。

JMP PRO **标准化 X** （仅可用于“拟合模型”启动窗口。）将构造模型效应时使用的所有列中心化和统一尺度。若未选择该选项，则使用原始数据表列构造高阶效应。然后基于所选的 “中心化”和 “统一尺度”选项将每个高阶效应中心化或统一尺度。请注意，“标准化 X”选项不将 Y 变量中心化或统一尺度。请参见 “标准化 X”。

JMP PRO **补缺缺失数据** 使用非缺失值替代 Y 或 X 中的缺失数据值。从**补缺方法**列表中选择合适的方法。

若未选择**补缺缺失数据**，则从分析中排除在任何 X 变量上具有缺失观测的行，而且不为这些行计算预测值。此外还会从分析中排除在 X 变量上不具有缺失观测但是在 Y 变量上具有缺失观测的行，但是计算预测值。

JMP PRO **补缺方法** （仅在**补缺缺失数据**时显示。）从以下补缺方法中选择：

均值 对于每个模型效应或响应列，使用非缺失值的均值替代缺失值。

EM 使用迭代期望值最大化 (EM) 方法来补缺缺失值。在第一次迭代时，使用均值替代效应或响应的缺失值，对数据拟合指定的模型。使用 Y 模型的预测值和 X 模型的预测值来补缺缺失值。对于后续迭代，在使用当前估计值给出条件分布时，使用其预测值替代缺失值。

为了进行补缺，多项式中的项被视为单独的预测变量处理。当指定了多项式中的项时，将根据原始数据计算该项；若选中了“标准化 X”，则根据标准化的列值计算该项。若某一行在定义多项式中的项所涉及的列中存在缺失值，则该条目对于多项式中的项是缺失的。使用这样定义的多项式中的项执行补缺。

有关 EM 方法的详细信息，请参见 Nelson, Taylor and MacGregor (1996)。

JMP PRO **最大迭代次数** （仅在将 EM 选作补缺方法时才显示。）支持您设置算法使用的最大迭代次数。若缺失值的当前估计值和前一估计值之间的最大差值不超过 10^{-8}，算法将终止。

在完成启动窗口并点击**确定**后，将显示“模型启动”控制面板。请参见“‘模型启动’控制面板”。

中心化和统一尺度

默认选定“偏最小二乘”的“中心化”和“统一尺度”选项。这意味着预测变量和响应会中心化和统一尺度以具有均值 0 和标准差 1。将预测变量和响应中心化可以使它们相对于变异是对等的。若没有中心化，则在构建后续因子中会涉及变量的均值以及围绕该均值的变异。为了进行说明，假定**时间**和**温度**是两个预测变量。对它们统一尺度表示**时间**的一个标准差变化大约等效于**温度**的一个标准差变化。

JMP PRO 标准化 X

在“拟合模型”窗口中选择“偏最小二乘”特质时，默认选定“标准化 X”选项。这确保作为模型效应输入的所有列以及交互作用项或多项式项涉及的所有列被标准化。

假定您有两个列 X1 和 X2，并且在 “拟合模型”窗口中输入交互作用项 X1*X2 作为模型效应。当选定 “标准化 X”选项时，在形成交互作用项前已将 X1 和 X2 中心化和统一尺度。形成的交互作用项按以下方式计算：

$$\left(\frac{X1 - mean(X1)}{std(X1)}\right) \times \left(\frac{X2 - mean(X2)}{std(X2)}\right)$$

在进入模型前，根据您对**中心化**和**统一尺度**选项的选择，将所有模型效应中心化或统一尺度。

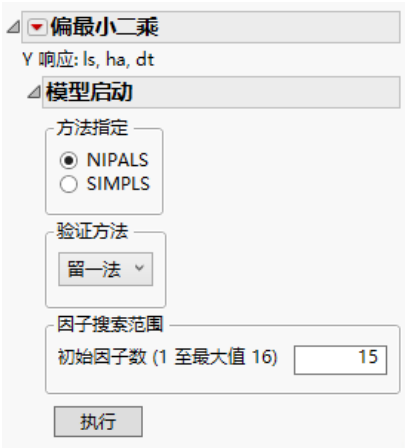
若未选择 “标准化 X”选项但是选择了**中心化**和**统一尺度**，则按以下方式计算进入模型的项：

$$\frac{X1 \times X2 - mean(X1 \times X2)}{std(X1 \times X2)}$$

“模型启动”控制面板

在“偏最小二乘”启动窗口中点击“确定”（或在“拟合模型”启动窗口中点击“运行”）后，将显示“模型启动”控制面板。

图 6.8 “偏最小二乘模型启动”控制面板



注意：在 JMP Pro 中，显示的“模型启动”控制面板的“验证方法”部分有所不同。

“模型启动”控制面板包含以下选项：

方法指定 选择模型拟合算法的类型。有两个算法选项：NIPALS 和 SIMPLS。当只有一个响应变量时，这两种方法生成相同的系数估计值。有关两种算法之间的差异的详细信息，请参见“‘偏最小二乘’平台的统计详细信息”。

验证方法 选择验证方法。验证用于确定要提取的最佳因子数。对于 JMP Pro，若在平台启动窗口中指定了验证列，则不显示这些选项。

JMP PRO **保留** 随机选择指定比例的数据作为验证集，并使用除验证集外的数据拟合模型。随机选择基于对模型因子的分层抽样，意图创建比基于简单随机抽样的训练集和验证集更平衡的训练集和验证集。

JMP PRO **K 重** 将数据分割为 *K* 个子集或**重**。依次使用每个重验证拟合其余数据的模型（拟合总共 *K* 个模型）。该方法最合适小数据集，因为它有效使用了有限的数据量。

留一法 执行留一法交叉验证。

无 不使用验证来选择要提取的因子数。在“因子搜索范围”中指定因子数。

因子搜索范围 指定在未使用验证时要提取多少个潜在因子。若使用验证，则这是在选择最佳因子数前平台尝试拟合的最大因子数。最大因子数是非缺失行数和因子数中的最小值。于是，初始因子数是 15 和最大因子数中的最小值。

因子规格 在点击**执行**以拟合初始模型后显示。指定要在拟合新模型中使用的因子数。

执行 使用给定的规格启动模型拟合。

“偏最小二乘”选项

在“偏最小二乘”红色小三角菜单中，您可以在点击“模型启动”控制面板中的“执行”之前使用以下选项：

JMP PRO **设置随机种子** 设置用于“K 重”和“保留”验证的随机化过程的种子。若您要重新生成分析，该选项会很有用。将种子设置为正值然后保存脚本，该种子将自动保存在脚本中。运行该脚本可以始终生成相同的交叉验证分析。当“验证方法”设置为“无”或使用验证列时，不显示该选项。

JMP PRO **SVD** 显示一个子菜单，该子菜单支持您在模型拟合算法中选择 SVD 算法的快速实现或经典实现。默认值为“快速 SVD”。

常规选项

请参见《使用 JMP》获取有关下列选项的信息：

本地数据过滤器 显示或隐藏支持您过滤特定报表中使用的数据的本地数据过滤器。

重新运行 包含使您可以重复或重新启动分析的选项。在支持该功能的平台中，“自动重新计算”选项立即在相应报表窗口中反映您对数据表所做的更改。

平台首选项 包含的选项支持您查看当前平台首选项或更新平台首选项以匹配当前 JMP 报表中的设置。

保存脚本 包含的选项支持您保存可将报表重现到若干目标的脚本。

保存“依据”组脚本 包含使您可以保存脚本的选项，可将为“依据”变量的所有水平重新生成平台报表的脚本保存到多个不同的位置。仅当在启动窗口中指定“依据”变量时才可用。

“偏最小二乘”报表

“偏最小二乘”报表的外观取决于是否使用验证方法。若您指定了验证列或在“验证方法”面板中选择了“保留”，则报表中的所有模型拟合均基于训练数据。否则，所有模型拟合均基于整个数据集。

若您使用了验证，则显示三个报表：

- 模型比较汇总
- “交叉验证”报表
- 带 <N> 个因子的 NIPALS（或 SIMPLS）拟合

若您选择了**无**作为交叉验证方法，则显示两个报表：

- 模型比较汇总

- 带 <N> 个因子的 NIPALS（或 SIMPLS）拟合

要拟合更多模型，请在“模型启动”面板中指定所需的因子数。

模型比较汇总

“偏最小二乘”报表中的“模型比较汇总”表显示每个拟合模型的汇总结果。

图 6.9 模型比较汇总

模型比较汇总

方法	SVD	行数	因子数	针对累积 X 解释的变异的百分比	针对累积 Y 解释的变异的百分比	VIP > 0.8 的数目
NIPALS	快速	16	7	99.995152	98.937438	22
NIPALS	快速	16	6	99.993471	97.768092	22

报表包含以下汇总信息：

方法 显示您在 “模型启动”控制面板中指定的分析方法。

SVD 显示指定的 SVD 算法实现。

行数 显示训练集中使用的观测数。

因子数 显示提取的因子数。

针对累积 X 解释的变异的百分比 显示模型解释的 X 的变异百分比。

针对累积 Y 解释的变异的百分比 显示模型解释的 Y 的变异百分比。

VIP>0.8 的数目 显示 VIP（变量投影重要性）值大于 0.8 的模型效应数。VIP 得分是衡量变量相对 X 和 Y 建模的重要性的指标 (Wold 1994; Eriksson et al. 2006)。

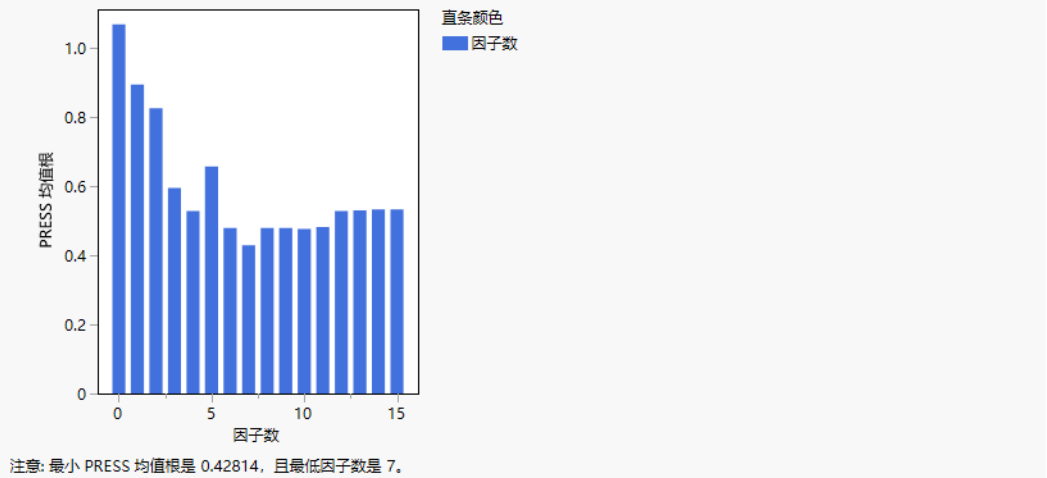
“交叉验证”报表

仅当在“偏最小二乘”的“模型启动”控制面板中将某种交叉验证方式选为“验证方法”时，才显示该报表。报表标题被动态地命名为 <交叉验证方法> 且方法 = <方法指定>，具体取决于在控制面板中选定的交叉验证和方法选项。它显示模型拟合的汇总统计量，拟合所用的因子数为 0 到提取的最大因子数（在“模型启动”控制面板中指定）。该报表还提供“PRESS 均值根值图”。请参见“PRESS 均值根图”。使用最小 PRESS 均值根统计量标识最佳因子数。

图 6.10 “交叉验证”报表

留一法交叉验证（方法为“NIPALS”，使用快速 SVD）

因子数	PRESS 均值根	van der Voet T²	概率 > van der Voet T²	Q²	累积 Q²	R²X	累积 R²X	R²Y	累积 R²Y
0	1.066667	10.070473	<.0001*	-0.137778	-0.137778	0.000000	0.000000	0.000000	0.000000
1	0.892851	11.141942	<.0001*	-0.098077	-0.098077	0.974480	0.974480	0.421421	0.421421
2	0.824551	8.605908	<.0001*	0.773385	0.751159	0.022034	0.996514	0.247724	0.669146
3	0.593922	8.034522	0.0080*	0.774570	0.943904	0.001772	0.998286	0.241683	0.910829
4	0.527742	7.879423	0.0020*	0.656229	0.980716	0.001153	0.999438	0.037043	0.947872
5	0.656013	6.883024	0.0250*	0.356257	0.987586	0.000386	0.999825	0.012502	0.960374
6	0.478479	3.446269	0.3360	0.640049	0.995532	0.000113	0.999938	0.019350	0.979724
7	0.428141	0.000000	1.0000	0.809320	0.999148	0.000017	0.999955	0.010643	0.990367
8	0.478562	7.031041	0.0130*	0.818516	0.999845	0.000011	0.999967	0.004666	0.995032
9	0.478739	3.066627	0.4040	0.810850	0.999971	0.000012	0.999979	0.001452	0.996485
10	0.476024	3.847943	0.2480	0.845631	0.999995	0.000007	0.999986	0.001417	0.997902
11	0.481440	4.822874	0.1310	0.820847	0.999999	0.000004	0.999990	0.001248	0.999150
12	0.527436	3.968884	0.2030	0.810490	1.000000	0.000004	0.999994	0.000637	0.999787
13	0.529249	4.028960	0.2100	0.812180	1.000000	0.000003	0.999997	0.000158	0.999945
14	0.531553	4.408278	0.1610	0.816103	1.000000	0.000003	1.000000	0.000055	1.000000
15	0.531553	4.408278	0.1640	0.816103	1.000000	0.000000	1.000000	0.000000	1.000000



注意: 最小 PRESS 均值根是 0.42814，且最低因子数是 7。

JMP PRO 当选定**标准化 X** 选项时，标准化会一次性应用到整个数据表。它不会重新应用到各个训练集。但是，当选定**中心化**或**统一尺度**选项的任意组合时，该选择组合会应用到每个交叉验证训练集。使用这些训练集执行交叉验证，若选定这些选项则会对它们单独进行中心化和统一尺度。

报表中显示以下统计量。若使用了任何形式的验证或交叉验证，则报告的结果是训练集统计量的汇总。

因子数 拟合模型时使用的因子数。

PRESS 均值根 所有响应的 PRESS 值的平均值的平方根。请参见 “PRESS 均值根”。

van der Voet T^2 van der Voet 检验的统计量，它检验具有不同提取因子数的模型是否与最佳模型存在显著差异。每个 van der Voet T^2 检验的原假设假定基于相应因子数的模型与最佳模型没有差异。备择假设是该模型与最佳模型有差异。请参见 “van der Voet T^2 检验的统计详细信息”。

概率 > van der Voet T^2 van der Voet T^2 检验的 *p* 值。请参见“van der Voet T^2 检验的统计详细信息”。

Q^2 预测能力的无量纲测度，它定义为：从 1 中减去 PRESS 值除以 Y 的总平方和所得的比值，即：

1 – PRESS/SSY

请参见“Q^2 的计算”。

累积 Q^2 具有给定因子数或更少因子数的模型的预测能力指标。对于给定的因子数 *f*，按以下方式定义累积 Q^2：

$$1 - \prod_{i=1}^{f} (\text{Press}_i/\text{SYY}_i)$$

此处 $PRESS_i$ 和 SSY_i 对应于它们针对 *i* 个因子的值。

R^2X 由指定因子解释的 X 变异的百分比。具有较大 R^2X 的成分解释 X 变量中的大部分变异。请参见“使用验证时 R^2X 和 R^2Y 的计算”。

累积 R^2X 具有给定因子数的模型所解释的 X 变异的百分比。它是 R^2X 值之和（$i = 1$ 到给定的因子数）。

R^2Y 由指定因子解释的 Y 变异的百分比。具有较大 R^2Y 的成分解释 Y 变量中的大部分变异。请参见“使用验证时 R^2X 和 R^2Y 的计算”。

累积 R^2Y 具有给定因子数的模型所解释的 Y 变异的百分比。它是 R^2Y 值之和（$i = 1$ 到给定的因子数）。

Q^2 和累积 R^2Y 的解释

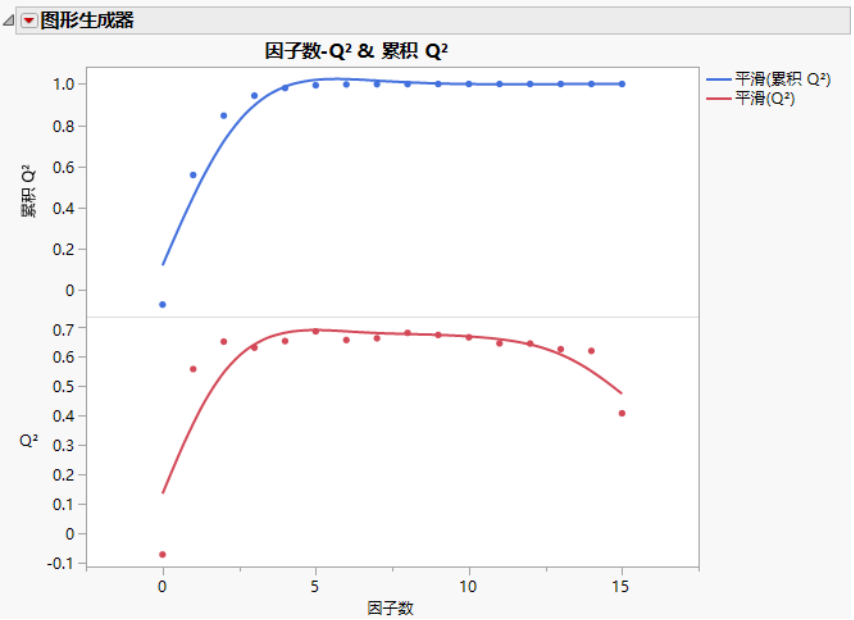
统计量 Q^2 和累积 R^2Y 均可衡量模型的预测能力，但方法不同。

- 累积 R^2Y 随因子数的增加而增大。这是因为随着因子添加至模型，更多的变异得到解释。
- Q^2 的趋势是随因子数的增加，先增大后减小（或至少停止增大）。这是因为随着更多的因子添加至模型，模型开始调整训练集，对于新数据却得不到较好的结果，导致 PRESS 统计量减小。

Q^2 和累积 R^2Y 分析是 van der Voet 检验的一个替代方法，用于确定要在模型中包括多少因子。选择一个因子数，使得 Q^2 较大且没有开始减小。您还希望累积 R^2Y 较大。

图 6.11 显示 Penta.jmp 数据表针对因子数标绘的累积 R^2Y 和 Q^2，验证方法为“留一法”。累积 R^2Y 增加，并且在大约四个因子时开始趋于平稳。统计量 Q^2 在两个因子时最大，然后开始趋于平稳。该图表明具有两个因子的模型将能够解释 Y 中的大部分变异，且不会过拟合数据。

图 6.11 Penta.jmp 的累积 R^2Y 和 Q^2



PRESS 均值根图

该条形图在水平轴上显示因子数，在垂直轴上显示“PRESS 均值根”值。它等效于在“交叉验证”报表（图 6.10）中显示在“PRESS 均值根”列右侧的水平条形图。

PRESS 均值根

对于指定的因子数 *a*，使用以下步骤计算 PRESS 均值根：

1. 对每个训练集拟合具有 *a* 个因子的模型。
2. 将得到的预测公式应用到验证集中的观测。
3. 对于每个 Y：
   - 对于每个验证集，计算每个验证集的观测值与其预测值的差值平方（预测误差平方）。
   - 对于每个验证集，求这些差值平方的平均值并将结果除以响应的方差估计值。对于“K 重”和“留一法”验证方法，除以整个响应列的方差。对于“保留”验证方法，除以训练集中响应值的方差。
   - 将这些均值相加，若有多个验证集时，则将它们的总和除以验证集数减 1。这是给定 Y 的 PRESS 统计量。
4. *a* 个因子的 “PRESS 均值根”是所有响应的 PRESS 值的平均值的平方根。
5. 多个 Y 的 PRESS 统计量通过计算在第 3 步中获得的全部响应的 PRESS 统计量的平均值得到。

Q^2 的计算

统计量 Q^2 定义为 1 - *PRESS* / *SSY*。*PRESS* 统计量是模型的所有响应的预测误差平方和平均值，该模型基于训练集构建，但是基于验证集计算。*SSY* 的值是所有响应的 Y 的平方和平均值，这些响应基于验证集中的观测值。

根据选择的验证方法，按以下方式计算 “交叉验证”报表中的统计量 Q^2：

留一法 Q^2 是 1 - 均值 (*PRESS*) / 均值 (*SSY*) 所得的值。每个 *Y* 的 *PRESS* 通过一次留一个观测构造的模型计算得出。 *SSY* 通过每个 *Y* 列中的所有值计算得出。

K 重 Q^2 是针对验证集计算的 1 - *PRESS* / *SSY* 值的平均值，指标计算基于通过每次留 *K* 个子集中的一个构造的 *K* 个模型。

保留或验证集 Q^2 是针对验证集计算的 1 - *PRESS* / *SSY* 值的平均值，指标计算基于使用单个训练数据集构造的模型。

使用验证时 R^2X 和 R^2Y 的计算

根据选择的验证方法，按以下方式计算“交叉验证”报表中的统计量 R^2X 和 R^2Y：

注意：对于所有这些计算，以类似方式计算 R^2Y。

留一法 通过每次留一个观测来构造模型，针对这些模型的 X 效应解释的变异百分比的平均值即为 R^2X。

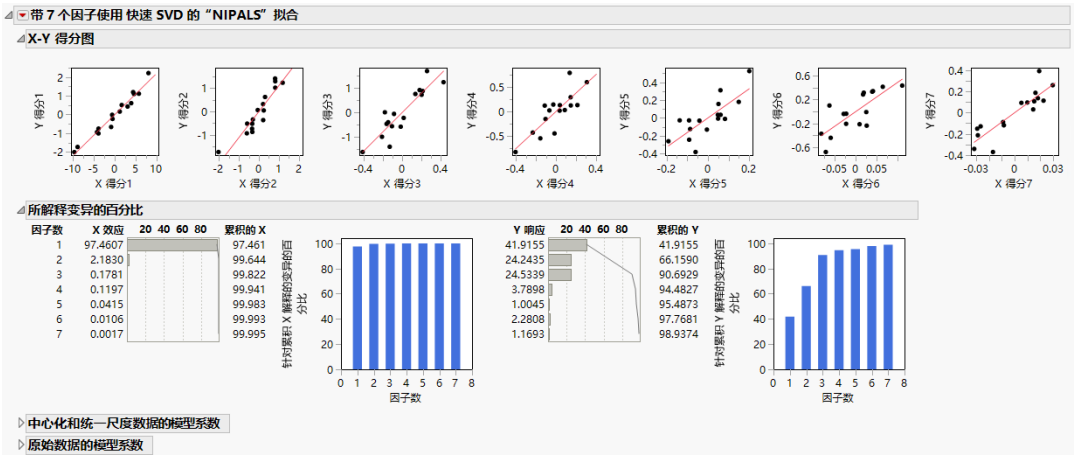
K 重 通过每次留一个子集来构造 K 个模型，针对这些模型的 X 效应解释的变异百分比的平均值即为 R^2X。

保留或验证集 使用训练数据来构造模型，针对该模型的 X 效应解释的变异百分比即为 R^2X。

“模型拟合”报表

在“偏最小二乘”平台中，“模型拟合”报表显示每个拟合模型的详细结果。该拟合使用基于交叉验证的最佳因子数，若未指定交叉验证方法，则使用指定的因子数。报表标题会指出使用的是 NIPALS 还是 SIMPLS，使用的是“快速 SVD”还是“经典 SVD”，并给出提取的因子数。

图 6.12 “模型拟合”报表



“模型拟合”报表包含以下汇总信息：

X-Y 得分图 提取的每个因子的 X 和 Y 得分的散点图。

所解释变异的百分比 显示针对 X 和 Y 解释的变异的百分比和累积百分比。针对每个提取的因子给出结果。

中心化和统一尺度数据的模型系数 对于每个 Y，显示基于中心化和统一尺度数据的模型的 X 的系数。

模型拟合选项

在 “偏最小二乘”平台中，“模型拟合”红色小三角菜单包含以下选项：

变异百分比图 添加名为“针对 X 效应解释的变异的百分比”和“针对 Y 效应解释的变异的百分比”的两个图。它们显示堆叠条形图，分别表示针对 *X* 和 *Y* 的每个提取的因子所解释的变异百分比。

变量重要性图 标绘每个 X 变量的 VIP 值。VIP 得分显示在“变量重要性表”中。请参见 “变量重要性图”。

系数 -VIP 图 针对模型系数标绘 VIP 统计量。您可以只显示与选定的 *Y* 对应的那些点。会额外提供标签选项。它们是根据中心化和统一尺度数据以及原始数据绘制的图。请参见 “系数 -VIP 图”。

设置 VIP 阈值 设置 “变量重要性图”、“方差重要性表”和 “系数 -VIP 图”的阈值水平。

系数图 针对 X 变量上的每个响应标绘模型系数。您可以只显示与选定的 *Y* 对应的那些点。它们是根据中心化和统一尺度数据以及原始数据绘制的图。

载荷图 对每个提取的因子标绘 X 和 Y 载荷。它们是针对 *X* 和 *Y* 单独绘制的图。

载荷散点图矩阵 显示 X 载荷和 Y 载荷的散点图矩阵。

载荷相关图 显示叠加在同一个图上的 X 和 Y 载荷的单个散点图或散点图矩阵。选择该选项时，您需要指定要标绘多少个因子。

- 若您指定两个因子，则显示单个载荷相关散点图。在图下面选择用于定义轴的两个因子。点击右箭头按钮可接连在图上显示因子的每种组合。
- 若您指定两个以上的因子，则显示一个散点图矩阵，每个矩阵单元内显示一对因子的散点图，显示的因子数由您选择的数目决定。

在这两种情况下，使用复选框来控制标签。

X-Y 得分图 包含以下选项：

拟合线 在“X-Y 得分图”上显示或隐藏穿过数据点的拟合线。

显示置信带 在“X-Y 得分图”上显示或隐藏拟合线的 95% 置信带。

得分散点图矩阵 显示 X 得分和 Y 得分的散点图矩阵。每个 X 得分散点图显示 95% 置信椭圆，该椭圆可用于离群值检测。有关置信椭圆的统计详细信息，请参见“X 得分散点图矩阵的置信椭圆的统计详细信息”。

距离图 显示以下距离的图：

- 每个观测到 X 模型的距离
- 每个观测到 Y 模型的距离
- 到 X 模型的距离与到 Y 模型的距离的散点图

在拟合效果好的模型中，X 和 Y 距离均很小，因此点接近原点 (0,0)。使用这些图可以查找相对 X 或 Y 的离群值。若一组点聚类在一起，则它们可能有共同的特性，可以单独分析。在使用验证集或验证集和测试集时，将为它们和训练集分别提供不同的报表。

T 方图 显示每个观测的 T^2 统计量以及控制限的图。基于观测在提取因子上的得分计算观测的 T^2 统计量。有关 T^2 和控制限计算的详细信息，请参见“T^2 图的统计详细信息”。

诊断图 显示诊断图以评估模型拟合效果。提供四种图：“预测值 - 实际值”图、“预测值 - 残差”图、“行号 - 残差”图和“残差正态分位数图”。这些图提供给每个响应。在使用验证集或验证集和测试集时，将为它们和训练集分别提供不同的报表。

刻画器 为每个 Y 变量显示一个刻画器。

谱刻画器 显示一个刻画器，其中所有响应变量都显示在图中的第一个单元中。该刻画器对于同时可视化 X 变量的变化对 Y 变量的影响很有用。

保存列 包含用于保存各种公式和结果的选项。

保存预测公式 将新公式列保存到原始数据表中。对于每个 Y 变量，有一个名为**预测公式: <响应>** 的列，其中包含作为 X 变量的函数的预测公式。

将预测保存为 X 得分公式 将新公式列保存到原始数据表中。对于每个 Y 变量，有一个名为**预测公式: <响应>** 的列，其中包含作为 X 得分公式的函数的预测公式。每个 X 得分的公式列也保存到数据表中。

保存预测公式的标准误差 将新公式列保存到原始数据表中。对于每个 Y 变量，有一个名为**预测值标准误差: <响应>** 的列，其中包含作为 X 变量函数的预测均值的标准误差公式。请参见“预测和置信限的统计详细信息”。

保存均值置信限公式 将新公式列保存到原始数据表中。对于每个 Y 变量，都有对应的列，其中包含作为 X 得分公式函数的响应均值的置信下限和置信上限。新列称为 **95% 均值下限: <响应>** 和 **95% 均值上限: <响应>**。这些列包含响应均值的 95% 置信限。请参见“预测和置信限的统计详细信息”。

保存单值置信限公式 将新公式列保存到原始数据表中。对于每个 Y 变量，都有对应的列，其中包含作为 X 得分公式函数的单个预测的置信下限和置信上限。新列称为 **95% 单值下限: <响应>** 和 **95% 单值上限: <响应>**。这些列包含单值的 95% 预测限。请参见“预测和置信限的统计详细信息”。

保存得分公式 将新公式列保存到原始数据表中。对于每个提取的因子，都有一个名为 **X 得分 <N> 公式**的列（其中包含 X 得分公式）以及一个名为 **Y 得分 <N> 公式**的列（其中包含 Y 得分公式）。X 得分公式是 X 变量的函数，Y 得分公式是 X 得分公式的函数。每个 X 得分公式列都具有 MDMCC 列属性，因此该列可以在“模型驱动的多元控制图” (MDMCC) 平台中使用。请参见《质量和过程方法》。有关公式的信息，请参见“偏最小二乘的统计详细信息”。

保存 Y 预测值 将新列保存到原始数据表中。对于每个 Y 变量，都有一个包含预测分类 Y 值的列。

保存 Y 残差 将新列保存到原始数据表中。对于每个 Y 变量，都有一个包含 Y 残差值的列。

保存 X 预测值 将新列保存到原始数据表中。对于每个 X 变量，都有一个包含预测 X 值的列。

保存 X 残差 将新列保存到原始数据表中。对于每个 X 变量，都有一个包含预测 X 残差值的列。

保存针对 X 效应解释的变异的百分比 将列保存到新数据表中。对于每个 X 变量，都有一列包含所有提取因子中解释的变异百分比。

保存针对 Y 响应解释的变异的百分比 将列保存到新数据表中。对于每个 Y 变量，都有一列包含所有提取因子中解释的变异百分比。

保存得分 将新列保存到原始数据表中。对于每个提取的因子，都有一列包含 X 得分，一列包含 Y 得分。

保存载荷 将列保存到新数据表中。有一个数据表包含 X 变量的载荷，还有一个数据表包含 Y 变量的载荷。

保存标准化得分 将新列保存到原始数据表中。新列包含每个提取因子的 X 和 Y 标准化得分。有关公式的信息，请参见“标准化得分和载荷的统计详细信息”。

保存标准化载荷 将列保存到新数据表中。有一个数据表包含 X 变量的标准化载荷，还有一个数据表包含 Y 变量的标准化载荷。有关公式的信息，请参见“标准化得分和载荷的统计详细信息”。

保存 T 方 将新公式列保存到原始数据表中。新列包含 T^2 公式，作为 X 变量的函数。该列中的值还用在 T 方图中。

将 T 方另存为 X 得分公式 将新公式列保存到原始数据表中。新列包含 T^2 公式，作为 X 得分公式的函数。

保存距离 将新列保存到原始数据表中。新列包含“到 X 模型的距离”(DModX) 和“到 Y 模型的距离”(DModY) 值。这些是在距离图中使用的值。

将距离另存为 X 得分公式 将新公式列保存到原始数据表中。新列包含“到 X 模型的距离”(DModX) 和“到 Y 模型的距离”(DModY) 公式，作为 X 得分公式的函数。

保存 X 权重 将列保存到新数据表中。对于每个提取的因子，有一列包含 X 变量的权重。

JMP PRO **保存验证** 将新列保存到原始数据表中。新列包含数字，这些数字指示如何在验证中使用每个观测。对于“保留”验证，该列标识行是用于训练还是验证。对于“K 重”验证，该列标识给行分配的子组编号。

JMP PRO **保存补缺** 将列保存到新数据表中。对于每个 X 和 Y 变量，都有一列包含原始数据列，其中的缺失值由其插补值替换。若指定一个验证列，则还包含该验证列。

JMP PRO **发布预测公式** 创建预测公式并将它们保存为“公式存储库”平台中的公式列脚本。若未打开“公式存储库”报表，该选项将创建“公式存储库”报表。请参见《预测和专业建模》。

JMP PRO **发布得分公式** 创建 X 和 Y 得分公式并在“公式存储库”平台中将它们保存为公式列脚本。若未打开“公式存储库”报表，该选项将创建“公式存储库”报表。请参见《预测和专业建模》。

删除拟合 从主平台报表中删除模型报表。

使用 VIP 生成模型 打开并填充一个启动窗口，其中相应的响应输入为 *Y*，其 VIP 超过指定阈值的变量输入为 *X*。执行与 “中心化和统一尺度数据的 ‘系数 - VIP’” 报表中的按钮相同的功能。请参见 “系数 -VIP 图”。

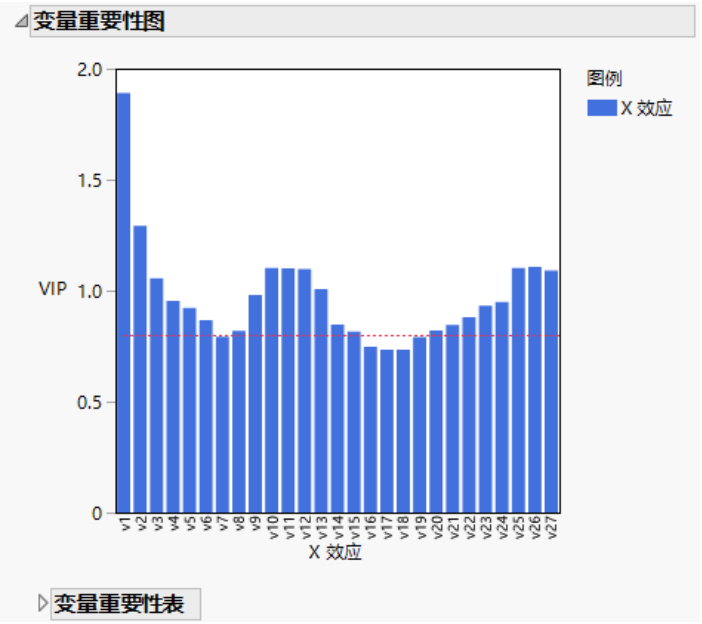
已保存 X 得分的模型驱动的多元控制图 保存每个 X 得分的公式，并启动“模型驱动多元控制图”(MDMCC) 启动窗口。在 MDMCC 启动窗口中，得分公式作为过程列分配。在点击“确定”之前，可以添加或删除过程、添加时间 ID 或设置历史数据的结束位置。请参见《质量和过程方法》。

预测值的刻画器 启动“刻画器”启动窗口。在“刻画器”启动窗口中，预测公式是针对指定模型拟合的每个变量的预测公式。您可以在点击“确定”之前添加噪声因子或其他预测公式。请参见《刻画器指南》。

变量重要性图

在“偏最小二乘”平台中，“变量重要性图”选项针对每个 X 变量的 VIP 值绘图。“变量重要性表”显示 VIP 得分。VIP 得分用于衡量变量在 X 和 Y 建模中的重要性。若变量有小的系数和小的 VIP，则它是要从模型中删除的候选项 (Wold 1994)。值为 0.8 通常视为小的 VIP (Eriksson et al. 2006)，在图上的 0.8 处绘制了一条红色虚线。

图 6.13 变量重要性图



系数 -VIP 图

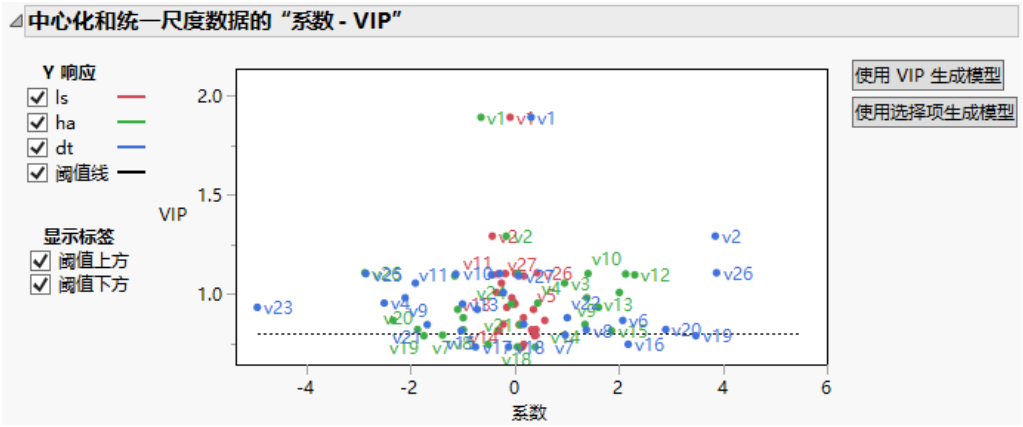
在“偏最小二乘”平台中，“系数 -VIP 图”选项针对模型系数标绘 VIP 统计量。图右侧有两个选项可帮助减少变量和生成模型。

使用 VIP 生成模型 打开并填充一个启动窗口，其中相应的响应输入为 *Y*，其 VIP 超过指定阈值的变量输入为 *X*。

使用选择项生成模型 允许您在图中直接选择 *X* 并将 *Y* 和选择的 *X* 输入启动窗口。

要基于当前的列选择使用另一平台，请打开所需的平台。请注意，在启动窗口中保留了选择内容。点击角色按钮，将填充所选的列。

图 6.14 中心化和统一尺度数据的 “系数 -VIP” 图



“偏最小二乘”的其他示例

在本例中，您使用 “拟合模型” 中的 “偏最小二乘” 特质对分类响应和大量变量进行分析。示例数据集包括从 165 名男性采集的血清样本：84 位患有前列腺癌，81 位没有前列腺癌。分析的目标是根据血清样本确定哪些男性患有前列腺癌。

1. 选择**帮助** > **样本数据文件夹**，然后打开 Prostate Cancer.jmp。
2. 选择**分析** > **拟合模型**。
3. 选择**状态**并点击 Y。
4. 选择**蛋白质**列组并点击**添加**。
5. 从 “特质” 列表中，选择**偏最小二乘**。
6. 点击**运行**。
7. （可选）点击 “偏最小二乘” 红色小三角并选择**设置随机种子**。
8. （可选）在 “指定随机种子” 旁边，输入 1234。

   通过指定随机种子，您可以重现本例中显示的结果。
9. （可选）点击**确定**。
10. 在 “模型启动” 中，点击**执行**。

图 6.15 “偏最小二乘”报表

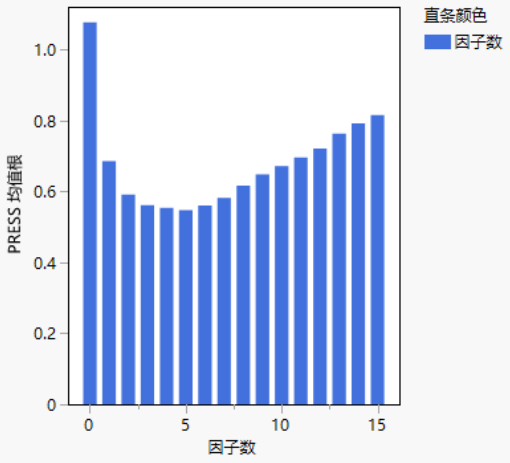
模型比较汇总

方法	SVD	行数	因子数	针对累积 X 解释的变异的百分比	针对累积 Y 解释的变异的百分比	VIP > 0.8 的数目
NIPALS	快速	165	5	52.886468	85.671251	626

K 重交叉验证，其中 K=7，方法=NIPALS，使用 快速 SVD

因子数	PRESS 均值根	van der Voet T²	概率 > van der Voet T²	Q²	累积 Q²	R²X	累积 R²X	R²Y	累积 R²Y
0	1.076926	130.46053	<.0001*	-0.000625	-0.000625	0.000000	0.000000	0.000000	0.000000
1	0.685794	16.150335	0.0010*	0.593265	0.593265	0.342651	0.342651	0.608907	0.608907
2	0.591463	2.927629	0.0790	0.696945	0.876737	0.077029	0.419680	0.132190	0.741097
3	0.561730	0.480087	0.4740	0.726683	0.966310	0.039013	0.458693	0.066929	0.808026
4	0.553949	0.278455	0.5910	0.734125	0.991043	0.043112	0.501805	0.034490	0.842515
5	0.547744	0.000000	1.0000	0.740188	0.997673	0.055223	0.557027	0.017535	0.860051
6	0.560870	1.357190	0.2350	0.728045	0.999367	0.048092	0.605120	0.015012	0.875063
7	0.581726	3.490817	0.0600	0.707162	0.999815	0.034371	0.639491	0.012864	0.887927
8	0.616081	7.673681	0.0010*	0.671501	0.999939	0.023573	0.663064	0.013162	0.901089
9	0.647871	12.869719	<.0001*	0.636867	0.999978	0.020427	0.683491	0.011323	0.912412
10	0.671625	14.823639	<.0001*	0.610463	0.999991	0.015562	0.699053	0.010699	0.923112
11	0.695867	17.385858	<.0001*	0.581816	0.999996	0.019040	0.718092	0.007945	0.931057
12	0.720865	19.399541	<.0001*	0.551351	0.999998	0.021476	0.739568	0.007050	0.938107
13	0.763351	22.554517	<.0001*	0.496903	0.999999	0.011706	0.751274	0.008857	0.946963
14	0.791689	26.405173	<.0001*	0.458986	1.000000	0.011510	0.762784	0.006236	0.953199
15	0.815424	28.291824	<.0001*	0.426187	1.000000	0.011048	0.773832	0.005424	0.958623

PRESS 均值根图



注意: 最小 PRESS 均值根是 0.54774，且最低因子数是 5。

“PRESS 均值根图”显示当因子数为 5 时， PRESS 均值根的值最小。这在“PRESS 均值根图”下的注释中有说明。生成名为**带 5 个因子使用快速 SVD 的 “NIPALS”拟合**的报表。

“偏最小二乘”平台的统计详细信息

本节包含有关“偏最小二乘”平台中使用的一些方法的统计详细信息。请参见 Hoskuldsson (1988)、Garthwaite (1994) 或 Cox and Gaudard (2013)。

偏最小二乘的统计详细信息

偏最小二乘基于解释变量 (*X*) 的线性组合（称为因子）来拟合线性模型。这些因子是通过将 *X* 与一个或多个响应 (*Y*) 之间的协方差最大化得到的。这样，PLS 可利用 *X* 和 *Y* 之间的相关性来揭示底层的潜在结构。这些因子实现了解释响应变异和预测变量变异的组合目标。当您的 X 变量数比观测数多或 X 变量高度相关时，偏最小二乘特别有用。

NIPALS

NIPALS 方法一次提取一个因子。用 $\mathbf{X} = \mathbf{X}_0$ 表示预测变量的中心化和统一尺度的矩阵，用 $\mathbf{Y} = \mathbf{Y}_0$ 表示响应值的中心化和统一尺度的矩阵。PLS 方法从预测变量的一个线性组合 $\mathbf{t} = \mathbf{X}_0\mathbf{w}$ 开始，其中 **t** 称为**得分向量**， **w** 是相关的**权重向量**。 PLS 方法通过 **t** 的回归来预测 $\mathbf{X}_0$ 和 $\mathbf{Y}_0$：

$\hat{\mathbf{X}}_0 = \mathbf{tp}'$，其中 $\mathbf{p}' = (\mathbf{t}'\mathbf{t})^{-1}\mathbf{t}'\mathbf{X}_0$

$\hat{\mathbf{Y}}_0 = \mathbf{tc}$，其中 $\mathbf{c}' = (\mathbf{t}'\mathbf{t})^{-1}\mathbf{t}'\mathbf{Y}_0$

向量 **p** 和 **c** 分别称为 *X* 和 *Y* **载荷**。

特定线性组合 $\mathbf{t} = \mathbf{X}_0\mathbf{w}$ 是在指定某些响应线性组合 $\mathbf{u} = \mathbf{Y}_0\mathbf{q}$ 的情况下具有最大协方差 $\mathbf{t}'\mathbf{u}$ 的组合。另一特性是 *X*- 和 *Y*-权重 **w** 和 **q** 与协方差矩阵 $\mathbf{X}_0'\mathbf{Y}_0$ 的第一个左奇异向量和右奇异向量成比例。或者分别等效于 $\mathbf{X}_0'\mathbf{Y}_0\mathbf{Y}_0'\mathbf{X}_0$ 和 $\mathbf{Y}_0'\mathbf{X}_0\mathbf{X}_0'\mathbf{Y}_0$ 的第一个特征向量。

这说明了如何提取第一个 PLS 因子。通过将 $\mathbf{X}_0$ 和 $\mathbf{Y}_0$ 替换为第一个因子中的 *X* 和 *Y* 残差，以相同方式提取第二个因子：

$$\mathbf{X}_1 = \mathbf{X}_0 - \hat{\mathbf{X}}_0$$

$$\mathbf{Y}_1 = \mathbf{Y}_0 - \hat{\mathbf{Y}}_0$$

这些残差还被称为**缩小的** *X* 和 *Y* 区组。根据要提取的因子数，重复提取得分向量和缩小数据矩阵的过程。

SIMPLS

SIMPLS 算法开发的目的在于优化统计准则：它寻找在 *X* 得分是正交的前提下使 *X* 和 *Y* 的线性组合之间的协方差最大化的得分向量。不同于 NIPALS，该算法缩小了矩阵 $\mathbf{X}_0$ 和 $\mathbf{Y}_0$，SIMPLS 缩小的是叉积矩阵 $\mathbf{X}_0'\mathbf{Y}_0$。

在单个 *Y* 变量的情况下，这两个算法是等效的。但是，对于多元 *Y*，模型有所不同。SIMPLS 是 De Jong (1993) 提出的。

van der Voet T^2 检验的统计详细信息

在“偏最小二乘”平台中，van der Voet T^2 检验帮助确定具有指定提取因子数的模型是否与建议的最优模型显著不同。该检验是基于以下原假设的随机化检验：两个模型的残差平方具有相同的分布。直观上，可以将该原假设表述为：两个模型具有相同的预测能力。

要获得“交叉验证”报表中给出的 van der Voet T^2 统计量，对每个验证集执行下面的计算。在单个验证集的情况下，结果是报告的值。若使用“留一法”和“K 重”验证，则对每个验证集的结果求平均值。

用 $R_{i,jk}$ 表示模型的响应 *k* 的第 *j* 个预测残差，该模型具有 *i* 个提取因子。用 $R_{opt,jk}$ 表示模型的相应量，该模型基于建议的最佳因子数 *opt*。检验统计量基于以下差值：

$$D_{i,jk} = R^2_{i,jk} - R^2_{opt,jk}$$

假定有 *K* 个响应。考虑以下符号：

$$\mathbf{d}_{i,j} = (D_{i,j1}, D_{i,j2}, \ldots, D_{i,jK})'$$

$$\mathbf{d}_{i,.} = \sum_j \mathbf{d}_{i,j}$$

$$\mathbf{S_i} = \sum_j \mathbf{d}_{i,j}\mathbf{d}_{i,j}'$$

按以下方式定义 *i* 个提取因子的 van der Voet 统计量：

$$C_i = \mathbf{d}_{i,.}'\mathbf{S_i}^{-1}\mathbf{d}_{i,.}$$

通过比较 C_i 与将 $R^2_{i,jk}$ 和 $R^2_{opt,jk}$ 随机交换所得值的分布，来获得显著性水平。模拟了这类值的 Monte Carlo 样本，并将显著性水平近似为模拟临界值大于等于 C_i 的比例。

T^2 图的统计详细信息

在“偏最小二乘”平台中，第 i 个观测的 T^2 值计算如下：

$$T_i^2 = (n-1)\sum_{j=1}^{p}\left(t_{ij}^2 / \sum_{k=1}^{n} t_{kj}^2\right)$$

其中 t_{ij} = 第 i 行和第 j 个提取因子的 X 得分，p = 提取的因子数，n = 用于训练模型的观测数。若未使用验证，则 n = 总观测数。

按以下方式计算 T^2 图的控制限：

$((n\text{-}1)^2/n)$*BetaQuantile(0.95, p/2, $(n\text{-}p\text{-}1)/2$)

其中 p = 提取的因子数，n = 用于训练模型的观测数。若未使用验证，则 n = 总观测数。请参见 Tracy et al. (1992)。

X 得分散点图矩阵的置信椭圆的统计详细信息

在“偏最小二乘”平台中，“得分散点图矩阵”选项将 95% 置信椭圆添加到 X 得分散点图。X 得分是不相关的，因为 NIPALS 和 SIMPLE 算法都生成正交得分向量。该椭圆假定每对 X 得分服从具有零相关性的二元正态分布。

考虑垂直轴上得分 i 和水平轴上得分 j 的散点图。椭圆的上、下、左、右端点的坐标定义如下：

- 上、下端点为 +/-sqrt(var(score i)*z)
- 左、右端点为 +/-sqrt(var(score j)*z)

其中 z = $((n\text{-}1)*(n\text{-}1)/n)$*BetaQuantile(0.95, 1, $(n\text{-}3)/2$)。有关 z 值的背景信息，请参见 Tracy et al. (1992)。

预测和置信限的统计详细信息

本节说明如何在“偏最小二乘”平台中计算预测标准误差和置信限。用 **X** 表示预测变量的矩阵，用 **Y** 表示响应值的矩阵，该矩阵可能基于您在启动窗口中的选择进行了中心化和统一尺度。假定 **Y** 的成分是独立的，且服从具有公共方差 σ^2 的正态分布。

Hoskuldsson (1988) 注意到基于得分构建的 **Y** 的 PLS 模型在形式上类似于多元线性回归模型。他使用这个相似性推导出预测值方差的近似公式。另见 Umetrics (1995)。但是，Denham (1997) 指出 PLS 预测的任何值是 Y 的非线性函数。他提出采用 Bootstrap 和交叉验证方法来获取预测区间。PLS 平台使用 Umetrics (1995) 中所述的基于正态性的方法。

用 **T** 表示其列为得分的矩阵，考虑 **X** 的新观测 $\mathbf{x}_0$。通过 **T** 对 **Y** 做回归来得到 **Y** 的预测模型。用 $\mathbf{t}_0$ 表示与 $\mathbf{x}_0$ 关联的得分向量。

用 a 表示因子数。将 s^2 定义为残差平方和除以自由度，若数据中心化，则自由度 $df = n - a - 1$，若数据未中心化，则自由度 $df = n - a$。s^2 的值是 σ^2 的估计值。

预测公式的标准误差

按以下方式估计 $\mathbf{x}_0$ 处的预测均值的标准误差：

$$SE(\overline{Y}_{x_0}) = s\sqrt{\left(\frac{1}{n} + \mathbf{t}_0(\mathbf{T}'\mathbf{T})^{-1}\mathbf{t}_0'\right)}$$

均值置信限公式

用 $t_{0.975,\,df}$ 表示 t 分布的 0.975 分位数，若数据已中心化，则该 t 分布的自由度 $df = n - a - 1$，若数据未中心化，则自由度 $df = n - a$。

按以下方式计算均值的 95% 置信区间：

$$\overline{Y}_{x_0} \pm t_{0.975,\,df} SE(\overline{Y}_{x_0})$$

单值置信限公式

按以下方式估计 $\mathbf{x}_0$ 处的单个响应预测值的标准误差：

$$SE(\hat{Y}_{x_0}) = s\sqrt{\left(\frac{1}{n} + 1 + \mathbf{t}_0(\mathbf{T}'\mathbf{T})^{-1}\mathbf{t}_0'\right)}$$

用 $t_{0.975,\,df}$ 表示 t 分布的 0.975 分位数，若数据已中心化，则该 t 分布的自由度 $df = n - a - 1$，若数据未中心化，则自由度 $df = n - a$。

按以下方式计算单个响应的 95% 预测区间：

$$\overline{Y}_{x_0} \pm t_{0.975,\,df} SE(\hat{Y}_{x_0})$$

标准化得分和载荷的统计详细信息

本节说明如何在“偏最小二乘”平台中计算标准化得分和载荷。

考虑以下符号：

- n_{tr} 是训练集中的观测数
- m 是 X 中的效应数
- k 是 Y 中的响应数
- $VarX_i$ 是第 i 个因子解释的 X 的变异百分比
- $VarY_i$ 是第 i 个因子解释的 Y 的变异百分比

- $\mathbf{XScore}_i$ 是第 i 个因子的 X 得分的向量
- $\mathbf{YScore}_i$ 是第 i 个因子的 Y 得分的向量
- $\mathbf{XLoad}_i$ 是第 i 个因子的 X 载荷的向量
- $\mathbf{YLoad}_i$ 是第 i 个因子的 Y 载荷的向量

标准化得分

按以下方式定义第 i 个标准化 X 得分的向量：

$$\frac{\mathbf{XScore}_i}{(n_{tr}-1)\sqrt{mVarX_i/n_{tr}}}$$

按以下方式定义第 i 个标准化 Y 得分的向量：

$$\frac{\mathbf{YScore}_i}{(n_{tr}-1)\sqrt{kVarY_i/n_{tr}}}$$

标准化载荷

按以下方式定义第 i 个标准化 X 载荷的向量：

$$\mathbf{XLoad}_i\sqrt{mVarX_i}$$

按以下方式定义第 i 个标准化 Y 载荷的向量：

$$\mathbf{YLoad}_i\sqrt{kVarY_i}$$

JMP PRO PLS 判别分析的统计详细信息

您可以通过使用“拟合模型”平台中的“偏最小二乘”特质执行“偏最小二乘判别分析”(PLS-DA)。在启动窗口中将某个分类变量输入为 Y 时，使用指示符编码对其编码。若有 k 个水平，则用一个指示符变量来表示每个水平，对于属于该水平的行用值 1 表示，不属于该水平的行用 0 表示。得到的 k 个指示符变量被视为连续变量，PLS 分析按处理连续 Y 的方式处理这些指示符变量。

第 7 章

多重对应分析

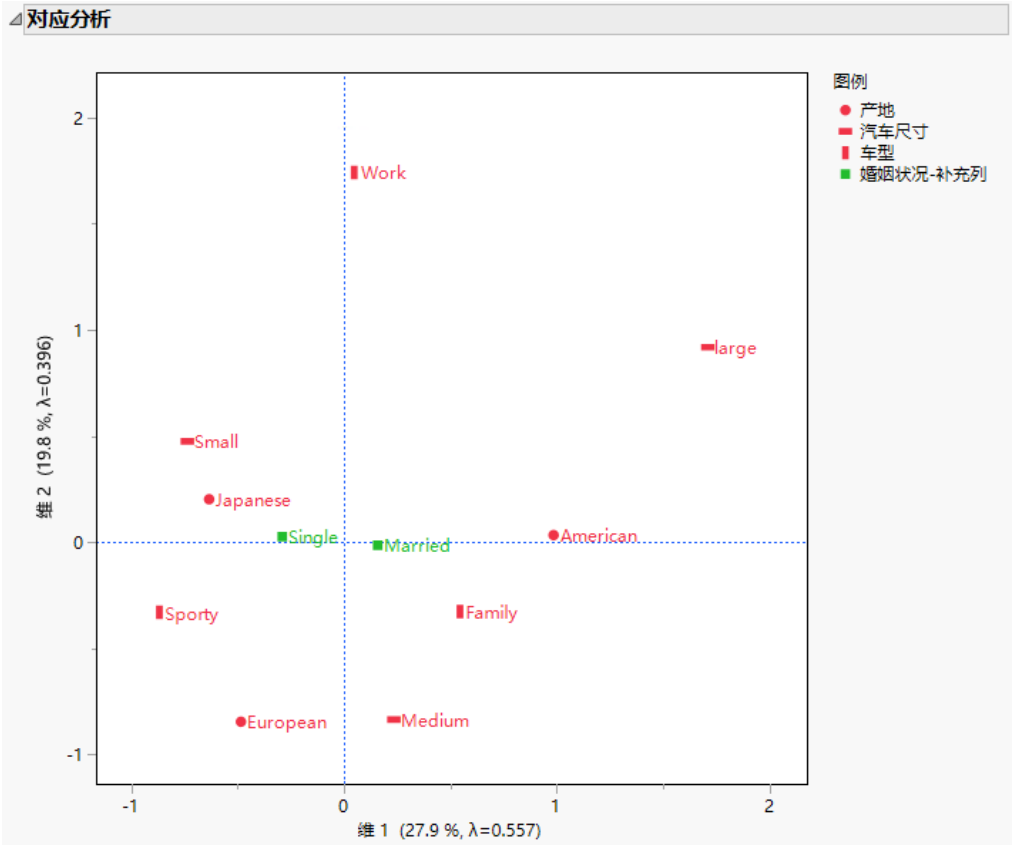
识别分类变量各水平之间的关联

“多重对应分析”(MCA) 针对多个分类变量，试图查找这些变量各水平之间的关联。MCA 将对应分析从两个变量扩展到多个变量的情况。您可以认为它就如同针对定量变量的主成分分析。与其他多元方法类似，多重对应分析它是一种降维方法，将原始的多维数据表示为二维或三维空间中的点。

社会科学领域经常应用多重对应分析。可将其用在调查分析中，找出测试对象对不同问题的态度一致性。还可以在消费者研究中使用该方法来确定产品的潜在市场。

有关多重对应分析的详细信息，请参见 LeRoux and Rouanet (2010)。

图 7.1 多重对应分析



目录

多重对应分析的示例

在本例中，您将探讨员工对四个问题的响应之间的关联，以了解不同的员工群体。本例使用从 55 名 JMP 员工收集的关于他们在以下领域的偏好（或品味）的数据：

- 偏好的电视节目（8 类）：新闻、喜剧、警察、自然、体育、电影、戏剧或肥皂剧。
- 偏好的影片（8 类）：动作片、喜剧、古装剧、纪录片、恐怖片、音乐剧、言情片或科幻片。
- 偏好的艺术类型（7 类）：表演、风景、文艺复兴、静物、肖像、现代或印象。
- 偏好的外出就餐场所（判别主成分 类）：汉堡和薯条、酒吧、印度餐厅、意大利餐厅、法国餐厅或牛排馆。

注意：偏好的测度基于 LeRoux and Rouanet (2010) 中的问题。

1. 选择**帮助** > **样本数据文件夹**，然后打开 Employee Taste.jmp。
2. 选择**分析** > **多元方法** > **多重对应分析**。
3. 选择**电视节目**、**电影**、**艺术**和**餐厅**，然后点击 **Y，响应**。

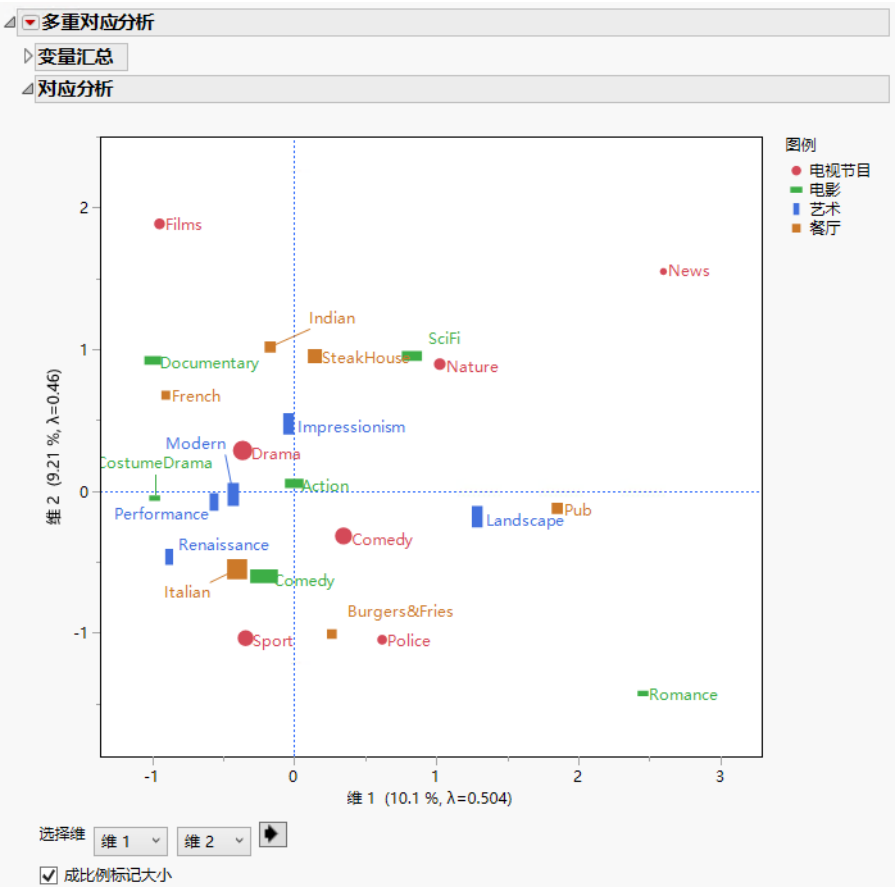
   在 MCA 中，与主成分分析一样，因子通常被认为是响应，而不是一些是响应，另一些是解释变量。
4. 点击**确定**。

提示：点击并拖动重叠标签将其重新排列。

5. 在该图下方，选择**成比例标记大小**。

   标记大小指示该类别中响应的相对比例。

图 7.2 最初的 “多重对应分析” 报表的一部分



“对应分析” 报表显示了四个变量的类别在前两个主轴或维上的投影。使用控件更改标绘的维。点与点之间的距离代表了员工响应模式之间的差异。

从该图中，您可以用您对主题的了解来解释以下发现：

- 聚集在 “Burgers and Fries” 餐厅偏好附近，显示有 “Sport” 和 “Police” 电视节目偏好的聚类。该聚类可称为 “大众文化” 群体。
- 从 “维 2” 来看，个人品味从大众文化群体转向可以归类为更 “复杂” 的品味 — 那些喜欢纪录片、偏好电视剧和特色餐厅（如法国菜和印度菜）的人群。

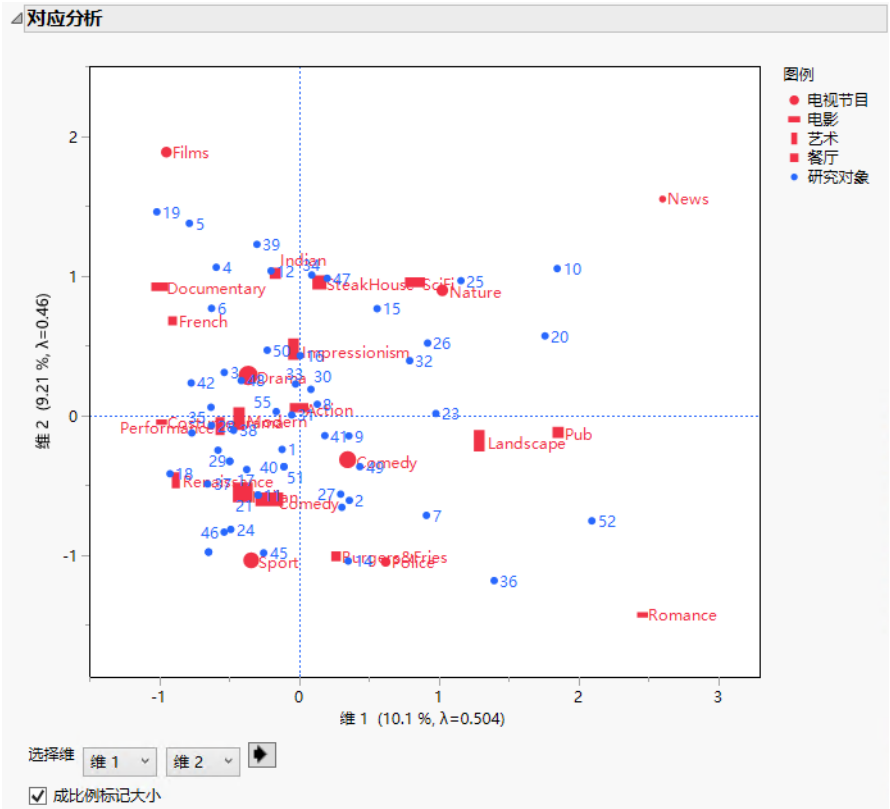
要获得个人得分，请将对象添加为 **X，因子**：

6. 打开 “变量汇总” 分级显示项。

   “变量汇总” 面板允许您修改分析而不必重新启动该平台。它还提供已完成分析的简明视图。

7. 选择**研究对象**并点击**添加 X**。分析自动更新。

图 7.3 包含研究对象的 “多重对应分析” 报表

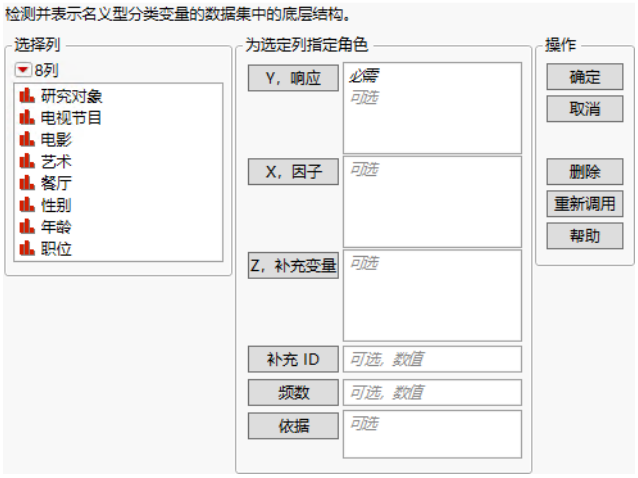


包括研究对象后，则可在图中左下象限突出显示有着相似品味的员工的聚类。这些员工属于我们认定的大众文化领域。

启动“多重对应分析”平台

通过选择**分析** > **多元方法** > **多重对应分析**启动“多重对应分析”平台。

图 7.4 “多重对应分析”启动窗口



有关“选择列”红色小三角菜单中选项的详细信息，请参见《使用 JMP》。

Y，响应 指定要分析的分类列。在 MCA 中，您通常关注变量之间的关联，但是没有明确的“解释”和“响应”变量。

X，因子 分配一个或多个要用作因子或解释变量的分类列。

注意：将某个对象 ID 列用作单个 X 来获取个人得分。

Z，补充变量 指定要用作补充变量的列。您对这些变量和响应的关系感兴趣，但进行计算时并不把它们包括在内。补充变量用于改进数据解释。

补充 ID 指定列用于标识不包含在分析计算中的行。“补充 ID”列通常具有值 1 或 0。与 ID 0 关联的行被视为补充行。若补充行中的 X 或 Y 变量水平在非补充行中并未出现，则忽略“补充 ID”列。

注意：一次只能指定“补充 ID”或“Z，补充变量”角色中的一种。

频数 指定一个频数变量。该选项适用于汇总数据。

依据 为“依据”变量的每个水平生成单独报表。若指定了多个“依据”变量，将为“依据”变量水平的每种可能组合生成单独的报表。

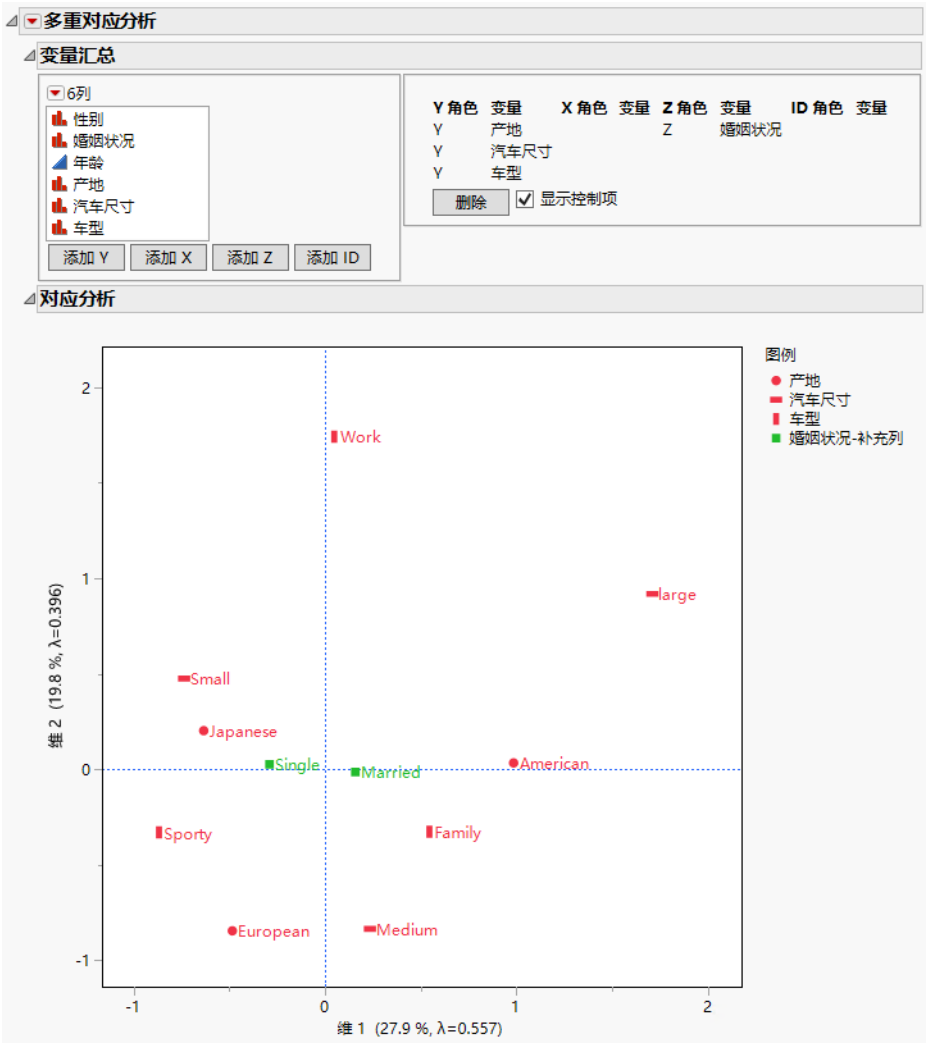
注意：“多重对应分析”平台处理缺失值的方式不同于其他许多 JMP 平台。该分析使用行中的所有非缺失单元格对。它不会从计算中删除整个行。

“多重对应分析”报表

初始 MCA 报表显示变量汇总、对应分析图和按重要性排列的数据维度的详细信息。从类别或个体的对应图中，您可以发现数据中存在的关联。这些细节提供了图中两个维是否足够解释数据内关系的信息。

交互式面板中的“变量汇总”允许您修改分析而不必重新启动该平台。该面板显示分析中使用的列和角色。若您选中“显示控制项”复选框，包含数据表中各列的列表将显示在左侧。您可以通过选择一列并点击“添加 Y”、“添加 X”、“添加 Z”或“添加 ID”来更改分析中的列。或者，您可以将该列拖到变量汇总表中的标题处。如此您不必返回到启动窗口就可以修改分析设置。

图 7.5 选中了 “显示控制项”的 “多重对应分析”报表



“多重对应分析”平台选项

MCA 红色小三角菜单包含的选项支持您根据需要定制报表。

交叉表 为选定的变量角色适当显示或隐藏 Burt 表或列联表。请参见 “交叉表”。

对应分析 显示对应分析报表的子菜单。根据您指定的分析类型，有以下报表可供您选择。这些报表给出图、详细信息、坐标和汇总统计量。

显示图 显示或隐藏将类别数据投影到提取的前两个主轴上的二维对应分析表。默认显示该图。该图使用等距尺度。请参见 “显示图”。

显示详细信息 显示或隐藏分析的详细信息，包括奇异值、惯量、卡方统计量、百分比和累积百分比。默认显示该报表。请参见“显示详细信息”。

显示调整惯量 显示或隐藏 Benzecri 和 Greenacre 调整惯量的报表。请参见 Benzécri (1979) 和 Greenacre (1984)。该选项在有一个或多个 X 变量时不可用。请参见“显示调整惯量”。

显示坐标 显示或隐藏一个报表，针对每个类别最多包含对应分析中前三个维度的坐标值，具体视情况而定。请参见“显示坐标”。

显示汇总统计量 显示或隐藏一个报表，它显示分析中每个类别对应的汇总统计量、质量、量和惯量。请参见“显示汇总统计量”。

显示对惯量的部分贡献 显示或隐藏一个报表，它显示每个类别对最多前三个维惯量的贡献。请参见“显示对惯量的部分贡献”。

显示平方余弦 显示或隐藏一个报表，针对每个类别，显示最多前三个维的平方余弦。该报表包括一个条形图，其中为每个 Y 变量的每个水平显示最多前三个维中每个维的平方余弦值。请参见“显示平方余弦”。

Cochran Q 检验 （仅在所有 Y 变量都仅包含相同的一组两水平并且 X 变量每行都具有唯一值的情况下，该选项才可用。）提供 Cochran *Q* 统计量，该统计量检验特定响应的边缘概率在各个 Y 变量中是否不变。Cochran *Q* 统计量是针对两个以上响应变量的广义的 McNemar 统计量。请参见 Agresti (2013)。请参见“Cochran Q 检验”。

三维对应分析 显示或隐藏将空间中 Y、X 和 Z 变量的类别数据投影到前三个主轴上的三维对应分析图。若少于三个维度，则该选项不可用。

保存坐标 将主坐标保存到一个或多个 JMP 数据表。列坐标、行坐标、补充列坐标和补充行坐标保存到单独的 JMP 数据表。您可以选择要保存的列数。

保存坐标公式 将多个维度的主坐标的公式列保存到数据表中。每个观测的值是以每个维的奇异值统一尺度的 Y 变量坐标的平均值。您可以选择要保存的列数。

补充行的交叉表 （在指定一个或多个补充 Z 和 X 变量或在指定 ID 时可用。）依据以下规则显示或隐藏 Burt 或列联表：

- 当指定一个或多个 X 和 Z 变量时，显示或隐藏“响应变量 - X 变量”列联表。
- 若指定了 ID 但未指定 X 变量，则显示或隐藏补充观测的 Burt 表。补充观测是 ID 变量等于 0 的那些观测。
- 若指定了 ID 并且指定了一个或多个 X 变量，则显示或隐藏补充观测的列联表。补充观测是 ID 变量等于 0 的那些观测。

补充列的交叉表 （当指定一个或多个补充 Z 变量时可用。）依据以下规则显示或隐藏列联表：

- 在未指定任何 X 变量时，显示或隐藏“补充变量 - 响应变量”列联表。
- 当指定一个或多个 X 变量时，显示或隐藏“补充变量 - X 变量”列联表。

马赛克图 （仅当有一个 X 变量和一个 Y 变量时才可用。）显示或隐藏 X - Y 马赛克图。马赛克图是堆叠的条形图，其中每一段与该组的频数计数成比例。

独立性检验 （仅当有一个 X 变量和一个 Y 变量时才可用。）显示或隐藏 X 和 Y 变量之间的独立性检验。该检验有两种方法：Pearson 形式和似然比形式，两者都计算卡方统计量。

请参见《使用 JMP》获取有关下列选项的信息：

本地数据过滤器 显示或隐藏支持您过滤特定报表中使用的数据的本地数据过滤器。

重新运行 包含使您可以重复或重新启动分析的选项。在支持该功能的平台中，“自动重新计算”选项立即在相应报表窗口中反映您对数据表所做的更改。

平台首选项 包含的选项支持您查看当前平台首选项或更新平台首选项以匹配当前 JMP 报表中的设置。

保存脚本 包含的选项支持您保存可将报表重现到若干目标的脚本。

保存 “依据”组脚本 包含使您可以保存脚本的选项，可将为“依据”变量的所有水平重新生成平台报表的脚本保存到多个不同的位置。仅当在启动窗口中指定“依据”变量时才可用。

显示图

在“多重对应分析”平台中，该选项显示类别或个体在前两个主轴所构成的平面上的投影中的一个投影图。该图使用等距尺度。您可以使用图下方的“选择维”控件来切换图中显示的维。第一个控件定义图的水平轴，第二个控件定义图的垂直轴。点击箭头按钮可以在图中显示的维之间循环切换。使用**成比例标记大小**复选框指定图中各点的大小是否应与对应每个点的观测计数成比例。

注意：选择对应分析图中的某个点也会选择报表窗口内其他表中的对应行，但并不选择该数据表中的行。要选择图中与特定变量关联的所有点，请在图例中选择变量的名称。

显示详细信息

在“多重对应分析”平台中，该选项显示奇异值表。

奇异值 显示列联表或 Burt 表的奇异值分解中的奇异值。有关该公式，请参见 “‘详细信息’报表的统计详细信息”。

惯量 列出奇异值的平方，反映了典型维中解释的相对变异。

卡方 列出针对 Burt 表或列联表计算的总卡方值拆分到当前维的部分。

百分比 各维惯量占总惯量的比例。

累积百分比 显示惯量的累积比例。若前两个奇异值捕获绝大部分惯量，则二维对应分析图足以显示表中的关系。

显示调整惯量

“多重对应分析”中的 Burt 表的主惯量为特征值。这些惯量存在的问题是对拟合好坏的判定过于悲观。Benzécri 提出了一种惯量调整方法，而 Greenacre 认为 Benzécri 调整过高估计了拟合

质量，因此他提出了新的调整方法来替代。我们计算了两种调整值供您参考。请参见 “调整惯量的统计详细信息”。

惯量 列出奇异值的平方，反映了典型维中解释的相对变异。

调整惯量 列出了根据 Benzécri 或 Greenacre 调整的调整惯量。

百分比 调整的惯量占总惯量的比例。

累积百分比 显示调整惯量的累积比例。若前两个奇异值捕获绝大部分惯量，则二维对应分析图足以显示表中的关系。

显示坐标

在 “多重对应分析”平台中，该选项显示 “列坐标”表，或 “行坐标”表和 “列坐标”表。

X 列出指定为 “X，因子”变量的列。

Y 列出指定为 “Y，响应”变量的列。

Z 列出指定为 “Z，补充变量”的列。

类别 列出 X、 Y 或 Z 变量的水平。

维 1、维 2、维 3 对于每个水平或每个响应，在相应的主轴上列出其坐标。默认情况下，表中显示最多前三个维的坐标。隐藏其他维的坐标列。要显示这些可选列，请右击某个表并从**列**子菜单中选择维列。

注意：若有指定为 “X，因子”变量的列，则 “坐标”报表会显示 X 表和 Y 表在同一个标题下。若指定了补充变量 Z，则会在 X 和 Y 坐标下列出相应的坐标（若适用）。

显示汇总统计量

在 “多重对应分析”平台中，该选项显示 “列点的汇总统计量”表，或 “行和列点的汇总统计量”表。 Y 表提供每个响应的每个水平的质量、量和惯量，这些统称为**列点**。 X 表为 “X，因子”变量的每个水平提供质量、量和惯量。请参见 “汇总统计量的统计详细信息”。

X 列出指定为 “X，因子”变量的列。

Y 列出指定为 “Y，响应”变量的列。

类别 列出 X 和 Y 变量的水平。

质量 （维 =2） 列出用解表示水平的质量。

量 列出响应水平的行频数除以总频数的结果。在 Burt 表中，等同于每行的 “合计百分比”。

惯量 列出响应水平所占总惯量的比例。各个水平及其响应的惯量值之和为 1。

注意：若有指定为“X，因子”变量的列，则“汇总统计量”报表会将 X 变量和 Y 变量相关的表显示在同一报表标题下。

显示对惯量的部分贡献

在“多重对应分析”平台中，该选项显示“关于列点对惯量的部分贡献”表，或“关于行和列点对惯量的部分贡献”表。还显示“关于列点对惯量的部分贡献图”。这是一个条形图，其中针对每个 Y 变量的每个水平，显示其对表中显示的每个维的部分贡献。

X 列出指定为“X，因子”变量的列。

Y 列出指定为“Y，响应”变量的列。

类别 列出 X 和 Y 变量的水平。

维 1、维 2、维 3 列出响应或因子水平对所指维的惯量的贡献。默认情况下，表中显示最多前三个维的列。其他列会隐藏起来。要显示这些可选列，请右击某个表并从**列**子菜单中选择维列。

每个响应的每个水平都对每个维的惯量做贡献。每个维中的部分贡献加总为 1。

行或列对维的惯量的贡献计算如下：

贡献 = (量)(坐标)2/(维惯量)

注意：若有指定为“X，因子”变量的列，则“对惯量的部分贡献”报表会显示 X 表和 Y 表在同一个标题下。

显示平方余弦

在“多重对应分析”平台中，该选项显示“列点的平方余弦”表，或“行和列点的平方余弦”。还显示“列点的平方余弦图”。这是一个条形图，其中针对每个 Y 变量的每个水平，显示最多前三个所示维中每个维的平方余弦值。

X 列出指定为“X，因子”变量的列。

Y 列出指定为“Y，响应”变量的列。

类别 列出 X 和 Y 变量的水平。

维 1、维 2、维 3 列出用所指维表示水平的**质量**。默认情况下，表中显示最多前三个维的结果。其他列会隐藏起来。要显示这些可选列，请右击某个表并从**列**子菜单中选择维列。

这些值表示每个列点在相应维上的质量。平方余弦可以解释为点与维的相关性的平方。前两个维的平方余弦之和等于“汇总统计量”报表中的“质量（维 =2）”。请参见“汇总统计量的统计详细信息”。

注意：若有指定为“X，因子”变量的列，则“平方余弦”报表会将 X 变量和 Y 变量相关的表显示在同一报表标题下。

Cochran Q 检验

在“多重对应分析”平台中，Cochran *Q* 检验是三个或更多二值结果的匹配样本的非参数齐性检验。您可以用它来检验配对中的比例差异。Cochran *Q* 检验是用于两种结果的 McNemar 检验的扩展。

交叉表

在“多重对应分析”平台中，“交叉表”选项显示或隐藏 Burt 表或列联表。当您选择多个“Y，响应”列且没有选择任何“X，因子”列时，将创建 Burt 表。若您选择任何“X，因子”列，则创建传统的列联表而非 Burt 表。分级显示项节点标题反映了交叉表的结构。

Burt 表是多重对应分析的基础。它是描述所有分类变量两两交叉列联的分区对称表。对角分区是对角矩阵（变量与其自身的交叉表）。非对角分区是普通的列联表。

Burt 表或列联表的红色小三角菜单包含要在表中显示的统计量的以下选项。

计数 单元格频数、边缘总频数和总计 （总样本大小）。默认显示该选项。

合计百分比 单元格计数和边缘合计占总计的百分比。默认显示该选项。

单元格卡方值 为每个单元格计算的卡方值，公式为 $(O - E)^2 / E$。

列百分比 每个单元格计数占列合计的百分比。

行百分比 每个单元格计数占行合计的百分比。

期望值 在独立性假设下每个单元格的期望频数 (E)。它由相应行合计与列合计之积除以总计得到。

偏差 观测的单元格频数 (O) 减去期望的单元格频数 (E) 所得的值。

列累积 累积列合计。

列累积百分比 累积列百分比。

行累积 累积行合计。

行累积百分比 累积行百分比。

制成数据表 为表中显示的每个统计量创建一个数据表。

多重对应分析的更多示例

本节包含使用“多重对应分析”平台的示例。

- “使用补充变量的示例”
- “使用补充 ID 的示例”
- “Cochran Q 检验示例”

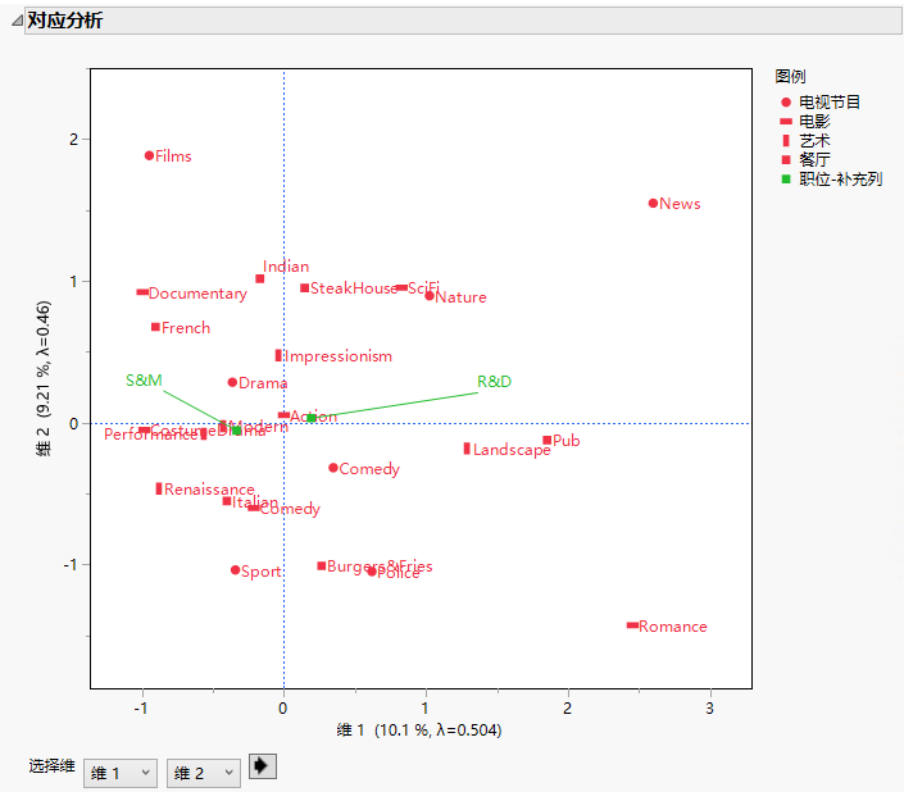
使用补充变量的示例

在“多重对应分析”平台中，使用一个补充变量来探索员工的工作类型与他们的偏好之间的关系。

1. 选择**帮助** > **样本数据文件夹**，然后打开 Employee Taste.jmp。
2. 选择**分析** > **多元方法** > **多重对应分析**。
3. 选择**电视节目**、**电影**、**艺术**和**餐厅**，然后点击 **Y，响应**。
4. 选择**职位**，然后点击 **Z，补充变量**。
5. 点击**确定**。

注意：由于**职位**是补充变量，所以在计算中不使用。计算完成后会标绘工作类型。

图 7.6 具有补充变量的 MCA 报表



从图中可以看出，R&D（研究和开发）和 S&M（销售和营销）员工所对应的数据点非常接近。我们认为这意味着没有由于员工工作职能的原因，其在偏好方面有很大的差异。换句话说，员工的工作角色并不是影响其偏好的显著因子。

使用补充 ID 的示例

使用 “多重对应分析”来确定阿拉斯加州和夏威夷州的人口增长是否与美国其他地区不同。这两个州被视为补充地区，因为在整个数据收集期间它们还没建州，而且不与美国大陆接壤。

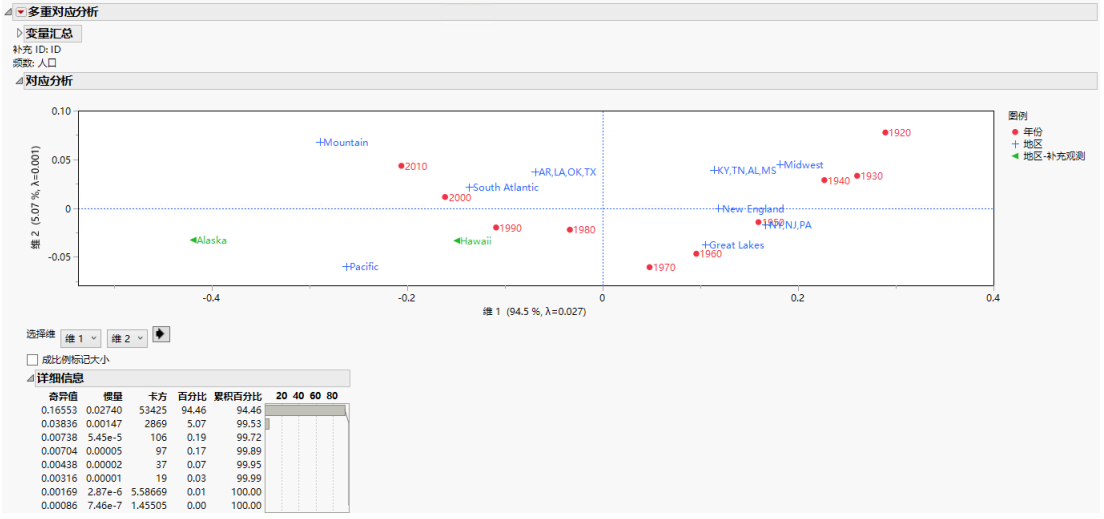
1. 选择**帮助** > **样本数据文件夹**，然后打开 US Regional Population.jmp。
2. 选择**分析** > **多元方法** > **多重对应分析**。
3. 选择**年份**，然后点击 **Y，响应**。
4. 选择**地区**，然后点击 **X，因子**。
5. 选择 ID，然后点击**补充 ID**。
6. 选择**人口**，然后点击**频数**。
7. 点击**确定**。

“详细信息”报表显示年份和地区之间的关联几乎可以完全使用第一个维来解释。从该图可以看出第一维度上的年份顺序符合自然规律，该排序在整个对应分析过程中自然地进行，事先没有提供排序信息。该图突出显示用于标绘数据的等距尺度。

从地区的排序可以看出人口的迁移规律是从中西部到东北部、再到南部、最后到山地和西部。

上述对应分析的计算没有包含阿拉斯加州和夏威夷州的数据，但我们可以根据结果来标绘它们。它们的增长模式类似于太平洋地区各州的模式，阿拉斯加州的增长率比太平洋地区更为极端。

图 7.7 具有补充 ID 的 MCA 报表

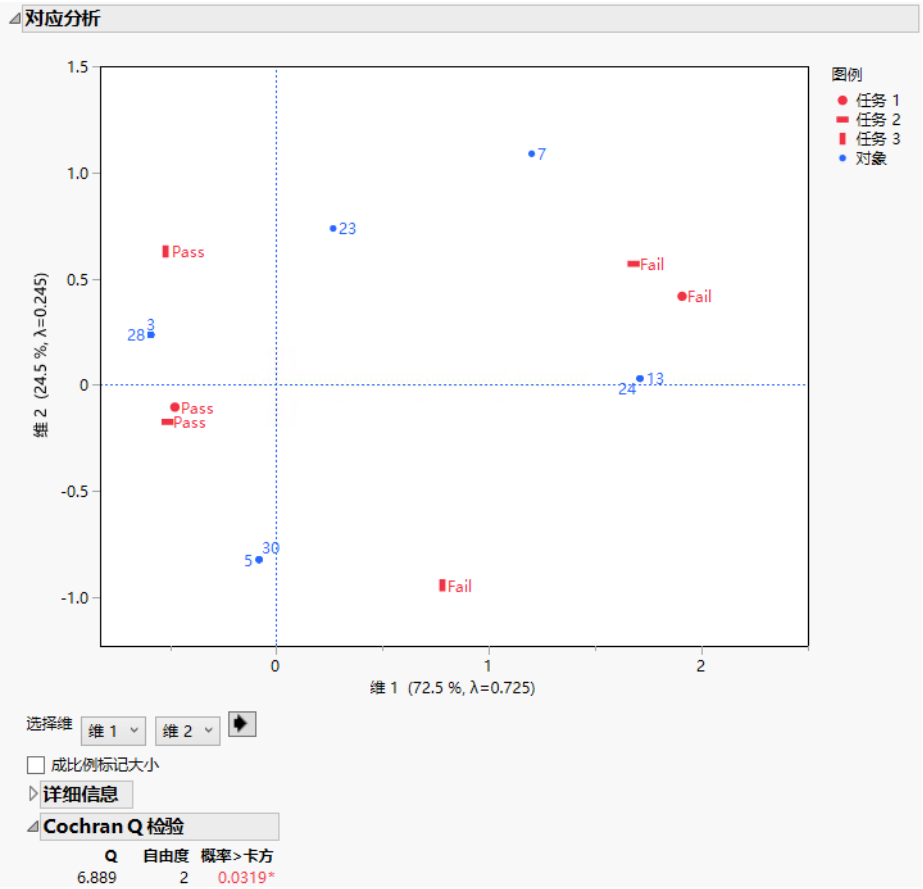


Cochran Q 检验示例

使用“多重对应分析”评估 30 个对象完成三项任务的容易程度。模拟任务 1 具有 80% 的完成率。模拟任务 2 和 3 具有与任务 1 相同的结果，但分别针对 95% 和 80% 的对象。

1. 选择**帮助** > **样本数据文件夹**，然后打开 Cohcrans Q.jmp。
2. 选择**分析** > **多元方法** > **多重对应分析**。
3. 选择**任务 1**、**任务 2** 和**任务 3**，然后点击 **Y，响应**。
4. 选择**对象**并点击 **X，因子**。
5. 点击**确定**。
6. 点击“多重对应分析”红色小三角菜单并选择**对应分析** > **Cochran Q 检验**。

图 7.8 Cochran Q 检验



该对应图显示了在将“Pass”和“Fail”结果聚在一起显示时任务 1 与任务 2 之间的相似性。不过，任务 3 更加远离任务 1 和 2。Cochran *Q* 检验统计量 判别主成分.889 关联的 *p* 值为 0.0319。该 *p* 值支持拒绝在 0.05 显著性水平下所有任务的通过率相同的假设。

“多重对应分析”平台的统计详细信息

本节包含“多重对应分析”平台的统计详细信息。

- “详细信息”报表的统计详细信息”
- “调整惯量的统计详细信息”
- “汇总统计量的统计详细信息”
- “Cochran Q 统计量的统计详细信息”

“详细信息”报表的统计详细信息

在“多重对应分析”平台中，“详细信息”报表列出奇异值。用于获得这些值的奇异值分解是简单“对应分析”中所用奇异值分解的扩展。

执行简单的“对应分析”时，报表列出的奇异值来自以下奇异值分解：

$$\mathbf{D}_r^{-0.5}(\mathbf{P} - rc')\mathbf{D}_c^{-0.5} = \mathbf{U}Diag(\Lambda)\mathbf{V}'$$

其中：

- **P** 是计数除以总频数所得的值构成的矩阵
- *r* 和 *c* 是 **P** 的行总和与列总和
- **D** 矩阵是 *r* 和 *c* 的值构成的对角矩阵
- Λ 是奇异值的列向量

执行 “多重对应分析”时，奇异值分解扩展为以下等式：

$$\mathbf{D}^{-0.5}(\mathbf{C} - dd')\mathbf{D}^{-0.5} = \mathbf{U}Diag(\Lambda)\mathbf{V}'$$

其中：

- **C** 是 Burt 表
- *d* 是 **C** 的列总和的列向量（*d* 还是行总和，因为 **C** 是对称的）
- **D** 是 *d* 的值的对角矩阵

在“详细信息”报表中，惯量是列向量 Λ。奇异值是惯量向量的平方根。列坐标计算如下：

$$\mathbf{D}^{-0.5}\mathbf{V}Diag\left(\Lambda^{0.5}\right)$$

调整惯量的统计详细信息

在 “多重对应分析”中，从 m 个分类变量构造的 Burt 表的常见主惯量为来自 Λ^2 的特征值 λ_k。这些惯量对于拟合好坏的判定过于悲观。Benzécri (1979) 提出了以下惯量调整方法； Greenacre (1984, p. 145) 也对该调整方法进行过描述。

$$\left(\frac{m}{m-1}\right)^2 \times \left(\lambda_k - \frac{1}{m}\right)^2 \text{ 对于 } \lambda_k > \frac{1}{m}$$

针对所有大于 $1/m$ 的惯量，该调整计算调整惯量占调整惯量之和的百分比。

Greenacre (1984, p. 15 判别主成分) 认为 Benzécri 调整过高估计了拟合的质量。Greenacre 建议计算调整惯量与以下值的百分比：

$$\frac{m}{m-1}\left(\text{tr}\left(Diag\left(\Lambda^4\right)\right) - \frac{n_c - m}{m^2}\right)$$

对于大于 $1/m$ 的所有惯量，其中 $\text{tr}(Diag(\Lambda^4))$ 是惯量平方和，n_c 是 m 个变量上的总类别数。

汇总统计量的统计详细信息

本节介绍“多重对应分析”平台中报告的汇总统计量。

质量是某点距离指定维数所定义空间中的原点的平方距离与具有最大维数的空间中的原点的距离之比。对于卡方这个量度，给定维中某点的质量可通过其向量与定义该维的向量所构成的余弦得到。质量还等于指定维中的惯量总和与所有维中的惯量总和的比值。质量表示低维空间中在多大程度上表示了该点的信息。

量是行或列的总频数与总频数之比。

惯量类似于主成分分析中的方差。整体惯量是二维频数表的总 Pearson 卡方除以表中所有观测值总和得到的值。

相对惯量是点对总惯量的贡献比例。在汇总统计量表中，相对惯量列在标记为“惯量”的列中。

Cochran Q 统计量的统计详细信息

“多重对应分析”平台提供 Cochran Q 检验。 Cochran Q 检验统计量计算如下：

$$Q = k(k-1)\frac{\sum_{j=1}^{k}\left(\sum_{i=1}^{b} X_{ij} - \frac{N}{k}\right)^2}{\sum_{i=1}^{b}\left(\sum_{j=1}^{k} X_{ij}\left(k - \sum_{j=1}^{k} X_{ij}\right)\right)}$$

其中

K 是处理数

b 是区组数

X_{ij} 是第 k 个处理的第 i 个对象的响应 （0 或 1）

N 是所有对象和处理的正值响应的总计

第 8 章

JMP PRO 结构化方程模型

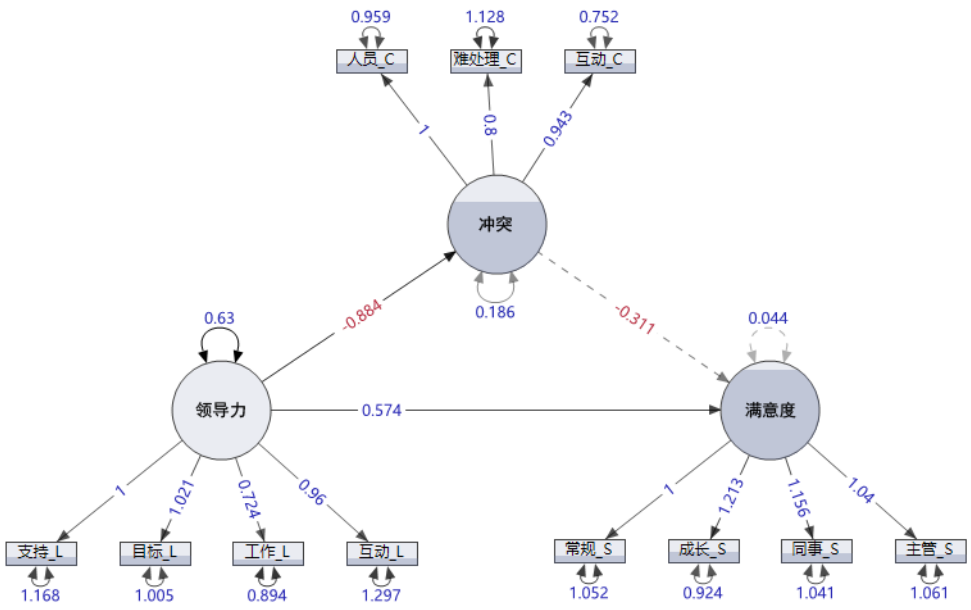
拟合结构化方程模型

“结构化方程模型”平台仅在 JMP-Pro 中提供。

“结构化方程模型”平台支持您拟合多种模型。这些模型包括确认性因子分析、具有或不具有潜在变量的路径模型、测量值误差模型以及潜在变量增长曲线模型。您可以通过选择**自**和**至**变量来指定模型并指明它们如何关联。在示意图中，这种关联通过单向箭头或双向箭头表示。这些选择填充了模型的两个备选视图：动态路径图和模型边缘列表。该平台还会在您指定模型时检查模型身份。

您还可以拟合多个组分析模型，从而可以对不同组的统计效应进行系统的模型比较。这对于检查不同总体之间的差异很有用。

图 8.1 结构化方程模型路径图



目录

JMP PRO 结构化方程模型概述

“结构化方程模型”(SEM) 平台支持您拟合可用于检验变量间关系理论的众多模型。模型中的变量可以是观测到的（**显**变量），也可以是未观测到的（**潜在**变量）。结构化方程建模在社会科学和行为科学中很普遍。

默认情况下，该平台为所有变量指定一个具有均值和方差的模型。然后，该平台提供了一个模型构建界面，使您能够在构建模型时看到模型的多个视图。它还提供模型构造过程中的一些模型细节，这些细节在运行模型之前会向您提醒哪些模型不受支持。

拟合一个或多个模型后，可以在“模型比较”报表中比较拟合的模型和两个基线模型。基线模型是不受限模型和独立模型。不受限模型是完全饱和模型，该模型拟合指定模型变量的所有均值、方差和协方差，而不对数据强加任何结构。独立模型拟合指定的模型变量的所有均值和方差。指定模型变量之间的所有协方差都被固定为零，这会导致高度受限的模型。

SEM 平台使用全信息最大似然 (Finkbeiner 1979) 方法。这使您能够充分利用数据中的所有可用信息，即使在具有随机缺失值的观测比例较高时也是如此。

有关结构化方程建模的详细信息，请参见 SAS Institute Inc.(2020a) 中的“CALIS 过程”一章、Bollen (1989) 和 Kline (201判别主成分)。

注意：“结构化方程模型”平台中的所有模型都用一个平均结构来估计，这意味着包含一个“常数”项。若不想让结构基于观测变量的均值，则应按照默认模型规格中的方式自由估计均值。

模型类型

本节介绍了可以在 “结构化方程模型”平台中拟合的一些不同类型的模型：

- **路径分析**支持您检验观测到的变量之间关联的备选解释模型。当研究中每个关注的结构只有一个变量可用时，通常使用此方法。也许最简单的“路径分析”模型是一个标准回归模型，在该模型中 X 预测 Y。SEM 平台支持您拟合该简单回归模型，但您也可以指定更有趣的模型。例如，您可能有变量 Z，根据理论或之前的研究，设定该变量为 $X \Rightarrow Y$ 关系的中介。换言之，X 预测 Z，然后 Z 预测 Y。因此，最初的 $X \Rightarrow Y$ 关系可能只因在原始模型中排除了 Z 才存在。可通过执行以下步骤在 SEM 平台中执行路径分析：
  1. 在启动窗口中选择所有观测到的变量，点击“模型变量”，然后点击“确定”。
  2. 在 “模型规格”报表中，在 “自列表”中选择预测变量，在 “至列表”中选择对应结果，然后点击单向箭头按钮。

  注意：所有外生变量 （那些没有指向它们的任何单向箭头的变量）必须在模型中自由相关，除非检验的是零相关的假设。使用双向箭头按钮指定这些协方差。

- **确认性因子分析** (CFA) 支持您检验备选测量模型。CFA 通常用于调查开发，并用作拟合结构化回归模型之前的初始步骤。SEM 平台通过执行以下步骤支持拟合确认性因子分析模型：

1. 在启动窗口中选择所有观测到的变量，点击“模型变量”，然后点击“确定”。
2. 使用 “模型规格”下的 “至列表”，选择假定要加载到潜在变量上的变量。
3. 在 “至列表”下方的框中输入潜在变量的名称，然后点击添加潜在变量 ⊕ 按钮。
4. 重复该过程，直到指定了模型的所有潜在变量。

请注意，SEM 平台总是包含一个平均结构，因此所有观测到的变量都列在“均值 / 截距”列表中，作为 “常数”项的结果。此外，若在启动窗口中选择了 “标准化潜在变量”选项，则通过将其第一个指标的载荷设置为 1 （默认值）或将其方差设置为 1，可自动识别所有潜在变量。最后，传统的 CFA 模型允许所有潜在变量共变。通过选择 “自列表”和 “至列表”中的所有潜在变量，然后点击双向箭头按钮，可以指定这些协方差。

- **结构化回归** (SR) 模型亦称**带潜在变量的路径分析**。这些模型通常在通过确认性因子分析 (CFA) 确定为合适的测量模型后使用。SR 模型支持您检验潜在变量之间特定的关系模式。换言之，尽管 CFA 不会针对潜在变量之间的效应强加任何方向（所有潜在变量都允许自由共变），但 SR 模型会强加。在假设管理层“领导力”在工作场所会导致较少的团队“冲突”和更高的员工“满意度”的例子中，“领导力”潜在变量可以预测“冲突”和“满意度”潜在变量。通过执行以下步骤，您可以在执行 CFA 后指定这些方向效应（回归）：

  1. 在启动窗口中选择所有观测到的变量，点击“模型变量”，然后点击“确定”。
  2. 使用 “模型规格”下的 “至列表”，选择假定要加载到潜在变量上的变量。
  3. 在 “至列表”下方的框中输入潜在变量的名称，然后点击添加潜在变量 ⊕ 按钮。
  4. 重复该过程，直到指定了模型的所有潜在变量。
  5. 在 “模型规格”报表中，在 “自列表”中选择预测变量，在 “至列表”中选择对应结果，然后点击单向箭头按钮。

- **潜在变量增长曲线** (LGC) 模型支持您拟合和检验重复测量数据的替代潜在轨迹。这些模型与混合模型框架中的随机效应模型非常相似。通常，您希望将无增长模型与线性模型进行比较。在无增长模型中，个体的起点可能不同，但轨迹平坦。在线性模型中，个体的起点和线性斜率都会随时间而变化。若有足够的数据可用，您还可以将这些模型与二次模型进行比较，在二次模型中，个体的起点以及线性和二次变化率均随时间而变化。您可以使用 SEM 平台模型规格来拟合 LGC 模型，或是使用模型快捷方式通过执行以下步骤来简化 LGC 模型的拟合：

  1. 在启动窗口中选择所有观测到的变量（重复测量），点击“模型变量”，然后点击“确定”。

  注意：为了使“潜在变量增长”模型快捷方式正确地指定模型，观测到的变量必须按时间顺序升序列出，并且必须具有相等的时间间隔。

  2. 使用 “模型快捷方式”选项，选择 “纵向分析” > “仅截距增长曲线”模型，然后点击“运行”。
  3. 使用 “模型快捷方式”选项，选择 “纵向分析” > “线性潜在变量增长曲线”模型，然后点击 “运行”。

4. 使用 “模型快捷方式” 选项，选择 “纵向分析” > “二次潜在变量增长曲线” 模型，然后点击 “运行”。

然后，“模型比较” 表会显示备选的拟合指数，并且可以选择最佳模型。

- 按照上述步骤确定理想的增长轨迹后，可以使用**条件潜在变量增长曲线**模型。此时，可以将截距和变化因子的预测变量添加到模型中。这些预测变量可能被证明是决定生长过程初始得分和随后变化的重要因子。要拟合条件 LGC，请在启动窗口中选择所有观测到的变量（重复测量），包括潜在变量的假设预测变量。确保预测变量是 “模型变量” 列表中最后的变量，以便于执行以下步骤：

  1. 使用 “模型快捷方式” 选项选择适当的增长轨迹。该选项指定 LGC 模型中的所有变量，包括预测变量。因此，您需要从增长过程中排除预测变量，并正确地将它们指定为预测变量。
  2. 在 “载荷” 列表中找到预测变量，选择涉及它们的所有效应，然后点击 “删除”。
  3. 在 “自列表” 中选择预测变量，在 “至列表” 中选择 “截距” 或 “斜率”。
  4. 点击单向箭头以指定条件 LGC。

  注意：若有多个预测变量，则必须通过在 “自列表” 和 “至列表” 中选择预测变量并点击双向箭头按钮来指定它们的协方差。

- **多组分析**模型支持您为 SEM 框架中的任意模型指定分组变量。然后在各组之间评估该模型，这样您就能够对不同的总体做出推断。您可以使用以下步骤指定一个多组分析模型：

  1. 选择您想在启动窗口中建模的观测变量，然后点击 “模型变量”。
  2. 选择一个分类分组变量（通常是一个水平很少的变量），点击 “组”，然后点击 “确定”。
  3. 使用 “模型规格” 报表指定您选择的模型。要添加回归路径，请在 “自列表” 中选择预测变量，在 “至列表” 中选定相应的结果，然后点击单向箭头 按钮。要添加协方差路径，请执行相同的步骤，但改为点击双向箭头 按钮。要添加潜在变量，请在 “至列表” 中选择其指标，然后点击 “至列表” 下方的添加潜在变量 按钮。
  4. 使用 “联合” 选项卡路径图选择边缘，并点击 “设置等于” 在组间应用等式约束。在 “联合” 选项卡中对模型规格所做的任何更改都将应用于所有组。可以使用组特定选项卡应用组特定约束或规格更改。默认情况下，模型中的路径可以跨组自由估计。

JMP PRO 结构化方程模型示例

在本例中，人力资源部门的一名员工希望开展一项调查，以衡量关键的工作场所结构。本例使用 “结构化方程模型” 平台构建了一个确认性因子分析模型。您可以使用该模型分析对 200 名员工的调查回复，了解他们工作场所的各个方面。该调查包含对与工作满意度相关的 11 个问题的回复。该确认性因子分析模型将调查问题的答复与领导力特征、角色冲突和整体工作满意度等潜在变量联系起来。

1. 选择**帮助** > **样本数据文件夹**，然后打开 Job Satisfaction.jmp。

2. 选择**分析 > 多元方法 > 结构化方程模型**。
3. 从**支持 _L** 一直选择到**主管 _S**，然后点击**模型变量**。
4. 点击**确定**。

   “结构化方程模型”报表 “模型规格”分级显示项随即显示。

创建潜在变量

5. 在 “至列表”中，从**支持 _L** 一直选择到**互动 _L**。在 “至列表”下方的框中键入 “领导力”，然后点击添加潜在变量 ⊕ 按钮。
6. 在 “至列表”中，从**人员 _C** 一直选择到**互动 _C**。在 “至列表”下方的框中键入 “冲突”，然后点击添加潜在变量 ⊕ 按钮。
7. 在 “至列表”中，从**常规 _S** 一直选择到**主管 _S**。在 “至列表”下方的框中键入 “满意度”，然后点击添加潜在变量 ⊕ 按钮。

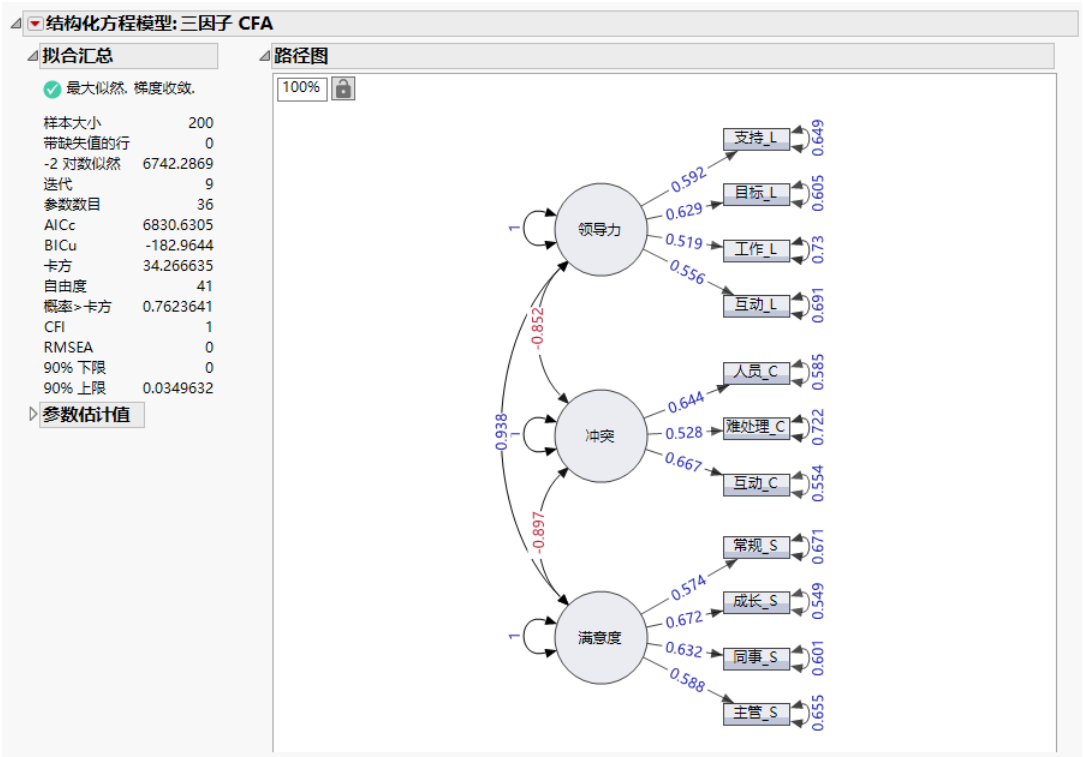
添加协方差并拟合模型

8. 同时在 “从列表”和 “至列表”中，选择**领导力**、**冲突**和**满意度**。点击双向箭头 ↔ 按钮。
9. 在 “模型名称”（位于 “模型规格”报表左上角）下方的文本框中，键入 “三因子 CFA”。
10. 点击**运行**。
11. （可选。）点击 “结构化方程模型 : 三因子 CFA”旁边的红色小三角，然后选择**路径图设置 > 布局 > 从上到下**。
12. （可选。）点击 “参数估计值”旁边的灰色展开图标。

    关闭 “参数估计值”报表将支持您查看整个路径图。
13. 右击路径图并选择**显示 > 显示估计值 > 标准化**。

    “路径图”中的数字现在表示该模型的标准化参数估计值。

图 8.2 “结构化方程模型”报表

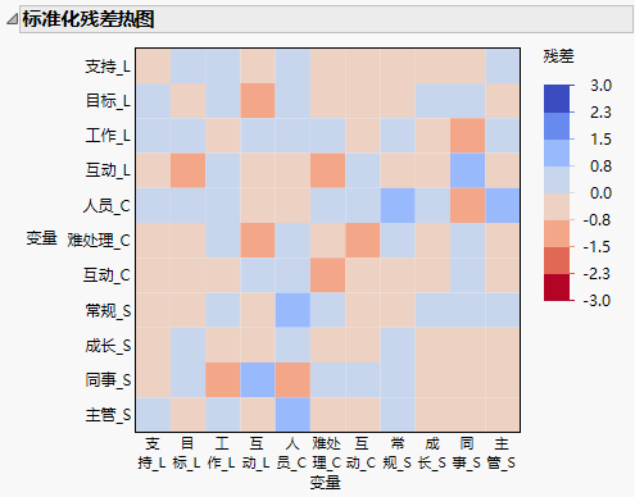


“拟合汇总”报表中列出的该模型的卡方统计量为 34.27，自由度为 41。请注意，相应的 p 值为 0.7 判别主成分 24，该值不显著。这表明，没有证据可以拒绝模型拟合良好这一原假设。因此，您可以得出结论：该模型对数据拟合良好。

卡方值取决于样本大小，因此，一些拟合良好的模型仍然可以生成显著的卡方值。比较拟合指数 (CFI) 和近似的均方根误差 (RMSEA) 为确定模型拟合提供了额外的指导。这些指数介于 0 和 1 之间。 CFI 值最好大于 0.90， RMSEA 值最好小于 0.10（Browne and Cudeck 1993 ；Hu and Bentler 1999）。在此， CFI 为 1 且 RMSEA 为 0，指示拟合极佳。您得出结论：该调查是测量领导力、冲突和满意度等潜在变量的一个不错的工具。

14. 点击 “结构化方程模型 : 三因子 CFA”旁边的红色小三角，然后选择**热图 > 标准化残差热图**。

图 8.3 标准化残差热图

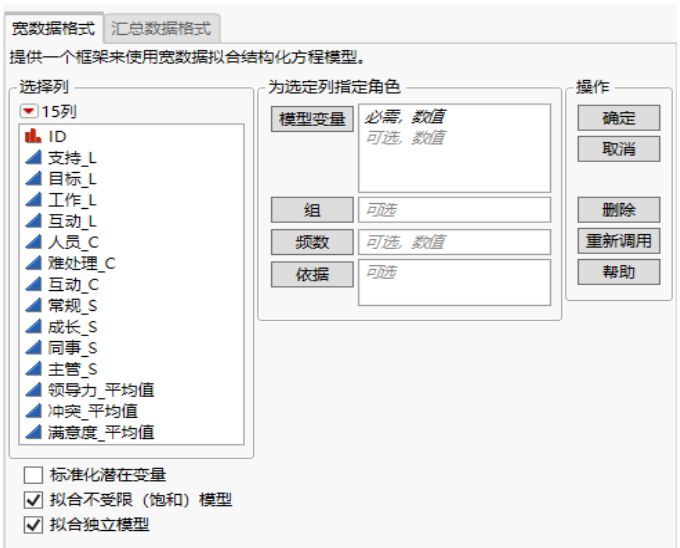


“标准化残差热图”在正方向或负方向上没有超过 2.0 个单位的值；这进一步证明了该模型拟合数据良好。残差也可以在更细粒度上诊断拟合不良的模型。该模型中的标准化残差没有显示出局部失拟的证据。

JMP PRO 启动“结构化方程模型”平台

通过选择**分析 > 多元方法 > 结构化方程模型**来启动 “结构化方程模型”平台。

图 8.4 “结构化方程模型”启动窗口



有关“选择列”红色小三角菜单中选项的详细信息，请参见《使用 JMP》。

启动窗口包括两种不同数据格式的选项卡。

宽数据格式 选择这些数据表：其表中每行对应于单个观测，而列包含了要在模型中使用的变量。仅包含缺失值的行会从分析中排除。

汇总数据格式 选择这些数据表：其数据汇总为相关性或协方差矩阵。均值与标准差也可以指定为列。若未指定它们，则均值假定为 0，标准差假定为输入矩阵对角线的平方根。

JMP PRO 启动窗口选项

模型变量 要包含在模型中的列。必须指定至少一列。所有列都必须具有数值型数据类型和连续型建模类型。

对于 “汇总数据格式”，“模型变量”列是包含汇总数据的相关性或协方差矩阵的列。

组 （仅可用于 “宽数据格式”。）为多组分析指定分组变量的列，该列支持您检验组间效应的等式约束。该变量必须为仅带有几个水平的分类变量。

频数 一列，列中的数值为分析中的每行分配一个频数。

依据 （仅可用于 “宽数据格式”。）一列，用于创建为变量的每个水平包含单独分析的报表。若分配了多个 “依据”变量，则为 “依据”变量水平的每个可能组合生成单独分析。

警告：使用 “依据”变量不能拟合多组分析模型。要执行多组分析，必须指定 “组”变量。

均值 （仅可用于 “汇总数据格式” 。 ）与相关性或协方差矩阵的每行中的变量相对应的均值列。若未指定，则均值假定为 0。

标准差 （仅可用于 “汇总数据格式”。）与相关性或协方差矩阵的每行中的变量相对应的标准差列。若未指定，则标准差假定为输入矩阵对角线的平方根。

标签 （仅可用于 “汇总数据格式” 。 ）与相关性或协方差矩阵的每行中的变量相对应的标签列。若未指定，则 “模型规格”报表中的变量将使用包含输入矩阵的列的名称。

标准化潜在变量 若选定，该选项会通过将潜在变量的方差固定为 1 并允许对所有载荷进行自由估计来设置潜在变量的尺度。

拟合不受限模型 若选定，则在拟合第一个模型时，将自动拟合不受限模型。随后可以在 “模型比较”报表中将拟合模型与不受限模型进行比较。不受限模型是完全饱和模型，该模型拟合指定模型变量的所有均值、方差和协方差，而不对数据强加任何结构。

拟合独立模型 若选定，则在拟合第一个模型时，将自动拟合独立模型。随后可以在 “模型比较”报表中将拟合模型与独立模型进行比较。独立模型拟合指定的模型变量的所有均值和方差。指定模型变量之间的所有协方差都被固定为零，这会导致高度受限的模型。

样本大小 （仅可用于 “汇总数据格式”。）指定汇总数据表示的观测数。

JMP PRO “结构化方程模型”报表

“结构化方程模型”报表包含以下项：

- “‘模型规格’报表”
- “‘模型比较’报表”
- “‘卡方差异检验’报表”
- “‘结构化方程模型拟合’报表”

JMP PRO “模型规格”报表

“结构化方程模型”平台中的“模型规格”报表包含用于指定模型的控件。在启动窗口中点击“确定”后，“模型规格”报表的“关系图”选项卡和“列表”选项卡中将显示默认的独立模型。独立模型包含指定的模型变量的所有均值和方差。

JMP PRO “规格”面板

“规格”面板包含用于构建模型的控件。

模型名称 支持您指定模型名称。在“操作”面板中点击“运行”后，将创建使用“模型名称”中的文本的模型报表。

自列表 列出在启动窗口中指定的模型变量以及“Constant”项。“Constant”项通过在“Constant”上回归每个变量来估计任何观测到的或潜在变量的均值。在点击箭头按钮之一将项添加到模型之前，先在该列表中选择一个或多个变量并且在“至列表”中选择一个或多个变量。若创建潜在变量，则潜在变量会添加到“自列表”和“至列表”。可以使用列表顶部的搜索过滤器控件过滤该列表中的项目。请参见“搜索过滤器选项”。

至列表 列出在启动窗口中指定的模型变量以及“Constant”项。“Constant”项通过在“Constant”上回归每个变量来估计任何观测到的或潜在变量的均值。在点击箭头按钮之一将项添加到模型之前，先在“自列表”中选择一个或多个变量并且在该列表中选择一个或多个变量。若创建潜在变量，则潜在变量会添加到“自列表”和“至列表”。必须选择“至列表”中的变量才能将潜在变量添加到模型中。可以使用列表顶部的搜索过滤器控件过滤该列表中的项目。请参见“搜索过滤器选项”。

单向箭头 → 指定“自列表”和“至列表”中选定的变量之间的关系类型。单向箭头等效于回归效应。

双向箭头 ↔ 指定“自列表”和“至列表”中选定的变量之间的关系类型。双向箭头等效于协方差效应。

添加潜在变量 + 向“自列表”和“至列表”添加潜在变量。变量的命名基于 + 按钮左侧框中的文本。

注意：在点击 ⊕ 按钮之前，必须在 “至列表”中选择潜在变量的指标。

删除潜在变量 ⊖ 删除 “自列表”或 “至列表”中当前选定的任何潜在变量。从 “自列表”和 “至列表”中删除某个潜在变量也会从模型中删除该潜在变量。

模型快捷方式 包含三类用于快速构建常见模型的选项。请注意，选择其中一个选项时，将清除当前所有模型规格。

规格帮助 包含的选项可快捷地处理指定模型的过程，例如为外生变量或内生变量添加协方差。“切换潜在变量尺度”选项还支持您更改潜在变量的尺度，这在拟合确认性因子分析时很有用。

共变外生变量 支持您快速指定现有外生变量之间的所有可能协方差。外生变量是模型中任何其他变量都无法预测的变量。大多数情况下，为使模型规格正确，这些变量之间的共变应该是非零的。

共变内生变量 支持您快速指定现有内生变量之间的所有可能协方差。内生变量是由模型中的一个或多个其他变量预测的变量。在某些情况下，为使模型规格正确，内生变量之间的残差方差应该是共变的。这种快捷方式最常适用于路径模型，这些路径模型中的一组结果对应一组预测变量，但既没有中介变量也没有潜在变量。

切换潜在变量尺度 在设置模型中潜在变量尺度的两种常用方法之间切换。要估计具有潜在变量的模型，必须设置潜在变量的尺度。以下两个选项是设置尺度的最常用方法：将第一个载荷固定为 1 或将潜在变量的方差固定为 1。第一种方法将潜在变量的尺度设置为指标的尺度，载荷固定为 1。第二种方法确保潜在变量的方差为 1，这可以与均值 0 配对，以确保潜在变量具有 *z* 得分量度。

固定测量值误差 支持您通过指定观测变量的误差方差估计值对无误差的潜在变量建模。您可以使用可靠性测度（如系数 omega）、不可靠性（1 - 可靠性）或测量误差方差来指定已知的误差方差。该选项可用于指定线性测量误差模型。

截面经典 包含用于确认性因子分析和中介分析的选项。

确认性因子分析 支持您选择潜在变量的指标并定制其名称。提供了在正交 （非相关）或斜交 （相关）规格之间进行选择的选项。

中介分析 支持您检验因果理论，因果理论假设通过一个中介变量将因果预测变量的效应传递到结果。这些模型的一个关键方面涉及预测变量通过中介变量对结果产生的间接 （或介导）效应的显著性。该快捷方式支持您为因果、中介和结果角色指定变量，从而可以指定简单的中介模型。可以使用复选框通过选择多个中介变量来指定多个中介模型。

纵向分析 包含用于常见潜在变量增长曲线模型的选项。潜在变量增长曲线模型是一系列纵向模型，这些模型用于描述观测单元随时间变化的轨迹特征。这些模型的标准规格支持您获得总体中平均轨迹的估计值以及与该平均值的离散测度。提供了四种常见的增长曲线快捷方式；“仅截距的潜在变量增长曲线”、“线性潜在变量增长曲线”、“二次潜在变量增长曲线”和“潜在基函数增长曲线”。

注意：若在选择某一潜在变量增长曲线选项时选择了“至列表”中的变量，则模型仅应用于选定的变量。否则，该模型将应用于“至列表”中的所有变量。

仅截距的潜在变量增长曲线 指定一个模型，该模型表示观测单元的平面轨迹。该模型有一个潜在变量，该变量可以解释为给定过程的截距。截距有一个估计的均值和方差，分别表征总体中过程的平均水平以及个体单元与该平均值的偏离。该模型通常用作与其他潜在变量增长曲线模型进行比较的基线。例如，若线性潜在变量增长曲线模型比仅截距模型的拟合效果更好，则可以确定存在线性趋势。

线性潜在变量增长曲线 指定一个模型，该模型表示观测单元的线性轨迹。该模型有两个潜在变量：截距和线性斜率。这些潜在变量中的每一个都有一个均值和一个方差，这使您能够分别解释观测单元中的平均水平和变化以及偏离该平均值的程度。所有潜在变量（或增长因子）都允许共变。斜率的第一个载荷固定为 0，这将截距集中在第一个时间点。请注意，该快捷方式假定变量按时间升序排列，因此斜率因子的载荷从 0 开始，对于每个变量都按 1 个单位递增。

二次潜在变量增长曲线 指定一个模型，该模型表示观测单元的二次轨迹。该模型有三个潜在变量：截距、线性斜率和二次斜率。这些潜在变量中的每一个都有一个均值和一个方差，这使您能够分别解释二次轨迹平均值以及与该平均值的各个偏差。所有潜在变量（或增长因子）都允许共变。线性和二次斜率的第一个载荷固定为 0，这将截距集中在第一个时间点。请注意，该快捷方式假定变量按时间升序排列，因此线性斜率因子的载荷从 0 开始，对于每个变量都按 1 个单位递增。二次斜率因子的载荷是线性斜率因子的荷载平方值。

潜在基函数增长曲线 指定不具有已知函数形式的增长曲线。该模型经常用于非线性轨迹。该模型有两个潜在变量：截距和非线性斜率，每个变量都有一个均值和一个方差，以便能够解释轨迹平均值以及与该平均值的各个偏差。这些潜在变量（或增长因子）都允许共变。非线性斜率的第一个载荷固定为 0，这将截距集中在第一个时间点。非线性斜率的最后一个载荷固定为 1，所有中间载荷均可自由估计。该规格支持您将载荷估计值解释为变化比例。请注意，该快捷方式假定变量按时间升序排列。

拟合并比较增长模型 （不可用于多组分析。）支持您使用卡方差异检验比较仅截距、线性和二次增长曲线模型的拟合。纵向分析的一个共同目标是描述数据变化模式的特征。通过比较这三种模型的拟合，可以确定哪种类型的轨迹能最佳拟合数据，从而实现这一目标。默认情况下，将折叠所有拟合模型报表，以便在检查特定模型结果之前可以识别最佳拟合模型。

多元潜在变量增长曲线 支持快速指定多元增长曲线模型。可以选择给定过程的变量子集以及将为该过程指定的增长曲线类型。⊞ 按钮支持您添加所指定的各个一元模型。默认情况下，允许所有潜在变量共变，这使您能够对增长因子之间的关联进行推断。请注意，该快捷方式假定每个过程的变量都按时间升序排列。

JMP PRO “操作”面板

“操作”面板中的按钮针对“列表”选项卡中的指定模型变量列表进行操作。可使用以下操作：

固定为 支持您将当前选定效应的参数值固定为常数值。为变量固定参数值时，固定值将显示在项名称后面的括号中。

注意：若点击“确定”时保持默认值为 0，则将从模型中删除选定效应。

设置等于 支持您将两个或更多选定效应的参数值限定为相等。若将两个或更多效应设置为具有相同的参数值，则在项名称后的括号中会显示任意字母数字标签 (“c1”)。若多组效应的参数值都设置为彼此相等，则使用顺序编号 （“c1”、“c2”等）。

注意：仅允许在同一类型的参数中使用等式约束。

自由 支持您删除针对模型中选定效应的约束。约束包括已设置为固定值的效应或已设置为与其他效应相等的效应。

删除 从模型中删除选定的效应。

撤销 撤销对模型的最后一次修改。

恢复 恢复对模型上次撤销的更改。

运行 为当前指定模型拟合和创建报表。

重置 将模型规格重置为独立模型，这是默认值。

JMP PRO “关系图”面板

“关系图”面板包含支持您调整模型关系图布局的按钮。第一个按钮支持您旋转关系图中的显变量（用矩形表示）。第二个按钮支持您循环查看关系图的两种不同排列。第三个按钮支持您定制关系图。请参见“定制路径图”。

JMP PRO “详细信息”面板

“详细信息”面板包含有关“模型规格”报表中当前指定的模型的信息。有关模型的信息包括显变量的数量、潜在变量的数量、自由估计参数的数量和自由度的数量。您还可以调整最大迭代次数。

JMP PRO “关系图”选项卡

“关系图”选项卡包含一个支持您对当前指定模型进行可视化的模型关系图。潜在变量用圆圈表示，显变量用矩形表示。单向箭头表示载荷和回归。双向箭头表示方差和协方差。方差由从一个变量到自身的双向箭头指定。协方差由两个变量之间的双向箭头指定。

路径图是交互式的，因此可以拖动各项来排列它们。您还可以使用箭头键移动路径图中的选定项。可以使用“关系图”面板中的按钮重新排列路径图中的项。可以使用路径图弹出式菜单中的选项来添加或删除路径图中的项。右击路径图本身时，会出现路径图弹出式菜单。弹出式菜单的内容根据您是点击空白绘制区的某个部分还是点击路径图的某个元素而有所不同。也可以通过调整路径图左上角的缩放设置来放大或缩小路径图。路径图左上角还包含一个“锁定”按钮，支持您锁定路径图中各项的位置。当路径图被锁定时，“锁定”按钮为蓝色。在路径图中进行选择后，可以通过在弹出式菜单中选择**选择 > 隐藏选择**选项来隐藏选择。当路径图中的项被隐藏时，路径图顶部会显示一个“全部取消隐藏”按钮，该按钮支持您显示所有隐藏的项。

注意：在启动窗口中指定“组”变量时，“关系图”选项卡嵌套在多个组选项卡中。您可以使用“关系图”选项卡左侧的面板选项卡在这些选项卡之间导航。有一个“联合”选项卡显示组模型集合，一个面板选项卡显示“组”变量的每个水平。

有关路径图定制选项的详细信息，请参见“定制路径图”。请参见“‘关系图’面板”、“‘详细信息’面板”和“路径图设置”。

提示：要将路径图复制为图像，请右击该图并选择**复制关系图**。要保持尽可能高的质量，请将剪贴板图像粘贴为向量图形。

JMP PRO “列表”选项卡

“列表”选项卡包含模型中每种类型变量的列表。这些列表按照在模型图中表示它们所用的箭头来分类。单向箭头用于指定均值或截距、载荷和回归。双向箭头用于指定方差和协方差。模型的均值和截距仅显示在“列表”选项卡中，而不出现在“关系图”选项卡中。

注意：在启动窗口中指定“组”变量时，“列表”选项卡嵌套在多个组选项卡中。您可以使用“列表”选项卡左侧的面板选项卡在这些选项卡之间导航。有一个“联合”选项卡显示组模型集合的“列表”选项卡，一个面板选项卡显示“组”变量的每个水平。

可以使用每个列表顶部的搜索过滤器控件过滤每个列表中的项目。请参见“搜索过滤器选项”。

JMP PRO “状态”选项卡

“状态”选项卡包含用于模型可识别性的检查。必须标识结构化方程模型以获取可靠估计值。SEM 上下文中的识别意味着：根据输入数据的总体协方差矩阵可对模型中的每个参数求出唯一解。因为可以在 SEM 框架中指定如此多的模型，所以有多种识别规则。“状态”选项卡包含三个信息面板：“识别规则”、“模型详细信息”和“数据详细信息”。

- “识别规则”面板包含一个列表，其中包含最多八个适用于指定模型的规则。一些规则是必要的，而另一些规则是保证模型可识别的充分条件。若必要规则失败，则必须在拟合模型之前采取步骤更正这些规则。若充分规则失败，您不一定需要解决任何问题。有时，充分规则可能会失败，而这并不能证明模型有任何错误。请注意，所有规则都假设协方差矩阵是正定的。若协方差矩阵不是正定的，则会在“模型详细信息”面板下显示警告。

提示：有关特定识别规则的详细信息，请点击该规则所在表中的行，该规则的说明将显示在该表的右侧。

- “模型详细信息”面板包含当前指定模型的描述性值列表。这些值可用于检测模型的潜在问题。
- “数据详细信息”面板包含用于输入数据的描述性值列表。这些值可用于检测数据的潜在问题。

若指定的数据列中存在奇异性，则报表包含“奇异性详细信息”表。

注意：在启动窗口中指定“组”变量时，“状态”选项卡嵌套在多个组选项卡中。您可以使用“状态”选项卡左侧的面板选项卡在这些选项卡之间导航。有一个“联合”选项卡显示组模型集合的“列表”选项卡，一个面板选项卡显示“组”变量的每个水平。

“状态”选项卡本身包含一个动态图标，该图标显示指定模型的当前状态。该图标指定以下状态：

	所有适用识别规则均通过。
	至少有一个必要识别规则未通过，您必须在拟合模型之前采取步骤更正问题。
	至少有一个非必要的识别规则未通过，需要进一步调查以确定模型是否指定无误。通常，SEM 的高级应用会导致这种情况；这并不意味着模型有问题。相反，它表明识别规则不能保证模型可识别。

图 8.5 “状态”选项卡示例

JMP PRO 搜索过滤器选项

“模型规格”报表的“列表”选项卡中显示的“自列表”、“至列表”和列表框包含搜索过滤器，这些过滤器支持您过滤特定列表中的项目。

点击搜索框旁边的下箭头按钮以优化搜索。

包含词条 返回包含一部分搜索条件的项。搜索“ease oom”返回如“Release Zoom”这样的消息。

包含短语 返回包含完全搜索条件的项。搜索“text box”返回包含“text”后面直接跟着“box”的条目（例如，“Context Box”和“Text Box”）。

以短语开头 返回以搜索条件开始的项。

以短语结尾 返回以搜索条件结束的项。

整个短语 返回包含整个字符串的项。搜索“text box”返回仅包含“text box”的条目。

正则表达式 允许您在搜索框中使用通配符 (*) 和句点 (.)。搜索“get.*name”查找包含“get”后面跟着一个或多个单词的项。它返回“Get Color Theme Names”、“Get Name Info”、“Get Effect Names”等。

反转结果 返回不匹配搜索条件的项。

匹配全部词条 返回同时包含两个字符串的项。搜索 “t test”返回包含任一搜索字符串或两者的元素：“Pat Test”、“Shortest Edit Script”和 “Paired t test”。

忽略大小写 忽略搜索条件中的大小写。

全字匹配 基于 “匹配全部词条”设置返回包含字符串中每个单词的项。若您搜索 “data filter”，并且选择了 “匹配全部词条”，则返回同时包含 “data”和 “filter”的条目。

JMP PRO “模型比较”报表

“结构化方程模型”平台中的“模型比较”报表包含已经拟合的所有模型的表。使用第二个列中的图标控制显示在“模型比较”报表下方的模型报表。“模型名称”右侧的其余列支持您根据各种条件比较模型。

模型的“AICc 权重”值可解释为在其中的一个拟合模型成立的前提下，特定模型为真实模型的概率。因此，AICc 权重最接近 1 的模型是最好的拟合。仅使用非缺失 AICc 值计算 AICc 权重，该权重定义如下：

AICc 权重 = exp[-0.5(AICc-min(AICc))] / sum(exp[-0.5(AICc-min(AICc))])

其中， min(AICc) 是拟合模型中的最小 AICc 值。

有关“模型比较”报表的其他准则的信息，请参见 “结构化方程模型拟合”报表”。

注意：若某个模型不收敛，则“模型比较”报表中模型名称的开头会显示一个星号。

为了提供拟合模型的性能的上下文，“模型比较”报表中默认显示以下两个模型：

不受限 （饱和） 拟合指定模型变量的所有均值、方差和协方差，而不对数据强加任何结构。

独立 拟合指定模型变量的所有均值和方差，并将所有协方差固定为零。

“模型比较表”选项

选择表中的某行后，“模型比较”表的弹出菜单中提供以下选项。

重命名模型 （仅当您在表中选择了“不受限（饱和）”或“独立”之外的模型时才适用。）支持您更改已经拟合的模型的名称。在“模型比较”表中更改模型的名称时，报表中相应分级显示项中的模型名称也将更新。

设置为独立模型 （仅当您在表中选择了 “不受限 （饱和） ”或 “独立”之外的模型时才适用。 ）使用报表中的选定模型替换 “独立”模型。该选项尤其适用于纵向模型。当您更改“独立”模型时，“模型比较”表中的比较拟合指数将随之更新。

警告：指定不正确的独立模型可能导致无效拟合指数。

重置默认独立模型 （仅当已将默认独立模型以外的模型设置为独立模型时才可用。）将默认独立模型恢复为“独立”模型以用于模型比较。该选项还将“模型比较”表中的比较拟合指数更新为其原始值。

“模型比较”报表选项

“模型比较”表下显示两个选项。在“模型比较”表中通过点击行使其突出显示后，这些选项可用。

比较选定模型 （仅当表的两行或更多行突出显示时才可用。）计算选定行的所有嵌套模型组合之间的嵌套卡方差异检验。

提示：若您选择的任何模型没有嵌套，则会出现警告，并且“卡方差异检验”报表中不会出现非嵌套模型组合。

清除选择 （仅当表的一个或多个行突出显示时才可用。）从表行中清除所有选择。

JMP PRO “卡方差异检验”报表

“结构化方程模型”平台中的“卡方差异检验”报表包含嵌套卡方检验的表。该表包含定义模型的两列。第一列包含两个模型中约束较强的模型，第二列包含两个模型中约束较弱的模型。较小模型嵌套在较大模型中。其余列显示了卡方值、自由度、CFI 和 RMSEA 中的差异，以及嵌套卡方检验的 p 值。列名中的 Δ 符号指示差异。显著的 Δ 卡方值指示嵌套模型中的附加约束在统计上显著增加了失配，应保留约束较少的模型。由于卡方检验受样本大小的影响，以致于随着样本量的增加，卡方检验更可能显著，所以还应该考虑 ΔCFI 和ΔRMSEA；理想情况下，ΔCFI 不应超过 –0.01，ΔRMSEA 不应超过 0.015 (Chen 2007)。

警告：该报表中的差异检验仅对嵌套模型有意义。

您可以通过点击红色的 X 按钮删除报表的任何行。要完全删除报表，必须删除表中的所有行。若在主报表中删除“结构化方程模型”节点，则将从表中删除涉及该模型的所有差异检验。

JMP PRO “结构化方程模型拟合”报表

每次在“模型规格”报表中点击“运行”时，都会显示指定模型的“结构化方程模型”报表。默认情况下，该报表包含“拟合汇总”报表、“参数估计值”报表和“路径图”。

注意：在启动窗口中指定“组”变量时，“组”变量的每个水平都有一个单独的模型拟合报表。您可以使用“结构化方程模型拟合”报表分级显示框上方的组选项卡在这些报表之间导航。

拟合汇总 有关模型拟合（包括收敛状态和估计方法）的信息表。在启动窗口中指定“组”变量时，该表另有一列包含统计量，这些统计量仅引用当前模型拟合选项卡中的“组”变量的水平。该表中报告以下统计量：

样本大小 用于拟合模型的观测（行）数。

带缺失值的行 包含至少一个缺失值的观测（行）数。所有缺失值都使用全信息最大似然 (Finkbeiner 1979) 来处理。

-2 对数似然 拟合模型的对数似然乘以 -2。该值可用于比较嵌套模型；两个模型的 -2 对数似然值之间的差值服从卡方分布，其自由度等于两个模型之间的自由度差值。请参见《拟合线性模型》。

迭代 用于拟合模型的迭代数。

参数数目 模型中的自由估计参数数目。

AICc 校正的 Akaike 信息准则。该值可用于比较模型，其中较小的数字表示模型拟合更好。请参见“AICc、BIC 和 BICu”。

BICu 相对于不受限模型 (BICu) 的 BIC 是 Bayes 信息准则的重新表述。BICu 定义为与不受限模型的比较。负值支持拟合模型，正值支持不受限模型。类似于其他信息准则，该值可用于比较模型，其中两个模型之间的较小数值指示拟合较好的模型。请参见“AICc、BIC 和 BICu”。

卡方 模型的卡方统计量。

自由度 模型拟合的卡方检验的自由度。

概率>卡方 模型的卡方统计量的 *p* 值。

CFI Bentler 比较拟合指数 (CFI) 为确定模型拟合提供了额外的指导。CFI 介于 0 和 1 之间。值最好大于 0.90（Browne and Cudeck 1993；Hu and Bentler 1999）。请参见“CFI”。

RMSEA 近似的均方根误差 (RMSEA) 为确定模型拟合提供了额外的指导。RMSEA 介于 0 和 1 之间。值最好小于 0.10（Browne and Cudeck 1993；Hu and Bentler 1999）。请参见“RMSEA”。

90% 下限 RMSEA 的 90% 置信下限。请参见“RMSEA”。

90% 上限 RMSEA 的 90% 置信上限。请参见“RMSEA”。

参数估计值 模型参数的估计值表。该表按均值 / 截距、载荷、回归和方差等部分来组织。对于每个估计值，给出标准误差 (Std Error)、Wald 检验统计量 (Wald Z) 和相应的 *p* 值（概率 >|Z|）。在启动窗口中指定“组”变量时，该表仅包含针对当前模型拟合选项卡中的“组”变量水平的参数估计值。

提示：“参数估计值”表另含一些隐藏列。要显示这些列，请右击该表，然后从**列**子菜单中选择其他列。

路径图 显示拟合模型的路径图表示。请参见“‘关系图’选项卡”。在启动窗口中指定“组”变量时，该关系图仅表示当前模型拟合选项卡中的“组”变量的水平。

JMP PRO “结构化方程模型”平台选项

“结构化方程模型”红色小三角菜单包含以下选项：

路径图设置 包含以下用于修改路径图显示的选项：

定制关系图 支持您定制路径图的许多方面。请参见“定制路径图的选项”。

布局 包含更改路径图总体形状的两个选项。您可以选择“从左到右”的布局或“从上到下”的布局。

提示：还可以拖动路径图中的项以更改特定项的排列方式。

复制关系图 将路径图的图像保存到剪贴板。要保持尽可能高的质量，请将剪贴板图像粘贴为向量图形。

复制关系图属性 将当前路径图属性复制到剪贴板。随后可将属性粘贴到另一个 SEM 路径图中。

粘贴关系图属性 将剪贴板中的路径图属性粘贴到当前 SEM 路径图中。

描述性统计量 包含以下用于生成描述性统计量的选项：

一元简单统计量 显示或隐藏一元简单统计量报表。该报表中的统计量是独立于可能具有缺失数据的其他列对每个列进行估计的。

全信息多元统计量 显示或隐藏多元简单统计量报表。该报表中的统计量使用全信息最大似然来估计，以解释缺失数据。

启动探索离群值 启动“探索离群值”平台。请参见《预测和专业建模》。

启动探索缺失值 （有缺失值时不可用。）启动“探索缺失值”平台。请参见《预测和专业建模》。

添加显变量 支持您向当前模型添加新的显变量。选择要添加的新的显变量后，将启动一个新的“结构化方程模型”报表。新报表对新添加的显变量使用现有的模型规格。

删除显变量 支持您在删除指定显变量的情况下启动新的“结构化方程模型”报表。选择要删除的显变量后，将启动一个新报表。新报表在已删除显变量的情况下使用现有模型规格。

复制模型规格 将当前结构化方程模型规格复制到剪贴板。随后可以将模型规格粘贴到另一个 SEM 平台报表中。

粘贴模型规格 将剪贴板中的模型规格粘贴到当前的模型规格中。

请参见《使用 JMP》获取有关下列选项的信息：

本地数据过滤器 显示或隐藏支持您过滤特定报表中使用的数据的本地数据过滤器。

重新运行 包含使您可以重复或重新启动分析的选项。在支持该功能的平台中，“自动重新计算”选项立即在相应报表窗口中反映您对数据表所做的更改。

平台首选项 包含的选项支持您查看当前平台首选项或更新平台首选项以匹配当前 JMP 报表中的设置。

保存脚本 包含的选项支持您保存可将报表重现到若干目标的脚本。

保存 “依据”组脚本 包含使您可以保存脚本的选项，可将为 “依据”变量的所有水平重新生成平台报表的脚本保存到多个不同的位置。仅当在启动窗口中指定 “依据”变量时才可用。

JMP PRO 模型选项

在 “模型规格”报表中点击 “运行”后，就会显示指定模型的 “结构化方程模型”报表。在启动窗口中指定 “组”变量时，模型选项会为 “组”变量的所有水平显示或隐藏报表中的元素。该报表有一个包含以下选项的红色小三角菜单：

显示路径图 在模型报表中显示或隐藏路径图。

路径图设置 包含以下用于修改模型路径图的选项：

定制关系图 支持您定制路径图的许多方面。请参见 “定制路径图的选项”。

布局 包含更改路径图总体形状的两个选项。您可以选择 “从左到右”的布局或 “从上到下”的布局。

提示：还可以拖动路径图中的项以更改特定项的排列方式。

复制关系图 将路径图的图像保存到剪贴板。要保持尽可能高的质量，请将剪贴板图像粘贴为向量图形。

复制关系图属性 将当前路径图属性复制到剪贴板。随后可将属性粘贴到另一个 SEM 路径图中。

粘贴关系图属性 将剪贴板中的路径图属性粘贴到当前 SEM 路径图中。

拟合指数 显示或隐藏包含各种指数值的报表，这些指数值支持您评估拟合的模型。除 “拟合汇总”报表中显示的值之外 （请参见 “‘结构化方程模型拟合’报表”），“拟合指数”报表还包含以下指数值：

BIC Bayes 信息准则。该值可用于比较模型，其中较小的数字表示模型拟合更好。请参见 “AICc、BIC 和 BICu”。

RNI 相对非中心指数 (RNI) 为确定模型拟合提供了额外的指导。该值等价于 CFI，但不以 1 为边界。值最好大于 0.90。请参见 “RNI”。

TLI Tucker-Lewis 指数 (TLI) 为确定模型拟合提供更多指导。该指数也称为非规范拟合指数 (NNFI)。TLI 介于 0 和 1 之间。值最好大于 0.95 (West et al. 2012)。请参见 “TLI”。

NFI 规范拟合指数 (NFI) 为确定模型拟合提供了额外的指导。NFI 介于 0 和 1 之间。值最好大于 0.95 (West et al. 2012)。请参见 “NFI”。

修正的 GFI 修正的拟合优度指数为确定模型拟合提供了额外的指导。修正的 GFI 介于 0 和 1 之间。值最好大于 0.95 (West et al. 2012)。请参见 “修正的 GFI 和修正的 AGFI”。

修正的 AGFI 修正的已调整拟合优度指数为确定模型拟合提供了额外的指导。修正的 AGFI 介于 0 和 1 之间 (West et al. 2012)。请参见“修正的 GFI 和修正的 AGFI”。

RMR 均方根残差 (RMR) 为确定模型拟合提供了额外的指导。RMR 的残差来自观测的协方差与隐含模型的协方差之间的差值。RMR 为正且最好为较小值 (West et al. 2012)。请参见“RMR 和 SRMR”。

SRMR 标准化均方根残差 (RMR) 为确定模型拟合提供了额外的指导。SRMR 的残差来自观测的协方差与隐含模型的协方差之间的标准化差值。SRMR 为正且最好为较小值 (West et al. 2012)。请参见“RMR 和 SRMR”。

注意：有关“拟合指数”报表中其他值的说明，请参见“‘结构化方程模型拟合’报表”。

拟合汇总 显示或隐藏包含模型拟合详细信息的报表。

参数估计值 显示或隐藏包含模型的非标准化参数估计值的报表。

标准化参数估计值 显示或隐藏包含模型的标准化参数估计值的报表。

置信区间 显示或隐藏“参数估计值”报表和“标准化参数估计值”报表中的置信区间。

总效应 （仅当模型包含至少一个回归或载荷变量且效应收敛时才可用。）显示或隐藏模型中总效应的未标准化和标准化估计值表。标准误差也包括在内。Bentler and Freeman (1983) 中对效应收敛检验进行了说明。表的右侧包含标准化估计值的条形图。

间接效应 （仅当模型包含中介变量且效果收敛时才可用。）显示或隐藏模型中间接效应的未标准化和标准化估计值表。标准误差也包括在内。Bentler and Freeman (1983) 中对效应收敛检验进行了说明。表的右侧包含标准化估计值的条形图。

提示：您可以在“间接效果”表中获取值的 bootstrap 估计值。要运行 bootstrap 分析，请右击包含您要 bootstrap 的统计量的表列，然后选择“Bootstrap”。请参见《基本分析》。

预测刻画器 支持您查看一组预测变量对一组结果变量的条件预期值的影响。选择该选项时，将出现一个窗口，您必须在其中选择一个或多个预测变量和一个或多个结果。预测和 95% 置信区间基于隐含模型的协方差矩阵。有关“预测刻画器”的详细信息，请参见《刻画器指南》。

注意：设置窗口中的初始变量列表仅限于与模型一致的变量。例如，“选择预测变量”列表仅包含预测模型中某项的变量，“选择结果”列表仅包含由模型中的其他某个变量预测的变量。选中**显示全部变量**框可查看两个列表中的所有模型变量。

隐含模型的协方差 显示或隐藏包含模型所隐含协方差矩阵的报表。

隐含模型的相关性 显示或隐藏包含模型所隐含相关性矩阵的报表。

隐含模型的均值 显示或隐藏包含模型所隐含每个变量均值的报表。

残差 显示或隐藏包含模型的残差矩阵的报表。该矩阵是隐含模型的协方差矩阵与样本协方差矩阵之间的差值。

标准化残差 显示或隐藏包含模型的标准化残差矩阵的报表。

RAM 矩阵 显示或隐藏包含网状动作模型 (RAM) 表示法所使用模型矩阵的报表。

估计值的协方差 显示或隐藏包含模型参数估计值的协方差矩阵的报表。

估计值的相关性 显示或隐藏包含模型参数估计值的相关性矩阵的报表。

内生变量的 R^2 （仅当模型为递归模型且包含内生变量时才可用。）显示或隐藏包含模型中每个内生变量的 R^2 值的报表。用 1 减去每个内生变量的残差方差与隐含模型的方差之比，即可计算出该值。R^2 值表示模型在内生变量中所解释的方差。内生变量是指在路径图中有一条指向它的路径的变量。

热图 支持您直观演示模型中的残差、协方差和相关性。

标准化残差热图 显示或隐藏包含模型的标准化残差热图的报表。

隐含模型的协方差热图 显示或隐藏一个报表，其中包含模型所隐含的协方差矩阵的热图。

隐含模型的相关性热图 显示或隐藏一个报表，其中包含模型所隐含的相关性矩阵的热图。

估计值的协方差热图 显示或隐藏一个报表，其中包含模型参数估计值的协方差矩阵的热图。

估计值的相关性热图 显示或隐藏一个报表，其中包含模型参数估计值的相关性矩阵的热图。

修改指标 支持您显示模型修改指标的全部或部分估计值。这些值可用于确定可以向模型添加哪些参数以改进模型拟合。每个表都按“卡方”列降序排序。

所有修改指标 显示或隐藏一个表，其中包含所有模型修改指标的估计值。该表包含指示每个估计值的参数类型的一列。

均值的修改指标 显示或隐藏一个表，其中包含均值和截距的模型修改指标的估计值。

载荷的修改指标 显示或隐藏一个表，其中包含载荷参数的模型修改指标的估计值。

回归的修改指标 显示或隐藏一个表，其中包含回归参数的模型修改指标的估计值。

协方差的修改指标 显示或隐藏一个表，其中包含协方差参数的模型修改指标的估计值。

评估测量模型 （仅可用于唯一因子之间没有协方差的确认性因子模型。）显示或隐藏用于量化关于检验和测度的可靠性和有效性的各种统计量和图形，包括指标可靠性、系数 omega 和 H 以及构造有效性矩阵。

“指标可靠性”图显示了潜在变量的平方标准化载荷以及可接受可靠性的建议最小阈值 (0.25)。变量的低值表示该变量在捕获相应潜在变量中的变异性方面表现欠佳。

“复合可靠性”报表和“构造最大可靠性”报表分别为每个潜在变量显示系数 Omega (McDonald 1999) 和 H (Hancock and Mueller 2001)。这些值的范围介于 0 到 1 之间，建议这些值约为 0.70 或更大。Omega 表示在观测到的复合得分中的潜在变量的方差比例。H 表示由指标表示的潜在变量方差的比例。这些估计值取决于模型；若拟合一个单因子模型，那么得到的 Omega 被称为一般 omega。若拟合具有一个以上潜在变量的因子模型，那么得到的 omega 估计值被称为分量 omega。不过，若拟合双因子模型，则一般因子估计值的 omega 称为分层 omega，而组因子称为分层分量 omega (Rodriguez et al. 2015)。建议的阈

值应在调查目标的上下文中使用；若您计划使用复合得分做出关于个体的决定，那么可靠性应该高于建议的阈值（大约 0.90 或更高），但若您计划将复合得分用于研究目的，那么阈值的下限是可以接受的 (Nunnally 1978)。

“构造有效性矩阵”报表有助于确定潜在变量是否在测量您认为它们在测量的内容：

- 下三角元素包含潜在变量相关性。这些元素支持您检查潜在变量之间的相关性有多强，并将其与假设的相关性强度进行比较。
- 上三角元素是潜在变量相关性的平方。这些元素让您能够关注潜在变量之间方差的重叠。在与矩阵中的对角线元素比较时，这些统计量尤其重要。
- 对角线元素包含每个潜在变量提取的平均方差量，相当于每个潜在变量的指标可靠性的平均值。一个好的潜在变量应该在对角线上有很高的值，因为它的指标有足够的系统方差来正确定义它。理想情况下，每个潜在变量的对角线元素应高于其上方和右侧的元素。

构造有效性矩阵的可视化支持您比较对角线元素和上三角元素。

请参见 “‘评估测量模型’报表的示例”。

预测值图 显示或隐藏模型中的内生变量预测值图。对于纵向数据，该图显示了模型隐含的随时间变化的增长轨迹。默认情况下，预测值显示为箱线图。选中**连接数据点**复选框可将显示切换为线图。

保存列 支持您将基于拟合结构化方程模型的列保存到数据表中。

保存因子得分 （仅当模型中有潜在变量时才可用。）将包含使用回归方法为每个潜在变量计算的因子得分的列保存到数据表中。因子得分在隐藏列中计算，该列也会添加到数据表中。该隐藏列使用 `Estimate Factor Score()` JSL 函数。有关该函数的详细信息，请参见“帮助”>“脚本索引”。

保存 Bartlett 因子得分 （仅当模型中有潜在变量时才可用。）将包含使用 Bartlett 方法为每个潜在变量计算的因子得分的列保存到数据表中。因子得分在隐藏列中计算，该列也会添加到数据表中。该隐藏列使用 `Estimate Bartlett Factor Score()` JSL 函数。有关该函数的详细信息，请参见“帮助”>“脚本索引”。

保存预测公式 （仅当模型中至少有一个内生或因变量时可用。）将包含每个变量观测结果预测值公式的列保存到数据表中。当模型中存在潜在变量时，使用 Bartlett 方法计算的因子得分也会保存到数据表中。

保存观测残差 （仅当模型中至少有一个内生或因变量时可用。）将包含每个变量观测结果残差值的列保存到数据表中。当模型中存在潜在变量时，使用 Bartlett 方法计算的因子得分也会保存到数据表中。

复制模型规格 将当前结构化方程模型规格复制到剪贴板。随后可以将模型规格粘贴到另一个 SEM 平台报表中。

在模型规格中重新调用 将 “模型规格”报表中的模型设置为指定的模型。

删除拟合 从报表窗口删除指定的模型报表。

JMP PRO 定制路径图

您可以采用多种方式定制结构化方程模型路径图。大多数定制选项位于以下一个或两个位置：路径图中的弹出式菜单和 “定制关系图外观” 窗口。

- “路径图弹出式菜单选项”
- “定制路径图的选项”

JMP PRO 路径图弹出式菜单选项

右击路径图本身时，会出现路径图弹出式菜单。菜单中可用的一组特定选项取决于在关系图中右击的位置，以及右击时是否选择了路径图的任何元素。

选择 包含在 “从列表” 和 “至列表” 中进行选择、向关系图中添加单向和双向箭头、创建新的潜在变量、隐藏选定项以及显示所有隐藏项等选项。在路径图中选择项之前，该子菜单中的许多选项都未启用。

提示：当路径图中的项被隐藏时，路径图顶部会显示一个 “全部取消隐藏” 按钮，该按钮支持您显示所有隐藏的项。

显示 包含显示或隐藏路径图的各种元素的选项。选择**显示默认值**可返回最初显示的一组元素。

注意：“显示默认值” 选项不影响 “显示均值 / 截距” 选项的设置。

对于拟合模型报表中的路径图，该子菜单还包含用于更改路径图中箭头上显示的估计值的选项。您可以在显示非标准化参数估计值、标准化参数估计值或不显示估计值之间进行选择。

选择该类型的全部项 （仅当右击路径图中的某项时才可用。）选择路径图中与右击的项类型相同的其他所有项。

选择潜在变量组 （仅当右击路径图中的潜在变量时才可用。）选择与所选潜在变量关联的整个项组。若未选择任何潜在变量，则该选项仅选择与右击的潜在变量关联的项组。

重命名变量 （仅当右击路径图中的潜在变量时才可用。 ）支持您更改路径图中潜在变量的名称。

注意：在路径图中重命名潜在变量时，将在 “从列表” 和 “至列表” 中更新该名称。

添加回归 （仅当右击路径图中的某个变量时才可用。）支持您添加单向箭头，该箭头表示右击的变量与路径图中的另一个变量之间的回归。选择该选项后，点击您希望箭头指向的变量。

添加协方差 （仅当右击路径图中的某个变量时才可用。）支持您添加双向箭头，该箭头表示右击的变量与路径图中的另一个变量之间的协方差。选择该选项后，点击您希望箭头指向的变量。

定制关系图 启动“定制关系图外观”窗口。请参见“定制路径图的选项”。

旋转潜在变量组 以 90 度角为增量顺时针旋转路径图中选定的潜在变量的潜在变量指标。若未选定任何潜在变量，该选项将旋转路径图中的所有潜在变量指标。

布局 支持您为路径图中的项选择两个排列选项之一。

复制关系图 将路径图的图像保存到剪贴板。要保持尽可能高的质量，请将剪贴板图像粘贴为向量图形。

复制关系图属性 将当前路径图属性复制到剪贴板。随后可将属性粘贴到另一个 SEM 路径图中。

粘贴关系图属性 将剪贴板中的路径图属性粘贴到当前 SEM 路径图中。

撤销 撤销对路径图上次进行的更改。

恢复 恢复对路径图上次撤销的更改。

编辑 包含用于 JMP 中图形的标准选项。有关这些选项的详细信息，请参见 《使用 JMP》。

重置布局 将路径图重置为原始设置。

JMP PRO 定制路径图的选项

点击“关系图”面板中的“定制”按钮，从红色小三角菜单中选择“定制关系图”，或从路径图弹出式菜单中选择“定制关系图”时，将显示“定制关系图外观”窗口。“定制关系图外观”窗口包含四个面板和一个下拉式菜单，该菜单支持您从两个预设颜色主题中进行选择。

主题预设 包含用于预设颜色主题的选项。您可以在“黑色和白色”主题或“蓝色”主题之间进行选择。选择其中一个可快速更新设置以遵循任一主题。

变量外观 包含用于设置路径图中的填充颜色、边框颜色、文本颜色和项的大小的选项。您可以设置显变量、潜在变量和常量变量的外观，以及这其中每个项的字体。

路径设置 包含用于箭头外观的选项。可以将粗细和透明度设置为固定值，或使其基于每个路径箭头的标准化估计值来设置粗细和透明度。也可以用虚线表示不显著的 p 值，将 α 水平设置为不显著，并更改与箭头关联的颜色和字体。默认情况下，箭头的粗细是固定的，箭头的透明度与标准化参数估计值成正比。

其他设置 包含以下选项：

启用网格 显示或隐藏路径图中的网格，该网格可用于帮助对齐路径图中的各项。

锁定关系图 锁定或解除锁定路径图中各项的位置。当路径图被锁定时，您不能四处拖动项，并且路径图顶部的锁定指示符将突出显示。

显示均值/截距 显示或隐藏路径图中的均值和截距。默认情况下，均值结构不会显示在路径图中。

显示潜在指示符 显示或隐藏路径图中的潜在变量指标。

显示回归 显示或隐藏路径图中的回归箭头。

显示方差 显示或隐藏路径图中的方差箭头。

显示协方差 显示或隐藏路径图中的协方差箭头。

显示等式约束 显示或隐藏路径图中的等式约束。

显示 R 方值 显示或隐藏路径图中每个节点的 R^2 值。

显示常数均方 显示或隐藏路径图中的常数箭头。

用 R 方填充节点 显示或隐藏路径图中每个节点对应于每个节点的 R^2 值的填充颜色。

预览 包含基于窗口中当前设置的路径图外观预览。

提示：使用“定制关系图外观”窗口底部的**保存至首选项**按钮将窗口中的当前设置保存到平台首选项，以便将来的路径图使用相同的设置。

JMP PRO 结构化方程模型的更多示例

本节包含使用 “结构化方程模型”平台的示例：

- “潜在路径变量模型的示例”
- “潜在变量增长曲线模型的示例”
- ““评估测量模型”报表的示例”
- “多组分析示例”

潜在路径变量模型的示例

在本例中，您要为 Bollen (1989) 中所述的工业化和政治民主构建一个结构化回归模型，文中使用来自 75 个发展中国家 / 地区的数据。数据表中的变量包括 19 判别主成分 0 年和 19 判别主成分 5 年的四个民主测度和 19 判别主成分 0 年的三个工业化测度。这些变量在数据表每一列中的“注释”列属性中说明。要查看 “注释”列属性，请右击列名，选择 “列信息”，然后在 “列属性”下选择 “注释”。

模型指定过程包含四个主要步骤：创建潜在变量、添加载荷和回归变量、添加协方差项以及对载荷变量设置约束。

1. 选择**帮助 > 样本数据文件夹**，然后打开 Political Democracy.jmp。
2. 选择**分析 > 多元方法 > 结构化方程模型**。
3. 从 Prod 判别主成分 0 一直选择到 Legis 判别主成分 5 并点击**模型变量**。
4. 点击**确定**。

   “结构化方程模型”报表 “模型规格”分级显示项随即显示。
5. 点击 “视图”面板框中**列表**选项卡。

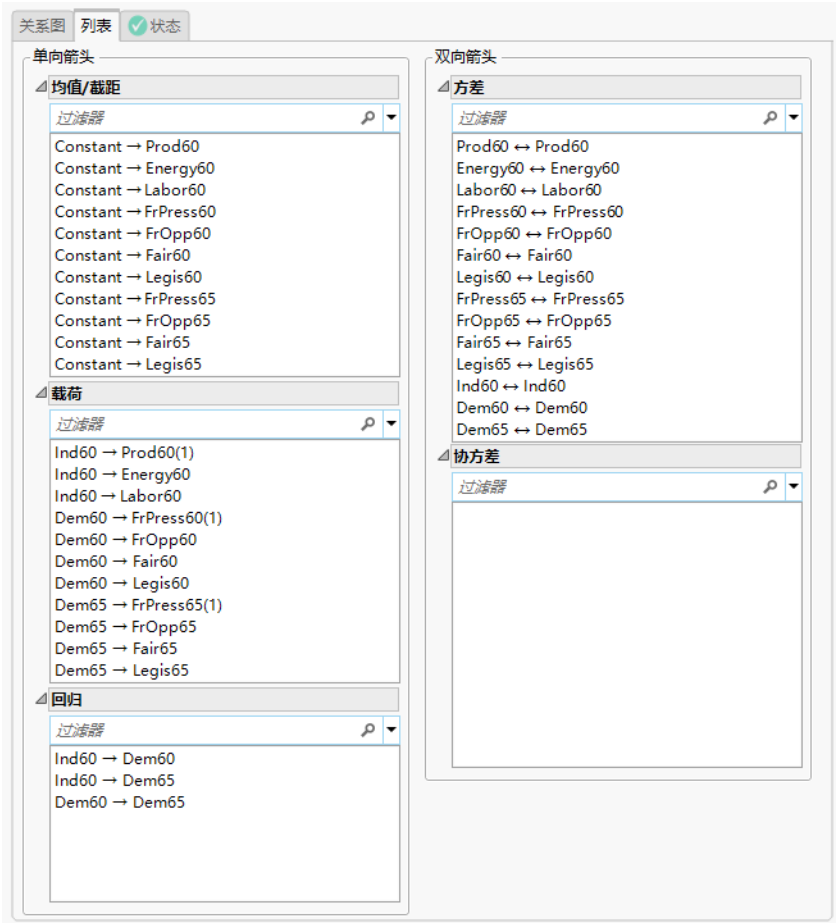
创建潜在变量

6. 在“至列表”中从 Prod 判别主成分 0 一直选择到 Labor 判别主成分 0，在“至列表”下方的框中键入“Ind 判别主成分 0”，然后点击添加潜在变量 按钮。
7. 在“至列表”中从 FrPress 判别主成分 0 一直选择到 Legis 判别主成分 0，在“至列表”下方的框中键入“Dem 判别主成分 0”，然后点击添加潜在变量 按钮。
8. 在“至列表”中从 FrPress 判别主成分 5 一直选择到 Legis 判别主成分 5，在“至列表”下方的框中键入“Dem 判别主成分 5”，然后点击添加潜在变量 按钮。

添加载荷和回归变量

9. 在“自列表”中选择 Ind 判别主成分 0，在“至列表”中选择 Dem 判别主成分 0，然后点击单向箭头 按钮。
10. 在“自列表”中选择 Ind 判别主成分 0，在“至列表”中选择 Dem 判别主成分 5，然后点击单向箭头 按钮。
11. 在“自列表”中选择 Dem 判别主成分 0，在“至列表”中选择 Dem 判别主成分 5，然后点击单向箭头 按钮。

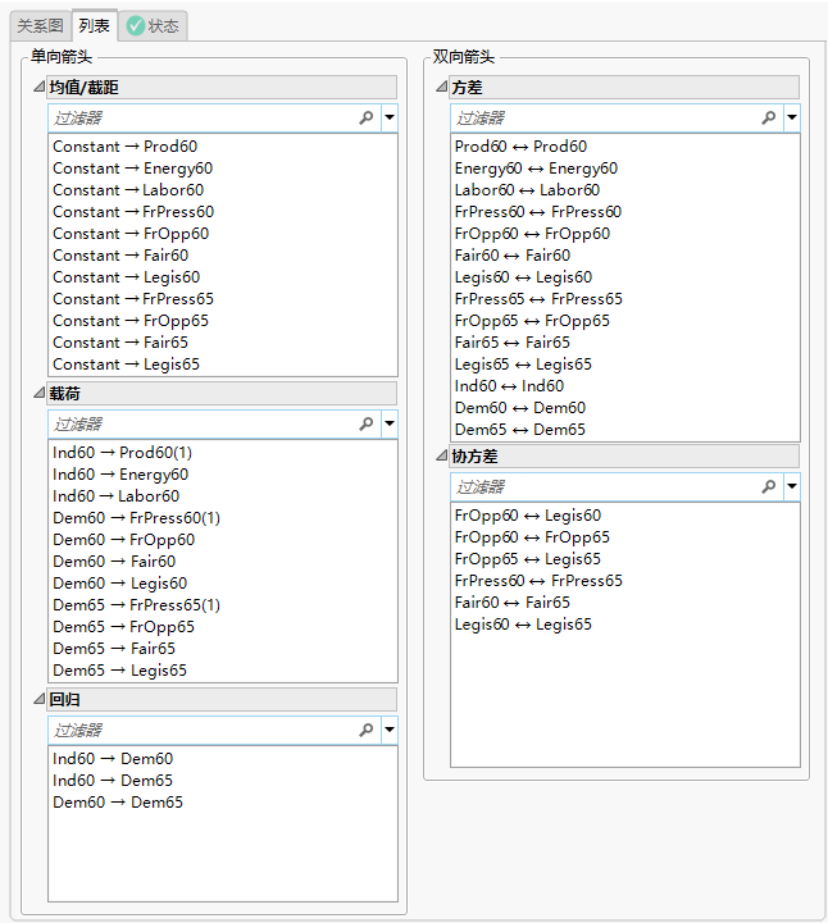
图 8.6 载荷和回归



添加协方差

12. 在“自列表”中选择 FrOpp 判别主成分 0，在“至列表”中选择 Legis 判别主成分 0 和 FrOpp 判别主成分 5，然后点击双向箭头 ↔ 按钮。
13. 在“自列表”中选择 FrOpp 判别主成分 5，在“至列表”中选择 Legis 判别主成分 5，然后点击双向箭头 ↔ 按钮。
14. 在“自列表”中选择 FrPress 判别主成分 0，在“至列表”中选择 FrPress 判别主成分 5，然后点击双向箭头 ↔ 按钮。
15. 在“自列表”中选择 Fair 判别主成分 0，在“至列表”中选择 Fair 判别主成分 5，然后点击双向箭头 ↔ 按钮。
16. 在“自列表”中选择 Legis 判别主成分 0，在“至列表”中选择 Legis 判别主成分 5，然后点击双向箭头 ↔ 按钮。

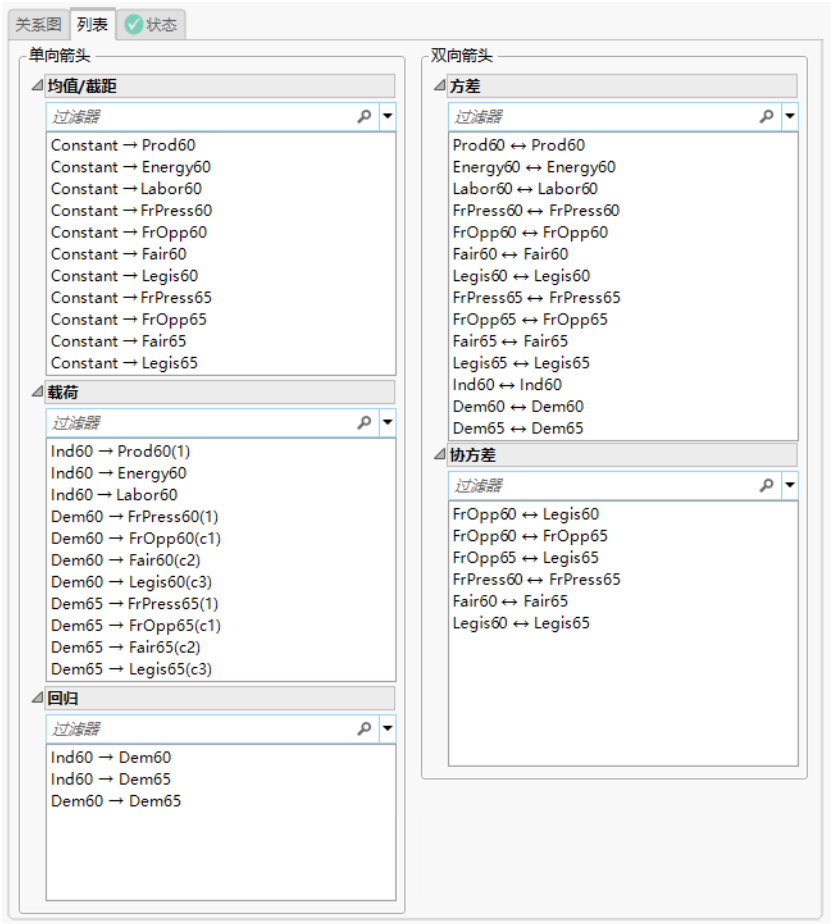
图 8.7 协方差



添加针对载荷的约束

17. 在“载荷”列表中选择 Dem判别主成分0->FrOpp判别主成分0 和 Dem判别主成分5->FrOpp判别主成分 5，然后点击**设置等于**。
18. 在 “载荷”列表中选择 Dem 判别主成分 0->Fair 判别主成分 0 和 Dem 判别主成分 5->Fair判别主成分 5，然后点击**设置等于**。
19. 在“载荷”列表中选择 Dem 判别主成分 0->Legis 判别主成分 0 和 Dem 判别主成分 5->Legis判别主成分 5，然后点击**设置等于**。

图 8.8 完成的模型规格



使用字母数字标签指定针对载荷的约束。例如，您可以看到 Dem 判别主成分 0->FrOpp 判别主成分 0 和 Dem 判别主成分 5-> FrOpp 判别主成分 5 被设置为相等，因为它们都被标记为 “c1”。

20. 在 “模型名称” 下方的文本框中，键入 “工业化和政治民主”。
21. 点击**运行**。

图 8.9 结构化方程模型 “拟合汇总”报表

拟合汇总

最大似然, 梯度收敛.

样本大小	75
带缺失值的行	0
-2 对数似然	3097.6362
迭代	7
参数数目	39
AICc	3264.779
BICu	-123.8851
卡方	40.17949
自由度	38
概率>卡方	0.3738824
CFI	0.9967743
RMSEA	0.0276538
90% 下限	0
90% 上限	0.0870678

“拟合汇总”报表中列出的该模型的卡方统计量为 40.18，自由度为 38。请注意，相应的 p 值为 0.3739，该值不显著。这表明，没有证据可以拒绝模型拟合良好这一原假设。因此，您可以得出结论：该模型对数据拟合良好。

卡方值取决于样本大小，因此，一些拟合良好的模型仍然可以生成显著的卡方值。比较拟合指数 (CFI) 和近似的均方根误差 (RMSEA) 为确定模型拟合提供了额外的指导。这些指数介于 0 和 1 之间。 CFI 值最好大于 0.90， RMSEA 值最好小于 0.10（Browne and Cudeck 1993 ； Hu and Bentler 1999）。在此， CFI 为 0.99 判别主成分 8， RMSEA 为 0.0277，这指示拟合极佳。

图 8.10 结构化方程模型 “参数估计值”报表

结构化方程模型: Industrialization and Political Democracy

参数估计值

均值/截距	估计值	标准误差	Wald Z	概率>|Z|
Constant → Prod60	5.0543838	0.0840624	60.126553	<.0001*
Constant → Energy60	4.7921946	0.1732697	27.657433	<.0001*
Constant → Labor60	3.5576898	0.1612318	22.065683	<.0001*
Constant → FrPress60	5.4646667	0.2989108	18.28193	<.0001*
Constant → FrOpp60	4.2564429	0.43899	9.6959907	<.0001*
Constant → Fair60	6.5631103	0.3939806	16.658459	<.0001*
Constant → Legis60	4.452533	0.37963	11.728613	<.0001*
Constant → FrPress65	5.1362519	0.304444	16.870925	<.0001*
Constant → FrOpp65	2.9780741	0.3923495	7.590361	<.0001*
Constant → Fair65	6.1962639	0.3643986	17.004083	<.0001*
Constant → Legis65	4.0433897	0.3753514	10.772279	<.0001*
载荷	**估计值**	**标准误差**	**Wald Z**	**概率>|Z|**
Ind60 → Prod60	1	.	.	.
Ind60 → Energy60	2.1796563	0.1389144	15.690643	<.0001*
Ind60 → Labor60	1.8182091	0.1521279	11.951848	<.0001*
Dem60 → FrPress60	1	.	.	.
Dem60 → FrOpp60	1.1907892	0.1416452	8.4068422	<.0001*
Dem60 → Fair60	1.1745429	0.1197994	9.8042513	<.0001*
Dem60 → Legis60	1.2509852	0.1229265	10.17669	<.0001*
Dem65 → FrPress65	1	.	.	.
Dem65 → FrOpp65	1.1907892	0.1416452	8.4068422	<.0001*
Dem65 → Fair65	1.1745429	0.1197994	9.8042513	<.0001*
Dem65 → Legis65	1.2509852	0.1229265	10.17669	<.0001*
回归	**估计值**	**标准误差**	**Wald Z**	**概率>|Z|**
Ind60 → Dem60	1.4713298	0.3914705	3.7584694	0.0002*
Ind60 → Dem65	0.6004651	0.2382793	2.520005	0.0117*
Dem60 → Dem65	0.8650429	0.0756829	11.429833	<.0001*
方差	**估计值**	**标准误差**	**Wald Z**	**概率>|Z|**
Prod60 ↔ Prod60	0.0813876	0.0196996	4.1314429	<.0001*
Energy60 ↔ Energy60	0.1204279	0.0699023	1.7228042	0.0849
Labor60 ↔ Labor60	0.4666598	0.0891232	5.2361204	<.0001*
FrPress60 ↔ FrPress60	1.8546635	0.4569832	4.0584942	<.0001*
FrOpp60 ↔ FrOpp60	7.5813057	1.3449558	5.6368437	<.0001*
Fair60 ↔ Fair60	4.9556824	0.9612174	5.1556312	<.0001*
Legis60 ↔ Legis60	3.2244586	0.7417042	4.3473645	<.0001*
FrPress65 ↔ FrPress65	2.3130445	0.4834091	4.7848595	<.0001*
FrOpp65 ↔ FrOpp65	4.968181	0.8945069	5.5541001	<.0001*
Fair65 ↔ Fair65	3.5600424	0.7379151	4.8244607	<.0001*
Legis65 ↔ Legis65	3.3076929	0.7128664	4.63999	<.0001*
Ind60 ↔ Ind60	0.4485992	0.0867473	5.1713357	<.0001*
Dem60 ↔ Dem60	3.8752798	0.888605	4.3610824	<.0001*
Dem65 ↔ Dem65	0.1644203	0.2333327	0.7046604	0.4810
协方差	**估计值**	**标准误差**	**Wald Z**	**概率>|Z|**
FrOpp60 ↔ Legis60	1.4400946	0.6909749	2.084149	0.0371*
FrOpp60 ↔ FrOpp65	2.1830127	0.731106	2.9859044	0.0028*
FrOpp65 ↔ Legis65	1.3717826	0.5781705	2.3726262	0.0177*
FrPress60 ↔ FrPress65	0.5825441	0.3644288	1.5985127	0.1099
Fair60 ↔ Fair65	0.711572	0.6194126	1.1487851	0.2506
Legis60 ↔ Legis65	0.3628	0.4607878	0.7873472	0.4311

接下来，“回归”下的参数估计值表明 Ind 判别主成分 0 对 Dem 判别主成分 0 和 Dem 判别主成分 5 具有正效应， Dem 判别主成分 0 对 Dem 判别主成分 5 也具有正效应。因此， Ind 判别主成分 0 的较高得分与 Dem 判别主成分 0 和 Dem 判别主成分 5 较高相关， Dem 判别主成分 0 的较高得分与 Dem 判别主成分 5 的较高得分相关。参数估计值相应的 p 值显示在“回归”下方。 3 个回归参数在 $\alpha = 0.05$ 水平下均显著。因此，您可以得出结论，潜在变量之间存在非零关系。

潜在变量增长曲线模型的示例

潜在变量增长曲线 (LGC) 模型支持分析人员对纵向数据进行建模，并描述随时间变化的轨迹特征。最常见的潜在变量增长曲线模型是线性增长模型，其中指定了截距和斜率潜在变量，以捕获过程的总体轨迹以及单个观测与总体轨迹的偏差。该模型非常类似于随机系数模型。LGC 模型的一个关键特性是它们对数据的均值结构施加约束。对均值结构建模支持您检验与增长相关的假设。

在本例中，您使用多项选择测试得分对参与一项学习计划 4 年以上的学生的成绩得分进行建模。您想要检验以下假设，即：线性轨迹可描述该过程的特征。

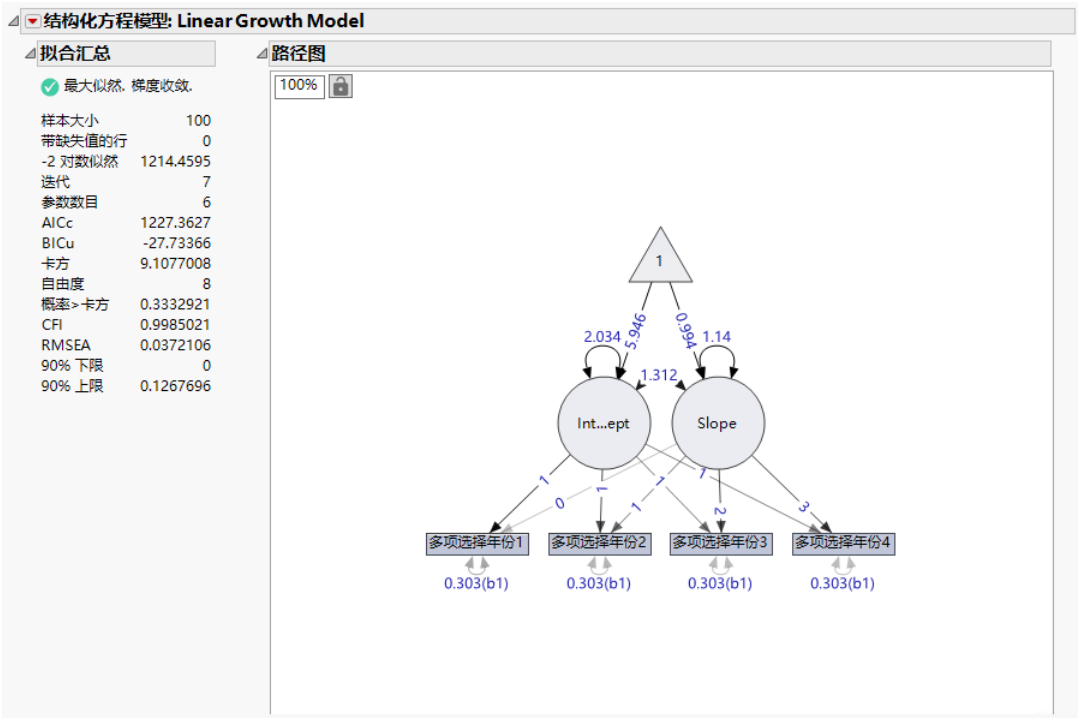
1. 选择**帮助 > 样本数据文件夹**，然后打开 Academic Achievement.jmp。
2. 选择**分析 > 多元方法 > 结构化方程模型**。
3. 从**多项选择年份 1** 一直选择到**多项选择年份 4**，然后点击**模型变量**。
4. 点击**确定**。

   “结构化方程模型”报表 “模型规格”分级显示项随即显示。
5. 选择**模型快捷方式 > 纵向分析 > 线性潜在变量增长曲线**。

   “模型规格”关系图现在显示线性潜在变量增长曲线模型。
6. 点击 “视图”面板框中**列表**选项卡。
7. 选择 “方差”列表中的前 4 项，然后点击**设置等于**。

   这会将每个变量的残差方差限制为相等，这类似于方差齐性假设。在 SEM 中，该假设是可检验的，这样另一个模型可以在没有等式约束的情况下进行拟合，并与卡方差异检验进行比较。
8. 在 “模型名称”下方的文本框中，键入 “方差相等的线性 LGC”。
9. 点击**运行**。

图 8.11 线性 LGC 模型的拟合汇总和路径图



“拟合汇总”报表中列出的该模型的卡方统计量为 9.11，自由度为 8。请注意，相应的 p 值为 0.3333，该值不显著。这表明，没有证据可以拒绝模型拟合良好这一原假设。因此，您可以得出结论：该模型对数据拟合良好。CFI 和 RMSEA 拟合指数也指示拟合极佳，因为它们分别大于 0.9 和小于 0.1。

图 8.12 线性 LGC 模型的参数估计值

结构化方程模型: Linear Growth Model

参数估计值

均值/截距	估计值	标准误差	Wald Z	概率>|Z|
Constant → Intercept	5.9461	0.1498668	39.675887	<.0001*
Constant → Slope	0.9942	0.1095503	9.075282	<.0001*
载荷	**估计值**	**标准误差**	**Wald Z**	**概率>|Z|**
Intercept → 多项选择年份1	1	.	.	.
Intercept → 多项选择年份2	1	.	.	.
Intercept → 多项选择年份3	1	.	.	.
Intercept → 多项选择年份4	1	.	.	.
Slope → 多项选择年份1	0	.	.	.
Slope → 多项选择年份2	1	.	.	.
Slope → 多项选择年份3	2	.	.	.
Slope → 多项选择年份4	3	.	.	.
方差	**估计值**	**标准误差**	**Wald Z**	**概率>|Z|**
多项选择年份1 ↔ 多项选择年份1	0.3030585	0.0303059	10	<.0001*
多项选择年份2 ↔ 多项选择年份2	0.3030585	0.0303059	10	<.0001*
多项选择年份3 ↔ 多项选择年份3	0.3030585	0.0303059	10	<.0001*
多项选择年份4 ↔ 多项选择年份4	0.3030585	0.0303059	10	<.0001*
Intercept ↔ Intercept	2.0338661	0.318341	6.3889543	<.0001*
Slope ↔ Slope	1.1395155	0.1698318	6.7096706	<.0001*
协方差	**估计值**	**标准误差**	**Wald Z**	**概率>|Z|**
Intercept ↔ Slope	1.3124985	0.2048419	6.4073734	<.0001*

截距的均值估计值为 5.95，这指示第一年的总体成绩得分。这意味着，在第一次测量时，学生的学习成绩平均为 5.9 判别主成分。截距的均值以第一次为中心，因为斜率因子在该变量上的载荷为零。

斜率的均值估计值为 0.99，这表明总体成绩得分每年增加 0.99。截距和斜率因子的显著方差估计值指向平均轨迹周围的显著变异性；并不是每个人的起点都一样，每个人的增长速度也都不尽相同。外部变量可以用作截距和斜率的预测变量，以了解导致个体轨迹差异的原因。

最后，截距和斜率之间的正协方差表明，那些第一年学习成绩较高的学生往往会随着时间的推移成绩进步更快。

LGC 模型为调查过程如何随时间变化提供了一种灵活的方法。“模型快捷方式”菜单提供了可相互拟合和检验的备选轨迹，以确定最佳模型拟合，包括无增长轨迹、二次轨迹，甚至是非线性轨迹（使用“潜在基函数”选项）。

“评估测量模型”报表的示例

确认性因子分析 (CFA) 模型支持您检验备选测量模型。“评估测量模型”报表提供用于量化检验和测度的可靠性和有效性的工具。结果包括指标可靠性、系数 Omega 和 H 以及构造有效性矩阵。

在本例中，您打算评估消费者数据调查的有效性和可靠性。您需要拟合以下五个潜在变量的确认性因子分析模型：隐私、安全、声誉、信任和购买意向。随后使用 “评估测量模型”选项检验调查可靠性。

1. 选择**帮助 > 样本数据文件夹**，然后打开 Online Consumer Data.jmp。
2. 点击 SEM: CFA 脚本旁边的绿色小三角。

   该脚本针对调查数据运行确认性因子分析模型。
3. 点击 “模型规格”和 “模型比较”旁边的灰色展开图标可隐藏这些报表部分。

图 8.13 CFA 模型的拟合汇总

拟合汇总

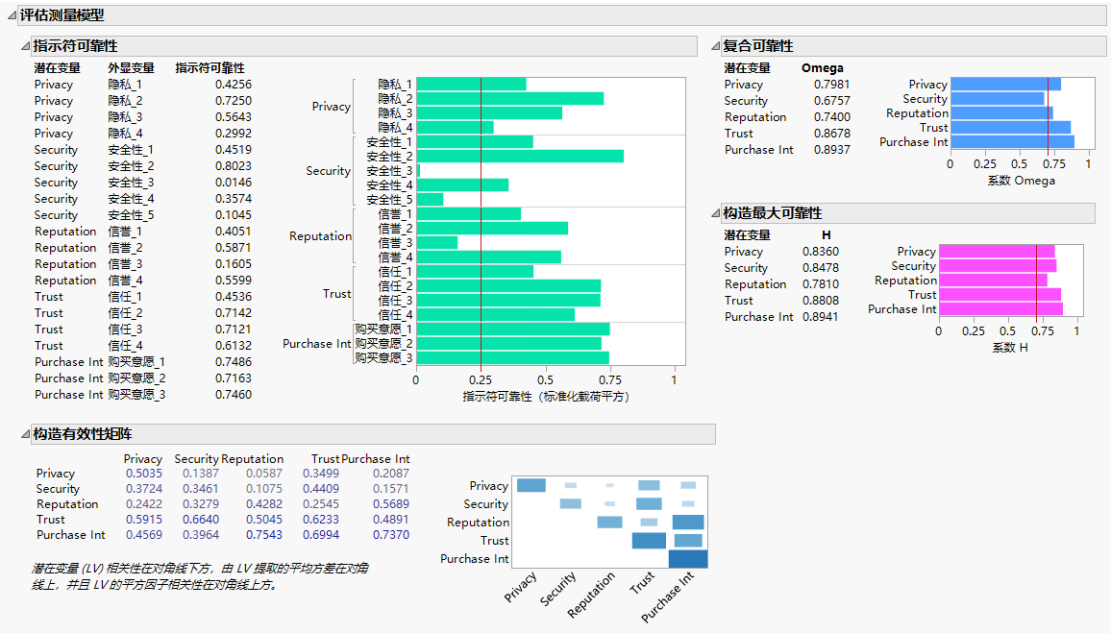
最大似然, 梯度收敛.

样本大小	843
带缺失值的行	0
-2 对数似然	45106.777
迭代	6
参数数目	70
AICc	45259.653
BICu	-269.8055
卡方	808.10917
自由度	160
概率>卡方	3.738e-87
CFI	0.9160319
RMSEA	0.0693187
90% 下限	0.0646117
90% 上限	0.0741036

“拟合汇总”报表中列出的该模型的卡方统计量为 808.11，自由度为 1 判别主成分 0。请注意相应的 p 值显著。这表明，有一些证据可以拒绝模型拟合良好这一原假设。不过，卡方统计量受样本大小影响极大，即使模型提供了对 200 到 300 个观测样本数据的良好拟合，卡方统计量也被证明是显著的。在本例中，样本大小为 843。因此，还应使用 CFI 和 RMSEA 拟合指数对模型拟合进行评估。在本例中，这两个指数一个大于 0.9，一个小于 0.1，都指示拟合良好。

4. 点击 “结构化方程模型 : CFA”旁边的红色小三角，然后选择**评估测量模型**。

图 8.14 CFA 模型的 “评估测量模型” 报表



“指标可靠性” 图显示了潜在变量的平方标准化载荷以及可接受可靠性的建议最小阈值 (0.25)。您可以看到，两个与安全相关的问题和一个与声誉相关的问题在捕获相应潜在变量的变异性方面做得不太好。“安全 _3” 问题的值很低，因此可以删除它，而不会对构造的可靠性造成任何损失；而 “安全 _5” 和 “声誉 _3” 这两个问题可以进行修改，以提高其可靠性。

“复合可靠性” 报表和 “构造最大可靠性” 报表分别显示每个潜在变量的系数 Omega 和 H。这些值的范围介于 0 到 1 之间，建议这些值约为 0.70 或更大。Omega 表示在观测到的复合得分中的潜在变量的方差比例。H 表示由指标表示的潜在变量方差的比例。通过这些测度，“安全” 方面的 Omega 系数略低于建议的阈值 0.70，但它足够接近，可以认为该复合对于研究目的是可靠的。建议的阈值应在调查目标的上下文中使用；若您计划使用复合得分做出关于个体的决定，那么可靠性应该高于建议的阈值（大约 0.90 或更高），但若您计划将复合得分用于研究目的，那么阈值的下限是可以接受的 (Nunnally 1978)。“安全” 和 “声誉” 的复合可靠性可以通过重点提高 “安全 _3”、“安全 _5” 和 “声誉 _4” 等问题的指标可靠性来提高。

“构造有效性矩阵” 报表提供了一种方式，用于确定潜在变量是否在测量您认为它们在测量的内容：请注意，在矩阵可视化视图中，“隐私”、“信任” 和 “购买意向” 的对角线值大于它们上方和右侧的值。不过，对于 “安全” 和 “声誉” 却不是这样。这进一步证明，可以改进 “安全” 相关和 “声誉” 相关的问题以衡量 “安全” 和 “声誉” 潜在变量。

您可以得出结论：可以通过删除或修改一些与 “安全” 和 “声誉” 相关的问题来改进调查。

有关报表组成部分的详细信息，请参见 “评估测量模型”。

多组分析示例

“结构化方程模型”平台中的 “组”变量角色支持您对多个组的统计效应进行模型比较。在本例中，您关注的是比较两组儿童的学习成绩增长轨迹。您有关于学生在多项选择测试中的分数的四个重复测度。

1. 选择**帮助 > 样本数据文件夹**，然后打开 Academic Achievement.jmp。
2. 选择**分析 > 多元方法 > 结构化方程模型**。
3. 从**多项选择年份 1** 一直选择到**多项选择年份 4**，然后点击**模型变量**。
4. 选择**性别**并点击**组**。
5. 点击**确定**。

   “结构化方程模型”报表 “模型规格”分级显示项随即显示。
6. 选择**模型快捷方式 > 纵向分析 > 线性潜在变量增长曲线**。

   “模型规格”关系图现在显示线性潜在变量增长曲线模型。
7. 在 “模型名称”下方的文本框中，键入 “分组线性 LGC 模型 （等式约束）”。
8. 在路径图中选择三角形图标与 “截距”和 “斜率”之间的箭头，然后点击**设置等于**。

   这将强制两组学生的 “截距”和 “斜率”潜在变量的均值相等。
9. 在 “设置等于”窗口中，点击**确定**在两组之间应用等式约束。
10. 点击**运行**。
11. 向上滚动至报表窗口的 “模型规格”部分。
12. 在 “模型名称”下方的文本框中，键入 “分组线性 LGC 模型”。
13. 选择路径图中标记为 a1 和 a2 的箭头，然后点击**自由**。
14. 点击**运行**。
15. 点击 “模型规格”旁边的灰色展开图标。
16. 在 “模型比较”表中，同时选定两个分组 LGC 模型并点击**比较选定模型**。

图 8.15 “模型比较”报表

模型比较

	模型名称	-2 对数似然	参数数目	AICc	AICc 权重	.2 .4 .6 .8	BICu	卡方	自由度	概率>卡方	CFI	RMSEA	90% 下限	90% 上限
1	不受限（饱和）	1170.3253	28	1249.1985	0.0000		0.0000	0.0000	0	.	1.0000	0.0000	0.0000	0.0000
2	独立	1852.9631	16	1891.5173	0.0000		627.3758	682.6378	12	<.0001*	0.0000	0.7476	0.7004	0.7958
3	Grouped Linear LGC Model (Equality Constraint)	1204.3683	16	1242.9225	0.0005		-21.2190	34.0430	12	0.0007*	0.9671	0.1917	0.1176	0.2690
4	Grouped Linear LGC Model	1183.2528	18	1227.6972	0.9995		-33.1242	12.9275	10	0.2278	0.9956	0.0765	0.0000	0.1812

比较选定模型 清除选择

卡方差异检验

模型嵌套...	...在模型中	Δ卡方	Δ自由度	概率>卡方	ΔCFI	ΔRMSEA
Grouped Linear LGC Model (Equality Constraint)	Grouped Linear LGC Model	21.1155	2	<.0001*	-0.029	0.1152

差异检验仅对嵌套模型有意义

卡方差异检验表明，等式约束导致模型失拟的统计上的显著性增加，因此是不合适的。与此相关，两个分组 LGC 模型的 BICu 值均为负值，这表明两个模型都倾向于不受限模型。您

选择继续使用无约束分组 LGC 模型，因为它受到卡方差异检验的支持，并且具有较小的 BICu 值。

17. 点击 “分组线性 LGC 模型 （等式约束）”旁边的 👁 按钮。

    这将隐藏报表中与具有等式约束的模型相关的部分。

18. 按 Alt 键并点击 “结构化方程模型 : 分组线性 LGC 模型”旁边的红色小三角。

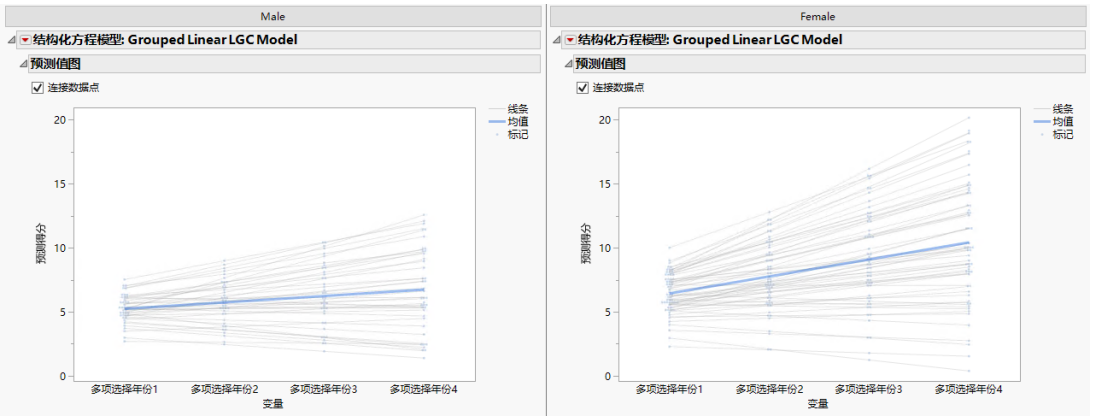
提示：Alt 键支持您同时选定多个红色小三角菜单选项。

19. 取消选择**显示路径图**、**拟合汇总**和**参数估计值**等选项。
20. 选择**预测值图**选项。
21. 点击**确定**。

注意：“结构化方程模型 : 分组线性 LGC 模型”红色小三角菜单中的选择将应用于 “男性”和 “女性”选项卡报表。

22. 右击 “男性”或 “女性”选项卡，然后选择**设置样式 > 水平展开**。
23. 选中**连接数据点**框。

图 8.16 预测值图



在过去 4 年中，与女性的预测值相比，男性的预测值具有更平缓的斜率和更小的变异。

JMP PRO “结构化方程模型”平台的统计详细信息

本节包含 “结构化方程模型”平台的统计详细信息。

- “估计方法”
- “拟合测度汇总”

估计方法

“结构化方程模型”平台中使用的默认估计方法取决于是否存在缺失值。若无缺失数据，“结构化方程模型”平台默认使用最大似然 (ML) 估计。若检测到任何缺失数据，则平台默认使用全信息最大似然 (FIML) 估计。在这两种情况下，都使用观测的信息矩阵获取标准误差。该平台在评估过程中组合使用多种优化算法，包括 Newton-Raphson、Quasi-Newton 和 Fisher 得分。

注意：“结构化方程模型”平台中的 ML 方法在估计中使用样本大小 N 而不是 N - 1。

拟合测度汇总

本节介绍 “结构化方程模型”平台中报告的拟合测度的汇总。

AICc、BIC 和 BICu

AICc 和 BIC 定义如下：

$$\text{AICc} = -2\text{logL} + 2k + \frac{2k(k+1)}{n-(k+1)}$$

$$\text{BIC} = -2\text{logL} + k\ln(n)$$

其中：

-2logL 是负对数似然的两倍。

n 是样本大小。

k 是参数个数。

有关“模型比较”报表中基于似然的测度的详细信息，请参见《拟合线性模型》。

与不受限模型 (BICu) 相关的 BIC 定义如下：

$$\text{BICu} = \chi^2_{min} - df_{min}\log(n)$$

其中：

χ^2_{min} 是拟合模型的卡方统计量。

df_{min} 是拟合模型的自由度。

n 是样本大小。

拟合模型的 BICu 等价于拟合模型的 BIC 减去不受限模型的 BIC。有关与不受限模型相关的 BIC 的详细信息，请参见 Bollen et al.(2014)。

CFI

比较拟合指数 (CFI) 定义如下：

$$\text{CFI} = \frac{max(\chi_0^2 - df_0, 0) - max(\chi_{min}^2 - df_{min}, 0)}{max(\chi_0^2 - df_0, 0)}$$

其中：

χ_0^2 是独立模型的卡方统计量。

df_0 是独立模型的自由度。

χ_{min}^2 是拟合模型的卡方统计量。

df_{min} 是拟合模型的自由度。

有关 CFI 的详细信息，请参见 Bentler (1990)。

RNI

相对非中心指数 (RNI) 定义如下：

$$\text{RNI} = \frac{(\chi_0^2 - df_0) - (\chi_{min}^2 - df_{min})}{\chi_0^2 - df_0}$$

其中：

χ_0^2 是独立模型的卡方统计量。

df_0 是独立模型的自由度。

χ_{min}^2 是拟合模型的卡方统计量。

df_{min} 是拟合模型的自由度。

有关 RNI 的详细信息，请参见 McDonald and Marsh (1990)。

TLI

Tucker-Lewis 指数 (TLI) 定义如下：

$$\text{TLI} = \frac{\frac{\chi_0^2}{df_0} - \frac{\chi_{min}^2}{df_{min}}}{\frac{\chi_0^2}{df_0} - 1}$$

其中：

χ_0^2 是独立模型的卡方统计量。

df_0 是独立模型的自由度。

χ_{min}^2 是拟合模型的卡方统计量。

df_{min} 是拟合模型的自由度。

详细信息，请参见 West et al.(2012)。

NFI

Bentler-Bonett 规范拟合指数 (NFI) 定义如下：

$$\text{NFI} = \frac{\chi_0^2 - \chi_{min}^2}{\chi_0^2}$$

其中：

χ_0^2 是独立模型的卡方统计量。

χ_{min}^2 是拟合模型的卡方统计量。

详细信息，请参见 West et al.(2012)。

修正的 GFI 和修正的 AGFI

修正的拟合优度指数 （修正的 GFI）定义如下：

$$\text{修正的 GFI} = \frac{p}{p + 2\left(\frac{\chi_{min}^2 - df_{min}}{n-1}\right)}$$

其中：

χ^2_{min} 是拟合模型的卡方统计量。

$df_{\min}$ 是拟合模型的自由度。

p 是拟合模型的观测变量数。

n 是样本大小。

修正的拟合优度指数 （修正的 AGFI）定义如下：

$$\text{修正的 AGFI} = 1 - \frac{p^*}{df_{min}}(1 - \text{修正的} GFI)$$

其中：

p^* 是协方差矩阵中唯一条目数和观测变量的均值向量。

$df_{\min}$ 是拟合模型的自由度。

有关详细信息，请参见 Maiti and Mukherjee (1991) 和 West et al.(2012)。

RMSEA

近似的均方根误差 (RMSEA) 定义如下：

$$\text{RMSEA} = \sqrt{\frac{\max(\chi^2_{min} - df_{min}, 0)}{n \times df_{min}}}$$

其中：

n 是样本大小。

$df_{\min}$ 是拟合模型的自由度。

χ^2_{min} 是拟合模型的卡方统计量。

利用非中心卡方分布的累积分布函数 $\Phi(x|\lambda, d)$ 计算 RMSEA 的置信限。90% 置信限计算如下：

$$\text{下限} = \sqrt{\frac{\lambda_L}{n \times df_{min}}}$$

$$\text{上限} = \sqrt{\frac{\lambda_U}{n \times df_{min}}}$$

其中：

λ_L 满足 $\Phi(\chi^2_{min} \mid \lambda_L, df_{min}) = 0.95$。

λ_U 满足 $\Phi(\chi^2_{min} \mid \lambda_U, df_{min}) = 0.05$。

详细信息，请参见 Maydeu-Olivares et al.(2017) 中的表 J.1a、J.1b、J.判别主成分a 和 J.判别主成分b。

RMR 和 SRMR

RMR 和 SRMR 的公式定义如下：

$$\text{RMR} = \sqrt{\frac{1}{b}\left[\sum_{i}^{p}\sum_{j}^{i}(s_{ij} - \hat{\sigma}_{ij})^2 + \sum_{i}^{p}(\bar{x}_i - \hat{\mu}_i)^2\right]}$$

$$\text{SRMR} = \sqrt{\frac{1}{b}\left[\sum_{i}^{p}\sum_{j}^{i}\frac{(s_{ij} - \hat{\sigma}_{ij})^2}{s_{ii}s_{jj}} + \sum_{i}^{p}\frac{(\bar{x}_i - \hat{\mu}_i)^2}{s_{ii}}\right]}$$

其中：

p 是显变量数。

b 是观测变量的协方差矩阵和均值向量中的唯一条目数：

$$b = \frac{p(p+1)}{2} + p$$

s_{ij} 是输入协方差矩阵的第 (i, j) 个元素。

$\hat{\sigma}_{ij}$ 是预测协方差矩阵的第 (i, j) 个元素。

$\bar{x}_i$ 是样本均值向量的第 i 个元素。

$\hat{\mu}_i$ 是向量预测均值的第 i 个元素。

有关详细信息，请参见 SAS Institute Inc.(2020b) 中的“CANDISC 过程”一章。

第 9 章

因子分析

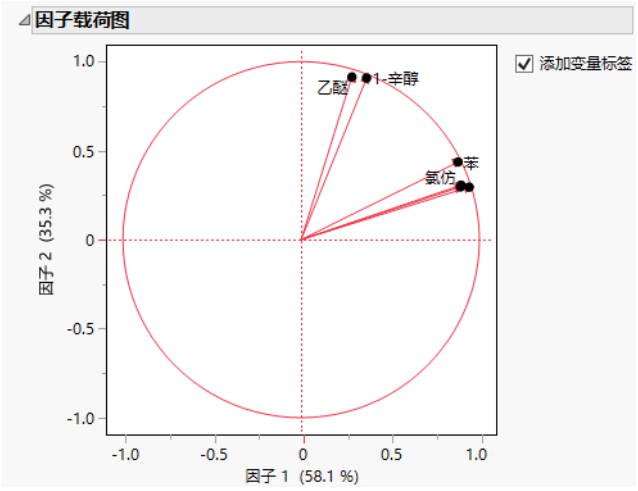
标识数据中的潜在变量

因子分析旨在用少量（不可观测）的潜在变量或因子描述可观测变量。因子分析也称为公因子分析和探索性因子分析。这些因子可以定义为观测变量的线性组合（加上误差）。旨在解释观测变量中的**共有**变异。因子分析的目的是通过未观测到的因子来发现针对观测变量的有意义的解释，除此之外我们还可以减少变量数。

因子分析在许多领域中都有广泛的应用，其发源于心理学、社会学和教育学。在这些领域中，因子分析有助于理解如何通过潜在模式和结构来解释表象行为。例如，用来衡量参与户外活动、爱好、锻炼和旅游的测度可能全都与可描述为“个性活跃/不活跃”的因子相关。

在您需要探索或解释数据中的潜在模式和结构时可使用因子分析。还可考虑通过因子分析借助少量的潜在变量来汇总变量中的信息。

图 9.1 旋转的因子载荷



目录

“因子分析”平台概述

因子分析根据较小数量的无法观测的因子对一组可观测的变量建模。构造这些因子可解释观测变量之间的相关性或协方差。因子旋转用于更改因子的参考轴，使其更加容易解释。

考虑有十个观测变量的情形：$X_1, X_2, \ldots, X_{10}$。假定您想要根据两个潜在因子 F_1 和 F_2 对这十个变量建模。为方便起见，假定这些因子之间不相关，而且每个因子的均值为 0，方差为 1。您想要得到的模型形式如下：

$$X_i = \beta_{i0} + \beta_{i1}F_1 + \beta_{i2}F_2 + \varepsilon_i$$

由此判定：$Var(X_i) = \beta_{i1}^2 + \beta_{i2}^2 + Var(\varepsilon_i)$。其中可归因于因子的 X_i 的方差部分为 $\beta_{i1}^2 + \beta_{i2}^2$，我们称之为公共方差或**公因子方差**。而剩余方差 $Var(\varepsilon_i)$ 是特殊方差，被视为 X_i 所特有的特定和误差方差的组合。

该平台为相关性或协方差矩阵的特征值提供了一张陡坡图。您可以根据陡坡图来确定要提取的因子数。该平台的默认因子数是超过 1 的特征值数。

该平台提供两种因子分解方法来估计该模型的参数：主轴和最大似然。通过两个“先验公因子方差”选项可估计每个变量的公因子对公差贡献的比例。这些选项针对相关性（或协方差）矩阵的对角线有不同的假设前提。“主成分”选项从相关性矩阵（对角线元素为 1）或协方差矩阵（对角线元素为变量的方差）出发进行后续分析，“公因子分析”选项将对角线元素设置为多重相关性的平方。这些值反映与其他变量分享的变异比例。

可利用因子旋转来支持对提取因子的解释。“因子分析”平台提供了多种旋转方法，其中包含正交旋转和斜交旋转。

与考虑公共方差的因子分析相比，主成分分析解释观测变量的总方差。请参见“主成分”。

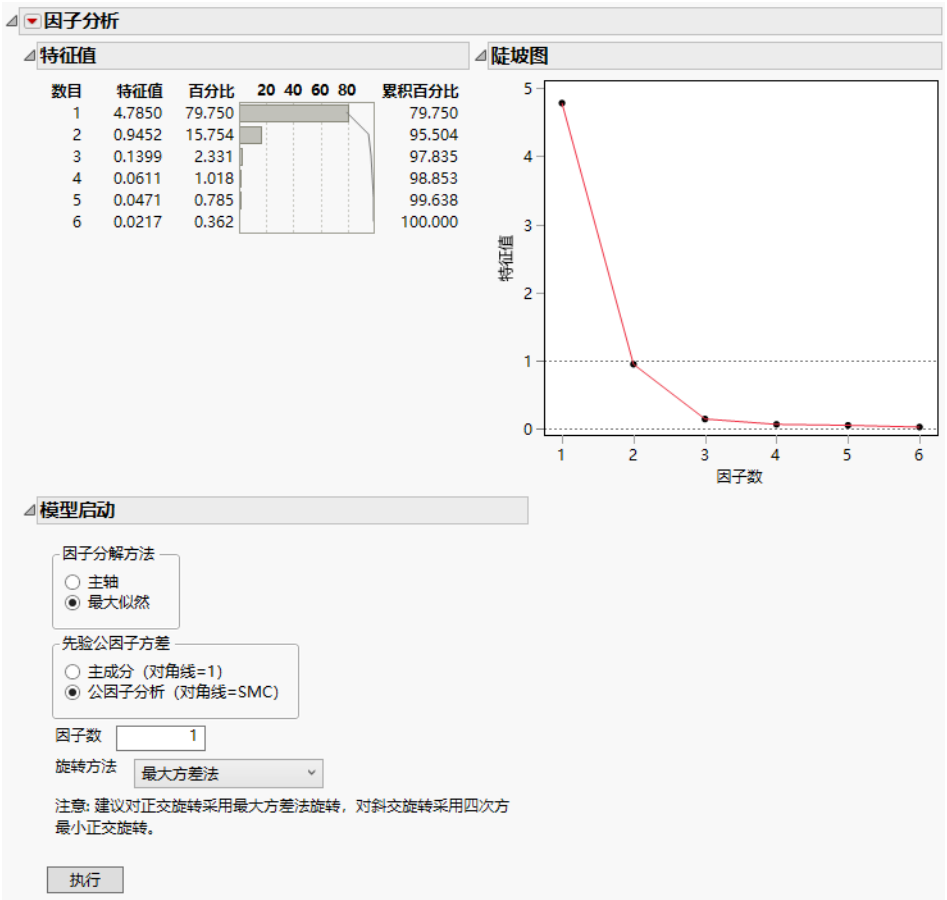
有关因子分析的详细信息，请参见 Jöreskog (1977) 或 Cudeck and MacCallum (2007)。

“因子分析”平台的示例

使用 “因子分析”平台提取由六种溶剂解释的两个因子。

1. 选择**帮助** > **样本数据文件夹**，然后打开 Solubility.jmp。
2. 选择**分析** > **多元方法** > **因子分析**。
3. 从 **1- 辛醇**一直选到**己烷**，然后点击 **Y，列**。
4. 点击**确定**。

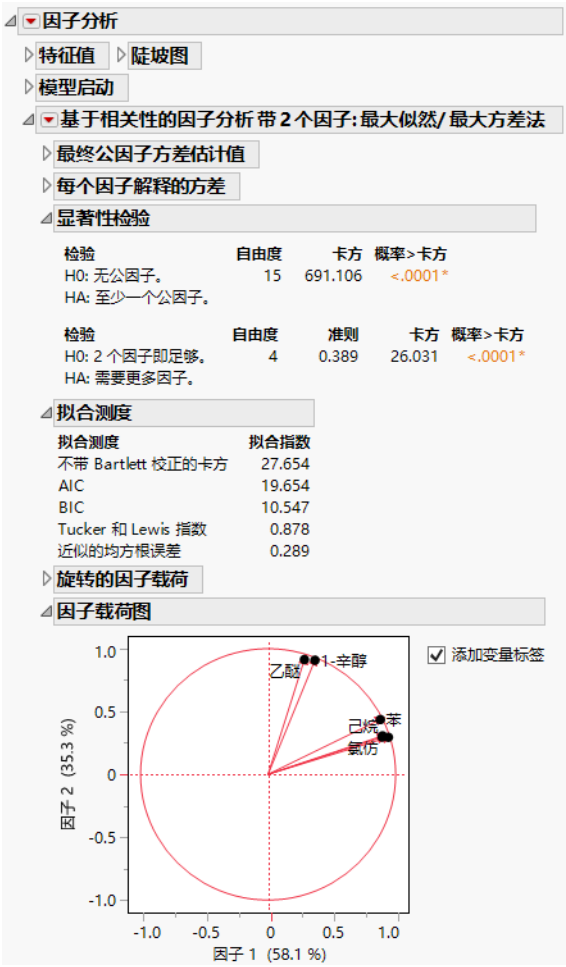
图 9.2 初始 “因子分析”报表



陡坡图的弯肘部指示包含两个因子的模型很合适。

5. 在 “模型启动”分级显示项中，做出以下选择：
   - 因子分解方法= **最大似然**
   - 先验公因子方差= **公因子分析**
   - 因子数 = 2
   - 旋转方法= **最大方差法**
6. 点击**执行**。

图 9.3 “因子分析”报表



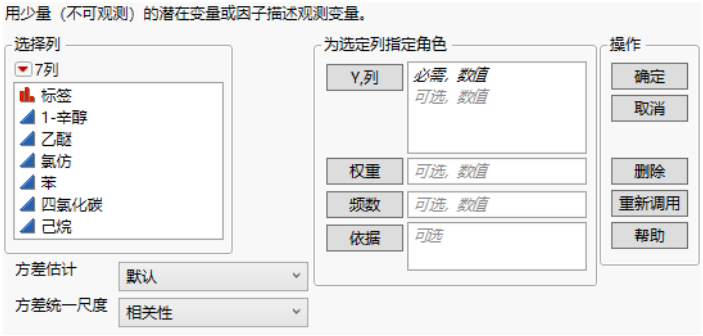
该报表列出公因子方差估计值、方差估计值、显著性检验、拟合测度、旋转的因子载荷和因子载荷图。“旋转的因子载荷”和“因子载荷图”表明：“因子 1”与**四氯化碳-氯仿-苯-己烷**这一组变量相关，“因子 2”与**乙醚–1- 辛醇**这一组变量相关。请参见“因子分析模型拟合选项”，了解有关报表中所显示信息的详细信息。

提示：点击“因子载荷图”中的点可以选择和移动标签。点击右下角可增大图的大小，以便更容易查看标签。

启动“因子分析”平台

通过选择**分析** > **多元方法** > **因子分析**启动“因子分析”平台。

图 9.4 “因子分析”启动窗口



有关“选择列”红色小三角菜单中选项的详细信息，请参见《使用 JMP》。

Y，列 要分析的列。这些列的数据类型必须为“数值”。

权重 包含数据表中每个观测的权重值的列。仅当行值大于零时才在分析中包含该行。

频数 为分析中的每行分配一个频数。该选项适用于汇总数据。

依据 为“依据”变量的每个水平生成单独报表。若指定了多个“依据”变量，将为“依据”变量水平的每种可能组合生成单独的报表。

方差估计 列出用于估计分析的方差 - 协方差矩阵的方法。有关这些方法的详细信息，请参见“‘多元’报表”。

方差统一尺度 列出用于执行因子分析的统一尺度方法。

相关性 支持相关性分析的默认方法。

协方差 支持针对加权相关性矩阵的分析，其中的权重是变量方差。

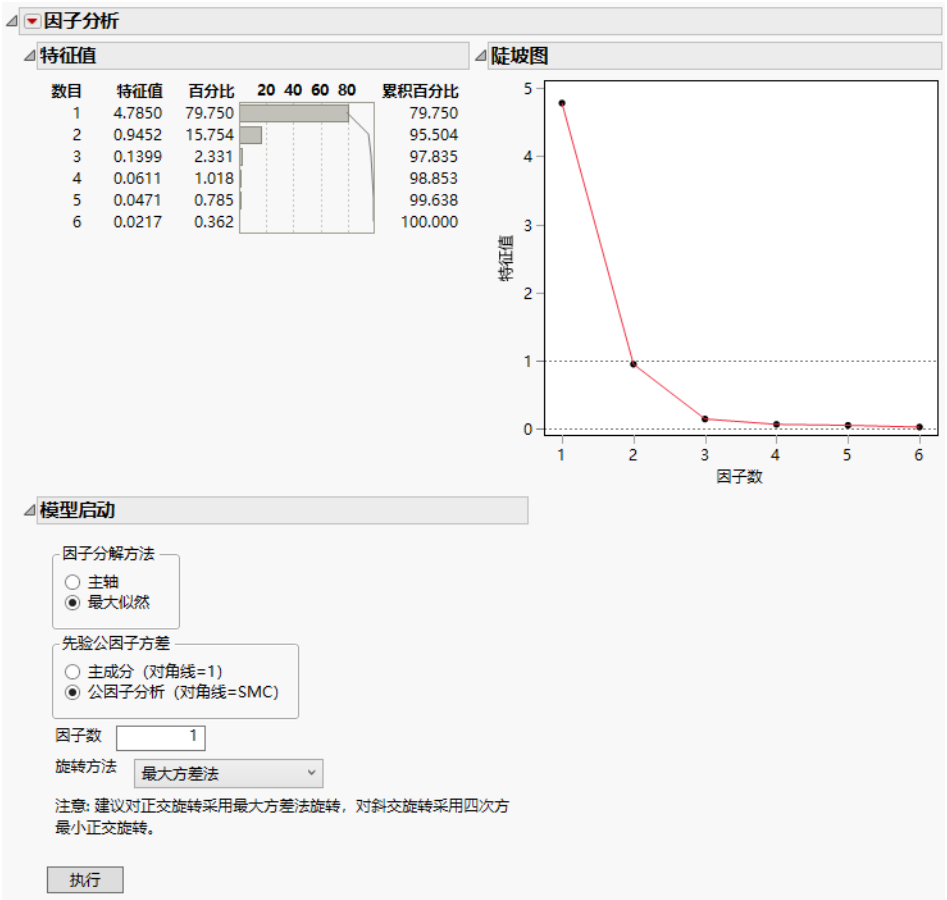
未统一尺度且未中心化 支持针对已经中心化或统一尺度的变量执行分析。

“因子分析”报表

初始“因子分析”报表显示了“特征值”和“陡坡图”。“特征值”从主成分分析获得。“陡坡图”是对这些特征值进行绘制的图形。“模型启动”中的初始因子数等于超过 1.0 的特征值的数目。您可以更改要提取的因子数。

可借助陡坡图的指导来选择因子数。陡坡图平坦之前的特征值数提供因子数的上限。对于本例，陡坡图建议两个成分（因子）。

图 9.5 “因子分析”报表

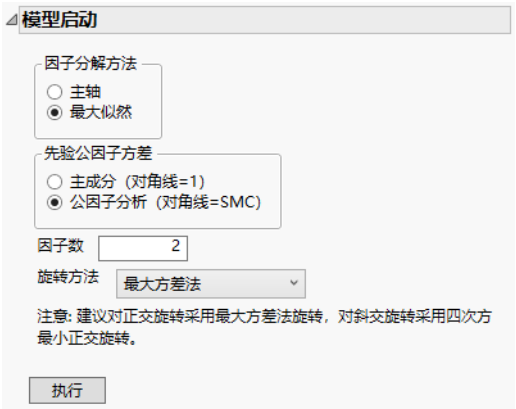


该“特征值”表显示第一个特征值解释了变异的 79.75%，第二个特征值解释了 15.75%。因此，前两个特征值共解释了总变异的 95.50%。第三个特征值仅解释了 2.33% 的变异，而其余特征值的贡献可忽略不计。尽管**因子数**框最初设置为 1，该分析建议提取 2 个因子才是合理的。

模型启动

使用 “模型启动”控制面板配置 “因子分析”模型。点击**执行**获取因子分析的结果。

图 9.6 模型启动



因子分解方法 定义提取因子的方法：

主轴 针对简化的相关性或协方差矩阵（其中矩阵的对角线被变量的公因子方差估计值取代）执行特征值分解。该方法计算效率高，但它不提供假设检验。

最大似然 支持您检验关于公因子数目的假设，并且获取模型拟合统计量。

注意: 最大似然方法要求正定相关性矩阵。若相关性矩阵不是正定矩阵，请选择“主轴”方法。

先验公因子方差 定义公因子对每个变量的方差贡献比例的估计方法。

主成分（对角线 = 1） 将所有公因子方差设置为 1，指示每个变量的方差 100% 都是由所有因子解释的。

提示：在 “因子分解方法”设置为 “主轴”时使用该选项将生成主成分分析。

公因子分析（对角线 = SMC） 将公因子方差设置为多重相关的平方 (SMC) 系数。对于给定的变量，SMC 即为该变量对其他所有变量作回归的 R 方。

因子数 指定从分析中提取的因子数。默认值是大于或等于 1.0 的特征值数。您可以设置因子数至少为 1 且不超过变量数。

旋转方法 定义旋转方法。默认方法为 “最大方差法”。请参见 “旋转方法”，了解可用旋转方法的说明。

旋转方法

在“因子分析”平台中，使用旋转可更改因子的参考轴，从而使因子更容易解释。旋转应用于从数据中提取的因子。旋转方法基于各种复杂性或简便性函数。有关旋转的详细信息，请

参见 SAS Institute Inc.(2020c) 中的“FACTOR 过程”一章、Browne (2001) 或 Frank and Todeschini (1994)。

初始提取之后的因子彼此之间是不相关的。若因子是通过正交变换旋转的，则旋转之后的因子也不相关。若因子是通过斜交变换旋转的，则旋转之后的因子就会彼此相关。斜交旋转往往生成比正交旋转更容易解释的因子。不过因子相关后，在考察对变量的解释时，很难衡量各因子的重要性。

正交旋转方法

最大方差法 将某个因子在所有变量上的载荷平方的方差之和最大化。这种公共方法导致每个变量在每个因子上都有或小或大的载荷。（$\gamma = 1$ 的直交旋转法。）

双四次幂极大法 最大方差法和四次方最大正交旋转法的等权解。（$\gamma = 0.5$ 的直交旋转法。）

相等最大值法 最大方差法旋转和四次方最大正交旋转之间的加权解。（$\gamma = N/2$ 的直交旋转法，其中 N = 因子数。）

Parsimax 因子法 旨在将因子复杂性最小化的一种解决方案。该方法可能导致交叉载荷，因为算法中未考虑变量复杂性。（$\gamma = N$ 的直交旋转法，其中 N = 因子数。）

直交旋转法 一种常规加权旋转方法，其中的权重用 γ 表示。许多特定正交旋转方法都是带有特定 γ 的直交旋转法。

Parsimax 平衡变量与因子复杂性。（$\gamma = (I(N-1))/(I+N-2)$ 的直交旋转法，其中 I = 项数， N = 因子数。）

四次方最大正交旋转 将解释每个变量所需的因子数最小化。（$\gamma = 1$ 的直交旋转法。）

斜交旋转方法

双四次幂极小法 将协方差比最小化的旋转方法 （$\tau = 0.5$ 的斜交转轴法。）

协方差极小法 “斜交最大方差法”旋转。（$\tau = 1$ 的斜交转轴法。）

斜交双四次幂极大法 “斜交双四次幂极大法”旋转。

斜交相等最大值法 “斜交相等最大值法”旋转。

斜交 Parsimax 因子法 “斜交 Parsimax 因子法”旋转。

斜交转轴法 一种常规加权斜交旋转方法，其中的权重用 τ 表示。许多特定斜交旋转方法都是带有特定 τ 的“斜交转轴法”旋转。

斜交 Parsimax 法 “斜交 Parsimax 法”旋转

四次最大正交旋转法 “斜交四次最大正交旋转法”旋转，等效于“四次方最小正交旋转”方法。

方差最大旋转法 “斜交最大方差法”旋转。

四次方最小正交旋转 “斜交四次方最小正交旋转”旋转，等效于“斜交四次方最大正交旋转法”（$\tau = 0$ 的斜交转轴法。）

斜交旋转法 两步旋转，首先执行最大方差法旋转，然后使用 Procrustes 旋转获得简单结构。这是一种计算高效的方法，可作为斜交转轴法的备选方法。

“因子分析”平台选项

“因子分析”红色小三角菜单包含以下选项：

特征值 显示或隐藏原始相关性矩阵、协方差矩阵或未统一尺度且未中心化的矩阵的特征值表。该表包含每个特征值所表示的总方差百分比、演示贡献百分比的条形图，以及每个后续特征值所贡献的累积百分比。大于等于 1.0 的特征值的数目可指导您选择合适的因子数进行分析。

陡坡图 显示或隐藏成分 （因子）数 - 特征值图。该图也可用作附加指导，用来确定提供了最大方差贡献率的因子数。标绘线变得平坦处的临界点可用作合适的因子数以进行分析。

Bartlett 球形检验 显示或隐藏 Bartlett 球形检验的结果。该检验是一个齐性检验，通过计算检验的卡方、自由度 (DF) 和 *p* 值 （概率 > 卡方）来确定特征值是否具有相同的方差。请参见 Bartlett (1937, 1954)。

Kaiser-Meyer-Olkin 检验 显示或隐藏 Kaiser-Meyer-Olkin (KMO) 检验的结果。该检验是一个方差比例指标，可能是公共方差，而该方差可能是由于内在因子导致的。该检验统计量针对每个单独变量和所有变量的集合计算，在 JMP 中称为抽样适当性测度 (MSA)。以下值指示变量是否适用于因子分析：

- 0.00 到 0.49 不可接受
- 0.50 到 0.59 较差
- 0.判别主成分0 到 0.判别主成分9 普通
- 0.70 到 0.79 中等
- 0.80 到 0.89 卓越
- 0.90 到 1.00 极好

注意：若相关性矩阵是奇异的，则 KMO 检验不可用。

请参见《使用 JMP》获取有关下列选项的信息：

本地数据过滤器 显示或隐藏支持您过滤特定报表中使用的数据的本地数据过滤器。

重新运行 包含使您可以重复或重新启动分析的选项。在支持该功能的平台中，“自动重新计算”选项立即在相应报表窗口中反映您对数据表所做的更改。

平台首选项 包含的选项支持您查看当前平台首选项或更新平台首选项以匹配当前 JMP 报表中的设置。

保存脚本 包含的选项支持您保存可将报表重现到若干目标的脚本。

保存 “依据”组脚本 包含使您可以保存脚本的选项，可将为“依据”变量的所有水平重新生成平台报表的脚本保存到多个不同的位置。仅当在启动窗口中指定“依据”变量时才可用。

因子分析模型拟合选项

“因子分析模型拟合”红色小三角菜单包含以下选项：

先验公因子方差 （仅适用于“公因子分析”。）显示或隐藏每个变量的公因子方差的初始估计值。对于给定的变量，该估计值是多重相关平方 (SMC) 系数，即该变量对其他所有变量构建回归模型得到的 R 方。

注意：当公因子方差等于或大于 1 时，会出现 Heywood 情形。这意味着唯一的因子方差（即，不由公共因子解释的残差方差）是负的。 Heywood 情形的处理方法是对唯一方差设定下限或对公因子方差设定上限。

特征值 （仅适用于“公因子分析”。）显示或隐藏简化相关性矩阵的特征值以及这些特征值解释的公共方差百分比。简化相关性矩阵是对角线元素被公因子方差估计值取代的相关性矩阵，特征值表明因子所解释的公共方差。“累积百分比”可超过 100%，因为简化相关性矩阵不一定是正定矩阵，可以包含负特征值。

请注意，该表指出了进行后续分析保留的因子数。

未排序和未旋转的因子载荷 显示或隐藏在排序和旋转之前的因子载荷矩阵。

未旋转的因子载荷 显示或隐藏旋转之前的因子载荷矩阵。因子载荷测量公因子对变量的影响。由于未旋转的因子是正交的，因此该因子载荷矩阵是变量与因子之间的相关性矩阵。载荷的绝对值越接近 1，因子对变量的影响就越大。

使用滑块或输入相应值来**隐藏小于该值的绝对载荷值**（该值即表中的指定值）。隐藏的值会根据**文本变暗**所指定的设置相应地变暗。

使用**文本变暗**滑块或输入相应值，控制绝对值比**隐藏小于该值的绝对载荷值**的指定值小的因子载荷值的字体透明度梯度。

图 9.7 带有文本变暗控件的未旋转的因子载荷

注意：“未旋转的因子载荷”矩阵经过排序，因此与同一因子关联的变量会显示在彼此上下。

旋转矩阵 显示或隐藏用于对因子载荷图和因子载荷矩阵进行旋转的值。

因子间相关性 （仅适用于斜交旋转。）显示或隐藏因子间的相关性矩阵。

目标矩阵 （仅适用于 “斜交旋转法”旋转。）显示或隐藏最大方差法因子模式要旋转到的矩阵。

因子结构 （仅适用于斜交旋转。）显示或隐藏变量与公因子之间的相关性的矩阵。

最终公因子方差估计值 显示或隐藏拟合因子模型后得到的公因子方差估计值。若因子是正交的，变量的最终公因子方差估计值等于该变量的因子载荷平方和。

标准得分系数 显示或隐藏将旋转后的因子保存至源数据表时用于估计因子得分的乘数表。

每个因子解释的方差 （仅适用于正交旋转。）显示或隐藏每个旋转因子解释的方差、公共方差的百分比和累积百分比。

每个因子解释的方差，忽略其他因子 （仅适用于斜交旋转。）显示或隐藏每个旋转因子解释的方差和公共方差百分比 （忽略其他因子）。

显著性检验 （仅适用于 “最大似然”因子分解方法。）提供两个卡方检验的结果。

第一个检验的 H_0：无公因子。该原假设指出没有任何公因子可以解释变量之间的交互相关。该检验为 Bartlett 球形检验，其原假设为：因子的相关性矩阵是一个单位矩阵 (Bartlett, 1954)。

第二个检验的 H_0：N 个因子足够多，其中的 N 是指定的因子数。拒绝该原假设即表明需要更多因子来解释变量之间的交互相关 (Bartlett, 1954)。准则即对数似然目标函数值。

拟合测度 （仅适用于 “最大似然”因子分解方法。）显示或隐藏拟合测度：未经 Bartlett 修正的卡方、 AIC、 BIC、 Tucker-Lewis 指数以及近似的均方根误差。

因子得分测度 显示或隐藏因子得分确定性测度：多重 R、多重 R 平方和最小相关性。这些测度用于评估因子得分对于二次分析是否有用。

未排序和旋转的因子载荷 显示或隐藏旋转之后的未排序因子载荷矩阵。

旋转的因子载荷 显示或隐藏旋转之后的因子载荷矩阵。若为正交旋转，这些值为变量与旋转因子之间的相关性。

使用滑块或输入相应值来**隐藏小于该值的绝对载荷值**（该值即表中的指定值）。隐藏的值会根据**文本变暗**所指定的设置相应地变暗。

使用**文本变暗**滑块和值控制该表的字体透明度梯度。值越小，绝对值比指定的**隐藏小于该值的绝对载荷值**小的因子值的字体就越透明。

图 9.8 带有文本变暗控件的旋转的因子载荷

注意：“旋转的因子载荷”矩阵经过排序，因此与同一因子关联的变量会显示在彼此上下。

因子载荷图 显示或隐藏旋转的因子载荷图。对两个以上因子建模时，载荷图是多个图的矩阵。

得分图 显示或隐藏估计的因子得分的散点图。对两个以上因子建模时，得分图是多个图的矩阵。

完成补缺的得分图 （仅适用于存在缺失值的情况。）显示或隐藏带有缺失值的补缺值的估计因子得分的散点图。

显示选项 支持您显示或隐藏载荷图上的箭头。

保存因子得分 将因子得分保存至数据表中的新公式列。因子得分由 Thurstone 方法估计。

注意：公式无法计算包含缺失值的行。

复制 SEM 的模型规格 将因子定义复制到剪贴板。随后可以将这些定义粘贴到 SEM 平台中。SEM JSL 脚本包含基于最终旋转载荷矩阵的固定载荷。该脚本仅包含大于**隐藏小于该值的绝对载荷值**阈值的载荷。

保存完成插补的因子得分 （可用于具有缺失数据的模型。）将使用插补缺失值计算的因子得分保存到数据表的新公式列中。

删除拟合 从“因子分析”报表中删除拟合模型结果。该选项支持您更改“模型启动”配置以生成新的报表。

第 10 章

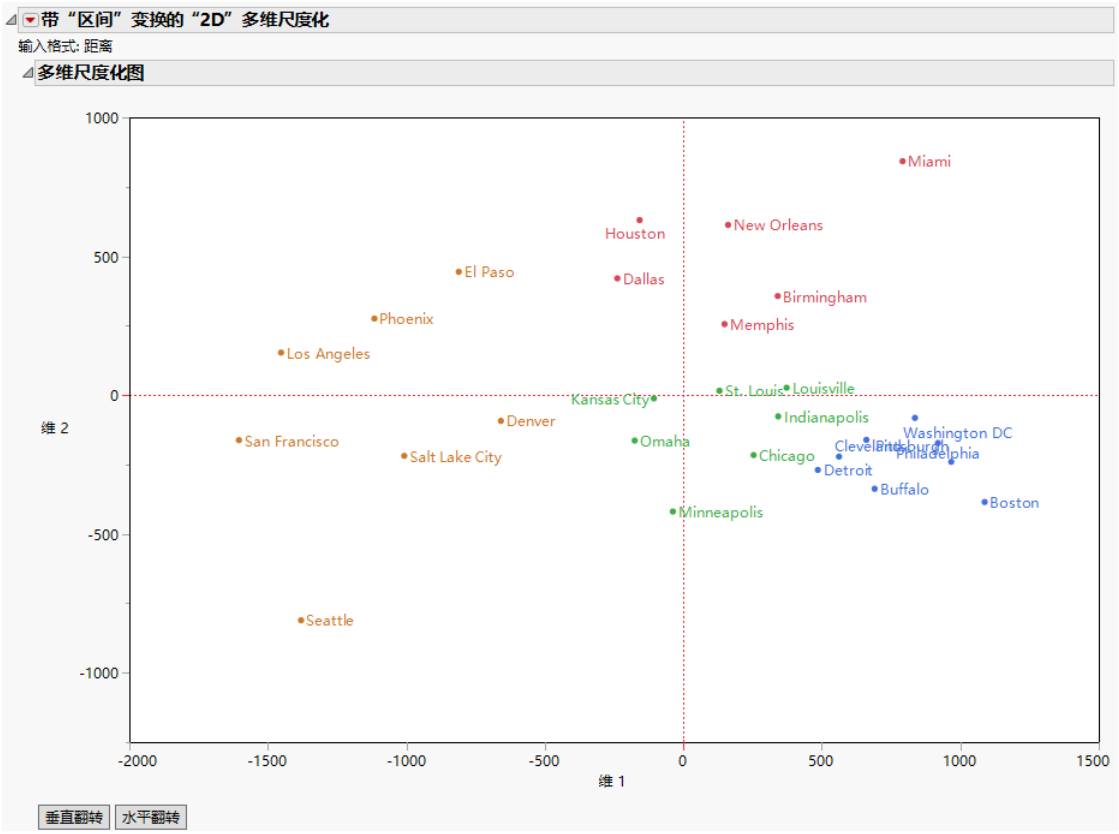
多维尺度化

直观表示一组对象间的邻近性

“多维尺度化 (MDS)”是一种方法，用于分析一组对象间的邻近性（相似性、相异性或距离），将它们之间的模式通过图形直观呈现出来。例如，考虑到城市之间的距离矩阵，MDS 可用于生成二维城市地图。

“多维尺度化”常用在消费者研究中，研究人员收集了关于品牌、口味或其他产品特性的感知测度数据。在其他很多领域，当人们想要根据一组特性（或邻近性）直观展示对象之间的邻近性时，MDS 也同样适用。

图 10.1　“多维尺度化”示例



目录

“多维尺度化”平台概述

“多维尺度化 (MDS)”平台生成一组对象之间的邻近性图。该图可用于直观探索数据集内的结构。MDS 是一种多元方法，用于分析一组对象间的邻近性（相似性或距离），以较少的维将它们之间的关系进行可视化展现。MDS 适用于距离矩阵。MDS 图的坐标通过最小化 Stress 函数（实际邻近性和预测邻近性的差异）来获取。

距离一词可指一种物理距离（比如城市之间的距离）测度。更多情况下，距离只是一种客观评估而不是精确测量。邻近性可测量不同品牌产品之间的感知相似性、犯罪率相关性或样本国家/地区的经济相似性。距离还可称为邻近性或相似性（相异性）。若数据作为特性列表提供，则首先从特性列表入手构造距离矩阵。

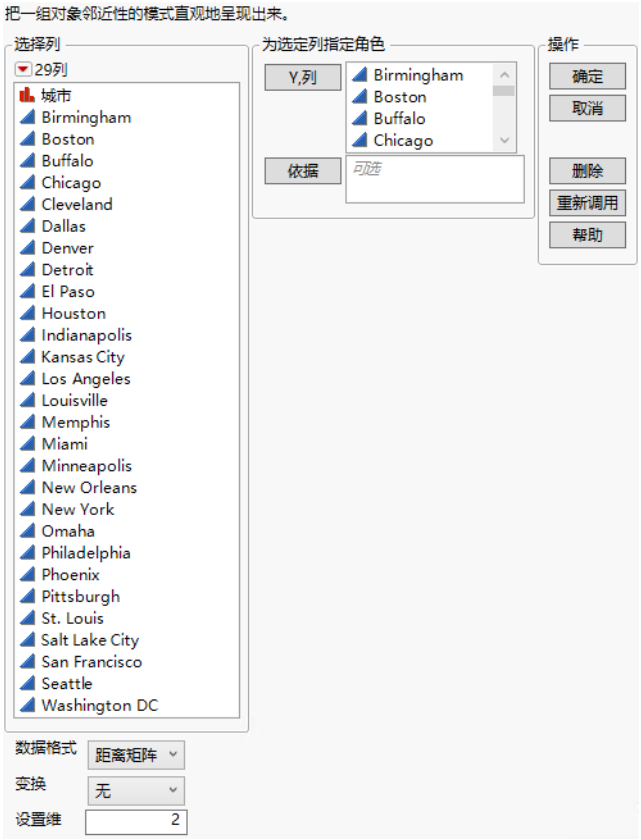
有关多维尺度化的详细信息，请参见 Borg and Groenen (2005) 或 Jackson (2003)。

“多维尺度化”示例

在本例中，使用 MDS 构造基于配对航距的 28 个城市的二维地图。包含航距的数据表是一个距离矩阵。

1. 选择**帮助 > 样本数据文件夹**，然后打开 Flight Distances.jmp。
2. 选择**分析 > 多元方法 > 多维尺度化**。
3. 从 Birmingham 一直选到 Washington DC，然后点击 **Y，列**。
4. 从“数据格式”菜单中选择**距离矩阵**。

图 10.2 完成的 “多维尺度化”启动窗口



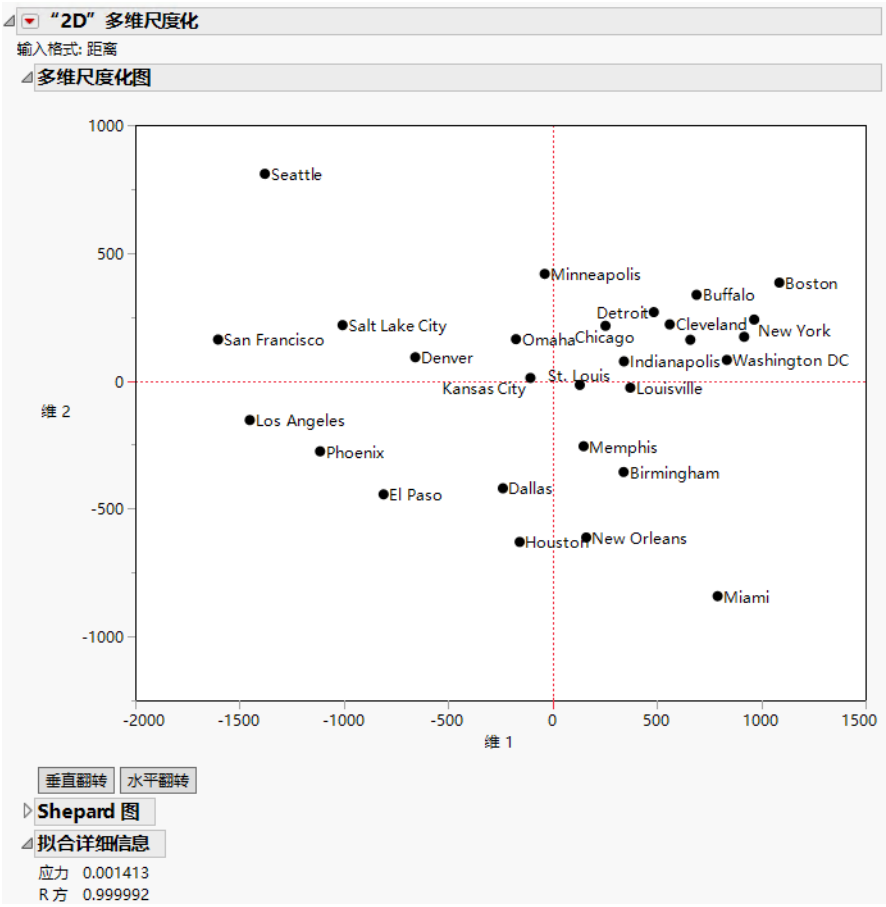
5. 点击**确定**。

   在 “多维尺度化图”中，悬停在数据点上方以查看行号或行标签。在接下来的 7 步中要对 MDS 图执行添加标签和旋转操作。

6. 选择 Flight Distances 数据表。
7. 右击列**城市**，然后选择**添加标签 / 撤销标签**。
8. 选择**行 > 行选择 > 选择所有行**。
9. 选择**行 > 添加标签 / 撤销标签**。
10. 选择 “多维尺度化图”。
11. 点击**垂直翻转**按钮。
12. 点击**水平翻转**按钮。

    “垂直翻转”和“水平翻转”按钮支持您更改 MDS 图的方向。MDS 结果不受方向影响。若方向在结果中是已知的，如物理位置，则您可能需要旋转或翻转您的图形。

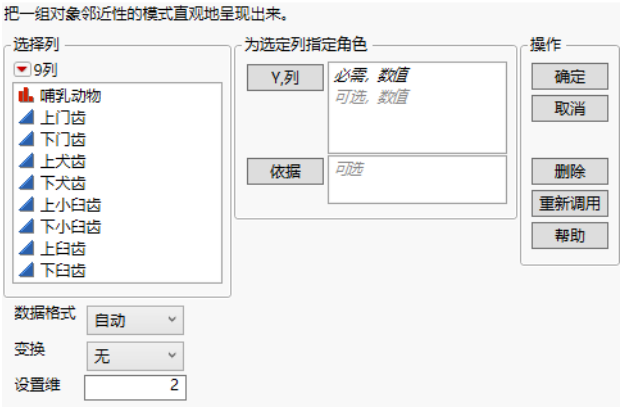
图 10.3 多维尺度化图


“2D” 多维尺度化
输入格式: 距离
多维尺度化图
1000
500
0
-500
-1000
维 2
Seattle
Minneapolis
Boston
Buffalo
Detroit
Cleveland
New York
Salt Lake City
San Francisco
Omaha
Chicago
Denver
Indianapolis
Washington DC
St. Louis
Kansas City
Louisville
Los Angeles
Phoenix
Memphis
Birmingham
El Paso
Dallas
Houston
New Orleans
Miami
-2000
-1500
-1000
-500
0
500
1000
1500
维 1
垂直翻转
水平翻转
Shepard 图
拟合详细信息
应力 0.001413
R 方 0.999992

启动“多维尺度化”平台

通过选择**分析** > **多元方法** > **多维尺度化**启动 “多维尺度化”平台。

图 10.4 “多维尺度化”启动窗口



有关“选择列”红色小三角菜单中选项的详细信息，请参见《使用 JMP》。

Y，列 要分析的列。这些列的数据类型必须为 “数值”。

依据 一个或多个列，其水平定义不同的分析。对于指定列的每个水平，都使用您已经指定的其他变量分析相应行。结果显示在不同的报表中。若指定了多个 “依据”变量，将为 “依据”变量水平的每种可能组合生成单独的报表。

注意：使用距离矩阵时，“依据”变量要求为 “依据”变量的每个水平生成完整矩阵。

数据格式 MDS 支持两种数据格式：您可以指定一种格式或允许 JMP 推断数据格式。

自动 数据格式是推断出的。使用启发方式识别距离矩阵表，并且假定其他所有数据表均具有特性列表格式。

距离矩阵 完全对称、下或上三角矩阵，其中的行数等于列数。对角线元素可以为零或缺失。

特性列表 包含质量测度或对象特征的一组列。假定测度都使用相同的尺度。对象通常在列中命名。对象列不用在分析中，但用作 MDS 图上数据点的标签。

变换 支持的变换包括 “比”、“区间”和 “有序型”。

无 不使用变换。

比 数据从小到大排序，值间差异有意义，并且尺度具有真实的零。用来对 MDS 图进行尺度变换。

区间 数据从小到大排序，并且值间差异有意义。用来对 MDS 图进行尺度变换和偏移。

有序型 数据从小到大排序。用于有序型数据。

设置维 用于直观表示对象间的邻近性的维数。通常使用二维或三维。三维以上的直观表示会变得较为复杂。

注意：选定的维可介于 1 到 n - 1 之间，其中 n = 对象数，否则维设置为 2。

处理距离矩阵和多维尺度优化是占用大量计算资源的过程。在处理大型距离矩阵时，进度窗口支持您监控或取消该过程。对于优化，MDS 使用非凸优化算法。重要的是使用多个起始值来避免局部最优解。对于中型到大型数据集，交互式进度窗口允许您监控优化进度、接受当前开始的估计值或接受当前解。

“多维尺度化”报表

初始 MDS 报表包含以下部分：

- “多维尺度化图”
- “Shepard 图”
- “拟合详细信息”

若在启动窗口中为拟合指定了三维或更多维，“MDS 图”则提供相应的控件用于选择您查看的维。

MDS 图上距离相近的对象具有类似的特征。向图中添加标签和颜色有助于发现类似组。Shepard 图和拟合汇总统计量提供一些测度，用于衡量 MDS 图在多大程度上反映了对象间的邻近性。

多维尺度化图

MDS 图将多维尺度化结果在二维空间进行展示。该图下方有两个按钮，分别用于在垂直和水平方向上翻转轴。可以反射、旋转或转换 MDS 解，而不会改变点间的邻近性。在处理具有已知地图方向的地理位置时，对轴进行旋转或反射是最常用的。

若在分析中使用了两个以上的维，则可以使用图下方的**选择维**控件切换或逐步浏览该图中显示的维。第一个控件定义图的水平轴，第二个控件定义图的垂直轴。

Shepard 图

在“多维尺度化”报表中，“Shepard 图”是实际或变换的邻近性与预测邻近性的对比图。该图反映了多维尺度化图在多大程度上反应了实际的邻近性。Shepard 图类似于“预测值 - 实际值”图。理想情况下，这些点落在 Y = X 线上（这条线显示为红色）。

拟合详细信息

在“多维尺度化”报表中，“拟合详细信息”部分提供了汇总统计量以衡量 MDS 邻近性与实际邻近性之间的吻合程度，同时还在使用变换时提供关于变换的详细信息。

Stress 在拟合过程中最小化的 Stress 函数 (Stress1) 的值。 Stress 可介于 0 和 1 之间，较低的值表示较好的拟合。

R 方 “预测的邻近性 - 实际或变换的邻近性”的线性拟合的 R^2 值。

斜率 若使用了比或区间变换，则提供变换的斜率。这是实际邻近性针对变换的邻近性的线性回归的斜率。

截距 若使用了区间变换，则提供变换的截距。这是实际邻近性针对变换的邻近性的线性回归的截距。

“多维尺度化”平台选项

使用 “多维尺度化”红色小三角菜单选项，您可以根据自己的需要定制报表。可用选项取决于您用于分析的数据类型和维数。

MDS 图 显示或隐藏 MDS 图。

诊断 提供 MDS 的诊断。

Shepard 图 显示实际邻近性（若使用变换，则为变换的邻近性）与预测的邻近性的对比图。默认显示该报表。请参见“Shepard 图”。

Waern 链接 在 MDS 图上显示 Waern 链接。选定该选项时，将可以使用针对此部分（最小或最大）的控件。请参见“Waern 链接”。

显示坐标 提供解坐标报表。这些是“多维尺度化图”上的点的坐标。该报表显示至多三个维的坐标。右击该报表并选择**列**可将更多维添加到报表中。最大维数是在启动窗口中设置的维数。

显示邻近性 显示邻近性报表。在每对对象之间都提供了原始和派生的邻近性 （距离） 。对象对在 “从对象”和 “至对象”列中标识。若使用了变换，变换的邻近性也包含在该表中。

保存邻近性 （仅当数据格式是 “特性列表”时可用。）将距离矩阵保存到数据表中。

3D 图 （仅当在启动窗口中为 “设置维”指定了三个或更多维时才可用。 ）显示前三个维的 3D 图。

保存坐标 将解坐标保存到数据表中的单独列内。

请参见《使用 JMP》获取有关下列选项的信息：

本地数据过滤器 显示或隐藏支持您过滤特定报表中使用的数据的本地数据过滤器。

重新运行 包含使您可以重复或重新启动分析的选项。在支持该功能的平台中，“自动重新计算”选项立即在相应报表窗口中反映您对数据表所做的更改。

平台首选项 包含的选项支持您查看当前平台首选项或更新平台首选项以匹配当前 JMP 报表中的设置。

保存脚本 包含的选项支持您保存可将报表重现到若干目标的脚本。

保存 “依据”组脚本 包含使您可以保存脚本的选项，可将为“依据”变量的所有水平重新生成平台报表的脚本保存到多个不同的位置。仅当在启动窗口中指定“依据”变量时才可用。

Waern 链接

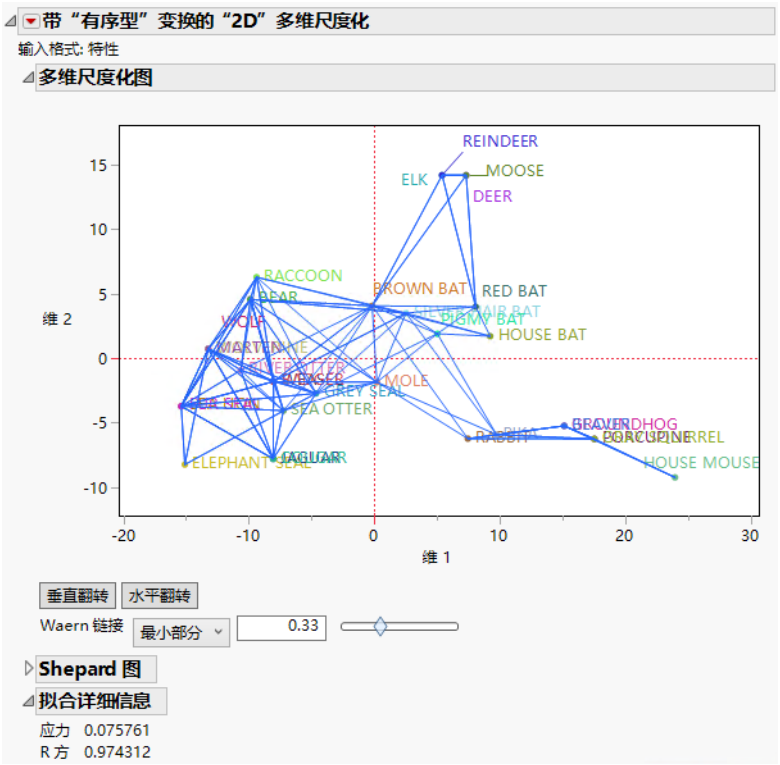
通过将实际邻近性与预测邻近性加以比较，Waern 链接提供对“多维尺度化”结果的可视检查。这些链接基于“多维尺度化图”上各点的实际邻近性连接各点。具有最小（最大）邻近性的对象将连接起来。可以考虑的典型方案是对象之间最小 33% 的邻近性。若 MDS 图能很好地表示邻近性，则针对最小实际邻近性的链接应连接图中最靠近的对象。若针对较小邻近性的链接在图中延伸，以至于连接了相距较远的对象，则需要质疑 MDS 拟合。

Waern 链接控件

您可以从一个列表中进行选择，以显示图中链接的“最小部分”或“最大部分”。通过在框中输入值或使用滑块可以控制显示的链接比例。图 10.5 显示针对 Teeth.jmp 数据表中最小的占比 33% 的 Waern 链接。

有关 Waern 链接的详细信息，请参见 Waern (1972)。

图 10.5 带 Waern 链接的 MDS 图

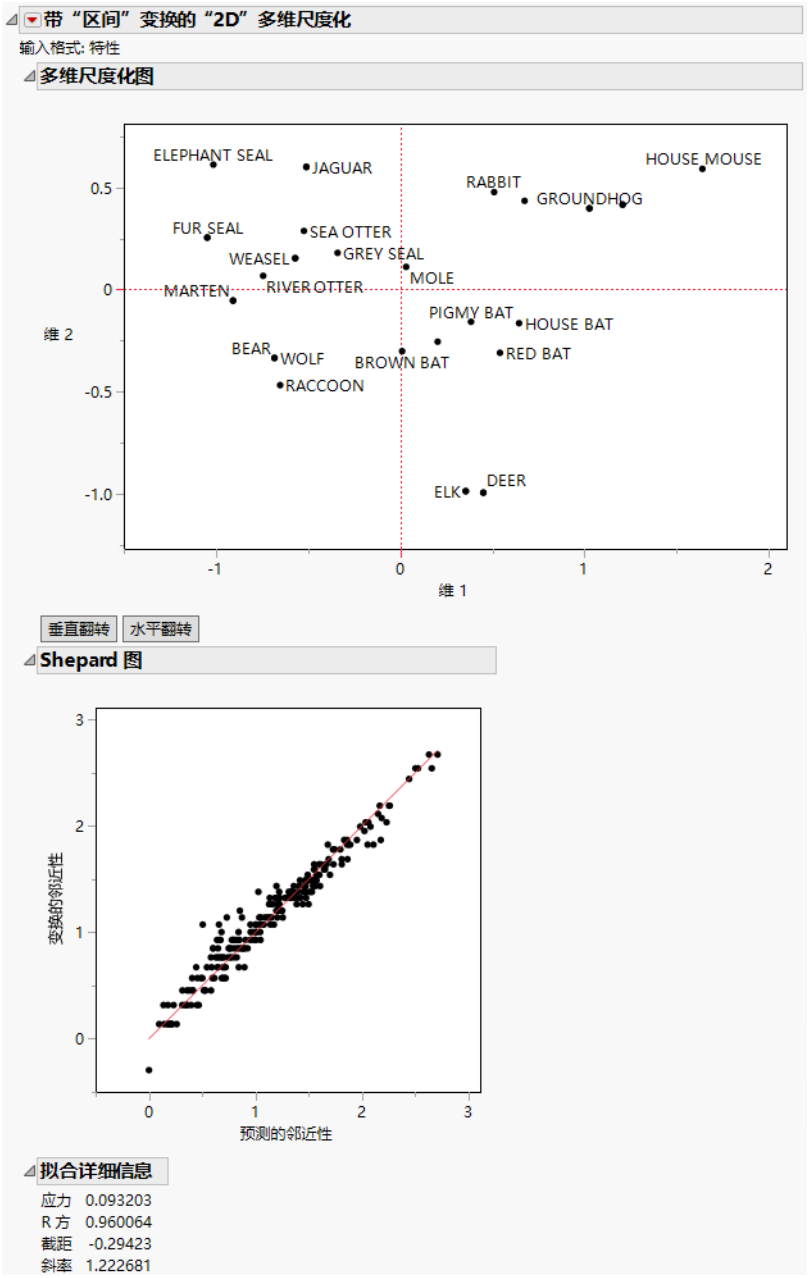


“多维尺度化”的更多示例

基于 32 种哺乳动物的牙齿特性数据使用 MDS 探索其相似性。使用区间变换来演示该变换的输出。数据确实具有有意义的排序（四颗牙齿的数量是两颗牙齿的两倍之多）。您可以探索其他变换，如有序型变换。

1. 选择**帮助** > **样本数据文件夹**，然后打开 Teeth.jmp。
2. 右击列**哺乳动物**，然后选择**添加标签 / 撤销标签**。
3. 选择**行** > **行选择** > **选择所有行**。
4. 选择**行** > **添加标签 / 撤销标签**。
5. 选择**分析** > **多元方法** > **多维尺度化**。
6. 从**上门齿**一直选择到**下臼齿**，然后点击 **Y，列**。
7. 选择**数据格式** > **特性列表**。
8. 选择**变换** > **区间**。
9. 点击**确定**。

图 10.6 “多维尺度化”报表



“Shepard 图”和“拟合详细信息”表明：由于动物牙齿的相似性， MDS 图可以很好地表示动物的相似性。“Stress”统计量 0.093 较低，“预测的邻近性 - 变换的邻近性”的 R^2 拟合高达 0.9 判别主成分。此外，“拟合详细信息”提供实际邻近性的变换的截距和斜率。

“多维尺度化”平台的统计详细信息

本节包含 “多维尺度化”平台的统计详细信息。

- “Stress 函数的统计详细信息”
- “变换的统计详细信息”
- “特性列表格式的统计详细信息”

Stress 函数的统计详细信息

在 “多维尺度化”平台中，使用 Quasi-Newton 优化法通过最小化 Stress 函数来确定 MDS 坐标。这种最小化会根据拟合前确定的维数生成一组坐标值，以最小化多维空间里对象间的派生的邻近性测度值。若数据为有序型，则使用单调回归。否则使用标准最小二乘回归。

使用以下符号定义 Stress：

- i, j - 对象数的下标
- d_{ij} - 对象 i 与 j 之间的派生距离

  δ_{ij} - 对象 i 与 j 之间的观测相对距离
- $f(\delta_{rs})$ - 距离的变换函数

Stress 函数定义如下：

$$\text{Stress} = \left[\frac{\sum_{i<j} [f(\delta_{ij}) - d_{ij}]^2}{\sum_{i<j} d_{ij}^2} \right]^{\frac{1}{2}}$$

这一 Stress 测度亦称 Kruskal Stress I 型，或简称 Stress1。

变换的统计详细信息

在“多维尺度化”平台中，使用变换来对实际的邻近性进行尺度变换。考虑到数据中的特定结构，变换被认为能够改进 MDS 对实际邻近性的呈现。变换函数中的参数在最小化算法中变为额外参数。本节使用“Stress 函数的统计详细信息”中所述的符号。

比变换

对于比数据：

$$f(\delta_{rs}) = b\delta_{rs}$$

区间变换

对于区间数据：

$$f(\delta_{rs}) = a + b\delta_{rs}$$

有序型变换

对于有序型数据，数据不进行变换，算法使用单调回归而不是最小二乘回归。

特性列表格式的统计详细信息

若数据表为特性列表，它将转换为距离矩阵，然后应用“多维尺度化”。使用欧氏距离计算距离矩阵。对于每对项，通过以下公式定义项之间的距离：

$$\delta_{ij} = \sqrt{\sum_{k} \frac{(x_{ki} - x_{kj})^2}{k}}$$

其中 k 是特性数目。假定测度都使用相同的尺度。

注意：有关 MDS 平台的高级示例，请参见 San Francisco Crime Distances.jmp 样本数据表和该表的源脚本。该脚本使用配对相关性创建距离矩阵。生成的距离矩阵随后用于探索犯罪类别之间的关系。

第 11 章

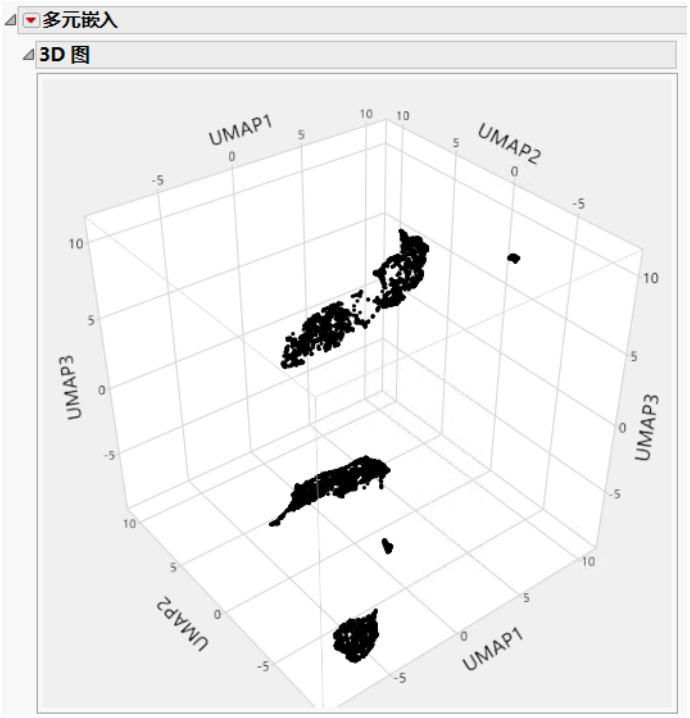
JMP PRO 多元嵌入

将数据从高维空间映射到低维空间

“多元嵌入”平台仅在 JMP Pro 中提供。

“多元嵌入”平台支持您将数据从极高维空间映射到低维空间。很多时候，您希望将数据映射到二维或三维，以便可以轻松地直观演示低维空间。在“多元嵌入”平台中，您可以使用“统一流形逼近与投影”(UMAP) 方法或“t 分布随机近邻嵌入”(t-SNE) 方法。这两种方法都尝试以更易识别近邻聚类的方式填充低维空间。

图 11.1 多元嵌入的示例



目录

JMP PRO “多元嵌入”平台概述

“多元嵌入”平台执行降维操作，这会将高维空间中的点 $\{x_1, x_2, \ldots, x_n\}$ 映射到低维空间中的点 $\{y_1, y_2, \ldots, y_n\}$。降维的目标是将点映射到低维，同时仍然保留高维数据中存在的重要信息。“多元嵌入”平台中使用的特定方法是“统一流形逼近与投影”(UMAP) 方法和“t 分布随机近邻嵌入”(t-SNE) 方法。UMAP 方法是一种流形学习方法，也称为非线性降维。该方法基于 Riemannian 几何和代数拓扑 (May, 1992)。t-SNE 方法是“随机近邻嵌入”(Hinton and Roweis, 2002) 的一种变化形式。

“多元嵌入”平台中提供的这两种降维方法都属于基于 k 近邻的学习算法。这些类型的算法首先找到每个点的近邻，以便在高维空间中创建 k 近邻图。然后，创建低维映射以将点从高维空间映射到低维空间，同时保持图的结构。

UMAP 方法概述

UMAP 方法首先找到每个点的近邻，然后创建 k 近邻图以构建拓扑结构。使用默认设置，每个点连接到至少一个其他点，即最近邻，并且在第 15 个近邻后不连接任何近邻。这之间的近邻形成了一个模糊区域。然后，通过将模糊区域的边缘合并在一起来创建高维数据的拓扑表示。有关如何合并边缘的详细信息，请参见 McInnes et al.(2018)。

为了创建低维映射，UMAP 使用梯度下降将高维拓扑表示与低维拓扑表示之间的交叉熵最小化 (McInnes et al., 2018)。UMAP 方法在尽量缩短计算时间的同时保留了数据的全局结构，并且能够处理非常大的数据集。

t-SNE 方法概述

t-SNE 方法基于点之间的配对相似性。每一配对相似性由两点是近邻的条件概率表示。在高维空间中，使用高斯分布将距离转换为条件概率。在低维映射中，使用自由度为 1 的 Student t 分布将距离转换为概率。t-SNE 方法因此而得名 (van der Maaten and Hinton, 2008)。

对于良好的低维映射，高维空间中 $\{x_i, x_j\}$ 之间的配对相似性与低维空间中 $\{y_i, y_j\}$ 之间的配对相似性相同。在这种假设下，t-SNE 方法找到了一个低维映射，该映射将高维相似性和低维相似性之间的差异最小化。使用 Kullback-Leibler 散度的一个版本来测量差异，然后使用梯度下降将其最小化。有关 t-SNE 方法的详细信息，请参见“‘多元嵌入’平台的统计详细信息”。

JMP PRO 多元嵌入的示例

使用 UMAP 方法将高维数据降低到二维。

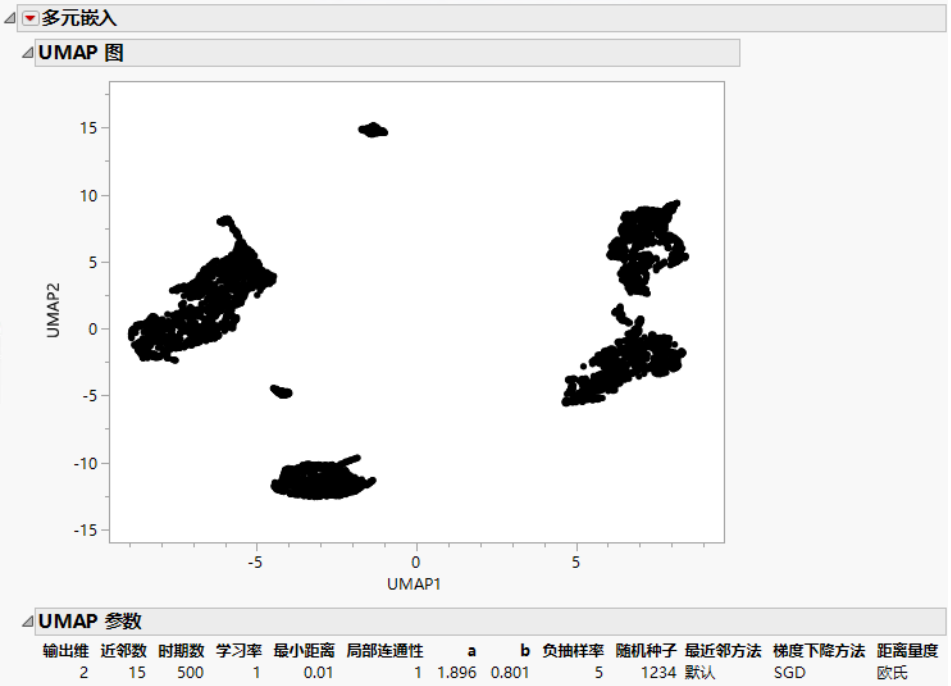
1. 选择**帮助** > **样本数据文件夹**，然后打开 Cytometry.jmp。
2. 选择**分析** > **多元方法** > **多元嵌入**。
3. 从 CD3 一直选择到 MCB，然后点击 **Y，列**。

4. 确认 “方法” 设置为 UMAP。
5. （可选。）在**随机种子**旁边键入 1234。

> **注意：**输入上面的种子可以重现本示例中所示的结果。

6. 点击**确定**。
7. 点击 “迭代历史记录” 旁边的灰色小三角。

图 11.2 “多元嵌入” 报表

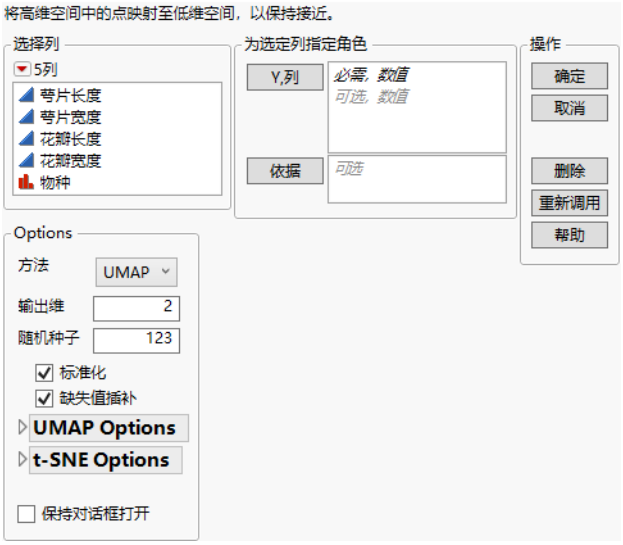


“多元嵌入” 报表窗口包含使用 UMAP 方法和计算算法的指定参数设置计算的二维图。

JMP PRO 启动 “多元嵌入” 平台

通过选择**分析 > 多元方法 > 多元嵌入**来启动 “多元嵌入” 平台。

图 11.3 “多元嵌入” 启动窗口



有关 “选择列” 红色小三角菜单中选项的详细信息，请参见 《使用 JMP》。“多元嵌入” 启动窗口包含以下选项：

Y，列 指定表示要映射到低维空间的高维数据的列。

依据 一列，其水平定义不同的分析。对于指定列的每个水平，都使用您已经指定的其他变量分析相应行。结果显示在单独的表和报表中。若分配了多个 “依据” 变量，则为 “依据” 变量水平的每个可能组合生成单独分析。

方法 指定将数据映射到低维空间的方法。在 UMAP 和 t-SNE 之间进行选择。

输出维 指定低维空间中的成分或维的数量。成分数必须大于或等于 2。

随机种子 指定一个随机种子，以便将来启动该平台时重现结果。

标准化 在计算用于降维的距离之前，在内部对数据进行标准化。

缺失值补缺 指定数据中使用多元奇异值分解 (SVD) 方法插补的缺失值。

注意：若您的数据包含缺失值且未选择 “缺失值插补” 选项，则在启动中点击 “确定” 后，将显示插补窗口。若数据中的每一行至少包含一个缺失值，则可以选择对缺失值进行插补、更改 Y 列的选择或取消分析。若数据中的某些行不包含缺失值，则可以选择对缺失值进行插补，继续而不插补或取消分析。

UMAP 选项 包含 UMAP 算法中使用的选项。有关在 UMAP 算法中如何使用以下参数的详细信息，请参见 McInnes et al.(2018)。

近邻数 指定为每个数据点找到的近邻数。指定的近邻数目越小，UMAP 算法就越集中于数据的局部结构。随着近邻数的增加，UMAP 算法会更多地捕获数据的全局结构。“近邻数”值的范围可介于 2 到数据中观测数的四分之一之间。默认值为 15。

时期数 指定优化低维表示时要使用的训练时期数。这是算法在整个训练数据中工作的次数。默认值为 500。

学习率 指定计算中的学习率的值。默认值为 1。学习率影响模型适应问题的速度。若学习率太大，算法可能会错过最优解。若学习率太小，算法可能会需要较长时间收敛。

提示：若算法不收敛或不生成具有极值的嵌入坐标，则考虑调整学习率的值。

最小距离 指定低维空间中的各点之间的最小标准化距离。该值可介于 0 到 0.99 之间。默认值为 0.01。

局部连通性 指定假定在局部级别连接的最近邻的数量。默认值为 1，这假设高维空间中的每个点都至少有一个与其连接的其他近邻。

A 指定控制嵌入优化算法的参数之一。若该值被指定为 0 或负数，则在算法中通过非线性最小二乘法过程计算 a。

b 指定控制嵌入优化算法的参数之一。若该值被指定为 0 或负数，则在算法中通过非线性最小二乘法过程计算 b。

负抽样率 指定在查找数据的低维表示时，每个正 1-单纯形样本要使用的负 1-单纯形样本数。“负抽样率”值可介于 2 到 20 之间。默认值为 5。

批处理模式，若数目大于 指定当样本大小大于指定数时，使用多线程优化嵌入坐标。默认值为 409判别主成分。

最近邻方法 指定用于查找最近邻的方法。

默认 根据样本大小和变量数选择最近邻方法。若观测数大于 409 判别主成分 且变量数小于或等于 1500，或者“距离量度”未设置为“欧氏”距离，则默认值为“ANNOY”。若非如此，默认值为 “VPTree”。

VPTree （精确） 使用有利点 (VP) 树查找一组最近邻。

ANNOY （近似） 使用 “近似最近邻” (ANN) 方法查找一组最近邻 (Bernhardsson, 2013)。这是两种用于大型数据集的方法中较快速的一种，但结果可能不如 VPTree 方法准确。

距离量度 （仅当将“ANNOY”指定为“最近邻方法”时才适用。）指定用于计算最近邻之间距离的量度。用于距离量度的选项包括“欧氏”、“角”、“Hamming”和“Manhattan”。默认情况下，“欧氏”被指定为“距离量度”。

提示：若数据包含二进制或分类变量，则非欧氏距离量度可能更合适。

梯度下降方法 指定优化算法中使用的梯度下降方法。

SGD 使用 “随机梯度下降”算法 (Saad, 1998)。这是默认方法。

ADAM 使用“自适应矩估计”方法 (Kingma, 2014)。仅当使用多线程时该选项才可用。

t-SNE 选项 包含 t-SNE 算法中使用的选项。“‘多元嵌入”平台的统计详细信息”中对其中很多选项进行了讨论。

稀疏 指定在计算高维空间中的条件概率时是否使用稀疏方法。稀疏方法支持计算高维数据。

困惑度 指定困惑度参数的值，该值与计算样本的相似性有关。困惑度参数的值应介于 5 和 50 之间，并且不应大于样本大小的八分之一。默认值是 30 或样本大小的八分之一中较小的那一个。

最多迭代次数 指定计算中使用的最大迭代次数。

初始主成分维度 指定初始随机主成分分析步骤中保留的维数。默认值为 50。

收敛准则 指定用于测量收敛程度的值。默认值为 1e-8。

初始尺度 指定推导出的成分的初始尺度。默认值为 0.0001。

Eta 指定计算中的学习率的值。默认值为 200。

膨胀迭代 指定迭代次数，在该次数之后就不再放大动量值。默认值为 250。

保持对话框打开 运行分析后保持启动窗口打开，以便您更新选项并重新运行分析。

JMP PRO “多元嵌入”报表

“多元嵌入”报表窗口包含一个图以及一个或多个表。该报表支持您直观演示低维空间，并查看算法参数和计算细节。

报表中图的形式取决于启动窗口中“输出维”选项中指定的维数。若指定两个维，则会显示“t-SNE 图”或“UMAP 图”报表。若指定三个或更多维，则会显示“三维图”报表。这些报表分别显示每个观测的前两个或三个嵌入成分得分的二维或三维散点图。

参数表包含在启动窗口中为 UMAP 方法或 t-SNE 方法指定的参数值。请参见“UMAP 选项”和“t-SNE 选项”。

若在启动窗口中指定了 t-SNE 方法，则报表中还会显示 “迭代历史记录”表。该表包含计算迭代的统计量。每行表示 100 次连续迭代结束时的值。“迭代历史记录”表包含以下列：

迭代 迭代编号。

成本 相应迭代处的成本函数的值。这是所有数据点上 Kullback-Leibler 散度的总和。

最大梯度 相应迭代处所有数据点上的梯度的最大值。

有关“迭代历史记录”表中列的详细信息，请参见“梯度下降算法的统计详细信息”。

JMP PRO “多元嵌入”平台选项

“多元嵌入”红色小三角菜单包含以下选项：

保存嵌入成分值 将推导出的 t-SNE 或 UMAP 成分保存为数据表中的新列。

显示 PQ 矩阵 （仅当使用 t-SNE 方法且未在启动窗口中选择“稀疏”选项时才可用。）创建两个新数据表，其中包含 t-SNE 计算中使用的 **P** 和 **Q** 量度。请参见 “t-SNE 方法的统计详细信息”。

请参见《使用 JMP》获取有关下列选项的信息：

本地数据过滤器 显示或隐藏支持您过滤特定报表中使用的数据的本地数据过滤器。

重新运行 包含使您可以重复或重新启动分析的选项。在支持该功能的平台中，“自动重新计算”选项立即在相应报表窗口中反映您对数据表所做的更改。

平台首选项 包含的选项支持您查看当前平台首选项或更新平台首选项以匹配当前 JMP 报表中的设置。

保存脚本 包含的选项支持您保存可将报表重现到若干目标的脚本。

保存 “依据”组脚本 包含使您可以保存脚本的选项，可将为“依据”变量的所有水平重新生成平台报表的脚本保存到多个不同的位置。仅当在启动窗口中指定“依据”变量时才可用。

JMP PRO 多元嵌入的更多示例

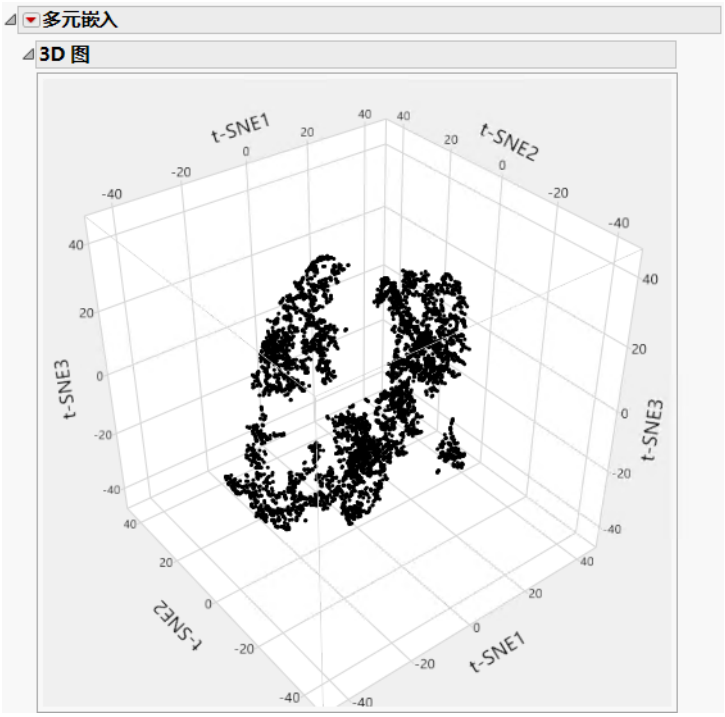
在本例中，您使用 “多元嵌入的示例”中的相同数据集，但将数据降至三维而不是二维。您还在 UMAP 方法和 t-SNE 方法之间比较了结果。

1. 选择**帮助 > 样本数据文件夹**，然后打开 Cytometry.jmp。
2. 选择**分析 > 多元方法 > 多元嵌入**。
3. 从 CD3 一直选择到 MCB，然后点击 **Y，列**。
4. 将 t-SNE 指定为 “方法”。
5. 在 “输出维”旁边键入 3。
6. （可选。）在 “随机种子”旁边键入 1234。

注意：输入上面的种子可以重现本示例中所示的结果。

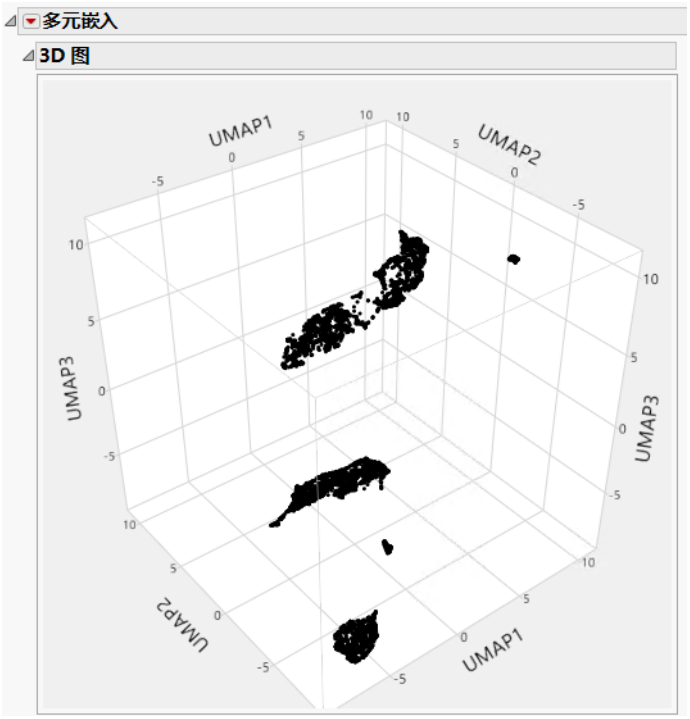
7. 选择**保持对话框打开**。
8. 点击**确定**。

图 11.4 三维 t-SNE 图的示例



9. 在启动窗口中，将 “方法”改为 “UMAP”。
10. 点击**确定**。

图 11.5 三维 UMAP 图的示例



每个三维图都显示了相应降维方法的前三个嵌入成分的坐标。将 UMAP 与 t-SNE 方法加以比较后发现，UMAP 方法所用的计算时间较少，同时也提供了更紧密的数据聚类。与图 11.4 中的三维 t-SNE 图相比，三维 UMAP 图中可见更多明显的聚类。

提示：您可以旋转该网格以便更清晰地查看聚类。

JMP PRO “多元嵌入”平台的统计详细信息

本节包含 “多元嵌入”平台的统计详细信息。

- “t-SNE 方法的统计详细信息”
- “梯度下降算法的统计详细信息”

t-SNE 方法的统计详细信息

通过将 $\{x_i, x_j\}$ 的高维相似性与 $\{y_i, y_j\}$ 的低维相似性之间的差异最小化，t-SNE 方法将高维空间 $\{x_1, x_2, \ldots, x_n\}$ 中的点映射到低维空间 $\{y_1, y_2, \ldots, y_n\}$ 中的点。配对相似性表示为概率分布。在高维空间中，使用高斯分布计算条件概率 $p_{j|i}$。“多元嵌入”平台提供两种计算条件概率的方法。

条件概率的稀疏近似计算

若在启动窗口中选定“稀疏”选项，则使用稀疏近似来计算 $p_{j|i}$。对于 n 个输入中的每一项，都使用有利点 (VP) 树发现一组最近邻。然后，仅为最近邻的那些子集计算条件概率：

$$p_{j|i} = \begin{cases} \dfrac{\exp(-\left\|x_i - x_j\right\|^2/(2\sigma_i^2))}{\sum_{k \in N_i} \exp(-\left\|x_i - x_k\right\|^2/(2\sigma_i^2))}, & \text{if } j \in N_i \\ 0 & \text{否则} \end{cases}$$

在该等式中，N_i 是 x_i 的 floor(3p) 最近邻集合，其中 p 是在启动窗口中定义的困惑度参数。高斯分布的方差 σ_i 也基于该困惑度参数。请参见 van der Maaten and Hinton (2008) 和 van der Maaten (2014)。

条件概率的非稀疏计算

若在启动窗口中未选定 “稀疏”选项，则为所有点计算 $p_{j|i}$：

$$p_{j|i} = \frac{\exp(-\left\|x_i - x_j\right\|^2/(2\sigma_i^2))}{\sum_{k \neq i} \exp(-\left\|x_i - x_k\right\|^2/(2\sigma_i^2))}$$

在该计算中，高斯分布的方差 σ_i 也基于该困惑度参数。

联合概率分布的计算

在 t-SNE 方法中，假定条件概率是对称的。因此，高维空间中的联合概率 p_{ij} 由对称条件相似性定义：

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

其中，对于所有 i 和 j，$p_{ij} = p_{ji}$。由于关注的是配对相似性，因此还假设 $p_{ii} = 0$。

使用自由度为 1 的 Student t 分布计算低维映射中的联合概率 q_{ij}：

$$q_{ij} = \frac{(1 + \left\|y_i - y_j\right\|^2)^{-1}}{\sum_{k \neq l} (1 + \left\|y_k - y_l\right\|^2)^{-1}}$$

这些概率具有与 p_{ij} 相同的属性，这意味着对于所有 i 和 j，$q_{ij} = q_{ji}$，并且 $q_{ii} = 0$。

通过最小化联合概率分布 **P** 与联合概率分布 **Q** 之间的单个 Kullback-Leibler 散度，t-SNE 方法将高维空间中配对相似性与低维空间中配对相似性之间的差异最小化。**P** 与 **Q** 之间的 Kullback-Leibler 散度计算如下：

$$\mathrm{KL}(\mathbf{P} \parallel \mathbf{Q}) = \sum_i \sum_j p_{ij} \log(p_{ij} / q_{ij})$$

梯度下降算法的统计详细信息

使用基于 Barnes-Hut 近似 (van der Maaten, 2014) 的梯度下降算法最小化 t-SNE 方法中使用的 Kullback-Leibler 散度。该算法使用以下符号：

$\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ 包含 n 个数据点的原始高维数据

p = 困惑度参数

T = 迭代次数

η = 学习率

$t_{inflate}$ = 迭代次数，在该次数之后动量值发生更改

$\alpha(t)$ = 迭代 t 处的动量，其中 $\alpha(t) = 0.5$ （对于 $t \leq t_{inflate}$），其他情况下 $\alpha(t) = 0.8$

$\mathbf{Y}^{(t)} = \{y_1, y_2, \ldots, y_n\}$ 迭代 t 处的低维映射解

梯度下降算法的步骤定义如下：

1. 计算 $p_{j|i}$，高维空间中的配对相似性。基于指定的困惑度 p 选择 σ_i。
2. 计算 p_{ij}。
3. 设置初始解 $\mathbf{Y}^{(0)} = \{y_1, y_2, \ldots, y_n\}$，初始解从均值为 0 且标准差为 10^{-4} 的正态分布生成。
4. 对于 $t = 1$ 到 T：
   - 计算 q_{ij}，低维映射中的配对相似性。
   - 计算成本函数：

     $$C = \mathrm{KL}(\mathbf{P} \parallel \mathbf{Q}) = \sum_i \sum_j p_{ij} \log(p_{ij} / q_{ij})$$

   - 使用 Barnes-Hut 近似计算梯度函数：

     $$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + \left\| y_i - y_j \right\|^2)^{-1}$$

   - 更新解：

$$\mathbf{Y}^{(t)} = \mathbf{Y}^{(t-1)} + \eta\left(\frac{\partial C}{\partial \mathbf{Y}}\right) + \alpha(t)(\mathbf{Y}^{(t-1)} - \mathbf{Y}^{(t-2)})$$

若满足以下条件之一，算法即停止：

- i 上的最大梯度值小于启动窗口中指定的收敛准则。
- 达到最大迭代次数 T。

第 12 章

项目分析

按项目和对象分析测验结果

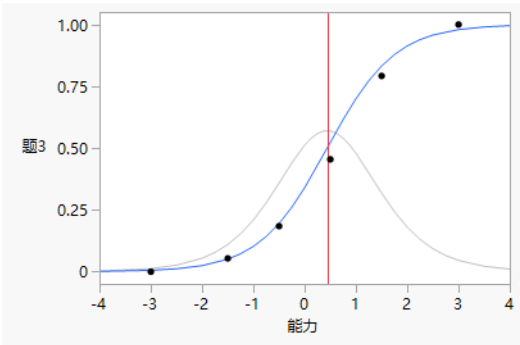
“项目分析”平台支持您拟合**项目反应理论**模型。项目反应理论 (IRT) 方法用于对测量手段（如：测试和问卷）进行分析和评分。项目反应理论 (IRT) 使用一系列等式将项目关联到未观测到的（潜在）特征或能力。项目（或问题）是无法直接观测到的底层潜在构造的指示符。收集数据时，对象能力和项目特征均未知。IRT 可用于研究标准化测验、认知发展和消费者偏好。IRT 是经典测验理论 (CTT) 的替代方法，CTT 侧重于观测得分合计而不是项目得分。

“项目分析”平台执行 IRT 方法，该方法生成以下结果：

- 在项目级别上对测量手段评分，可深入了解每个项目在潜在响应中的贡献。
- 在同一尺度上生成响应者和项目的得分。
- 响应者和项目得分显示在单个图中。
- 显示项目特征曲线。这些曲线可用于探索项目与响应者的潜在特征或能力之间的关系。

有关项目反应理论的详细信息，请参见 de Ayala (2009)。

图 12.1 项目分析特征图



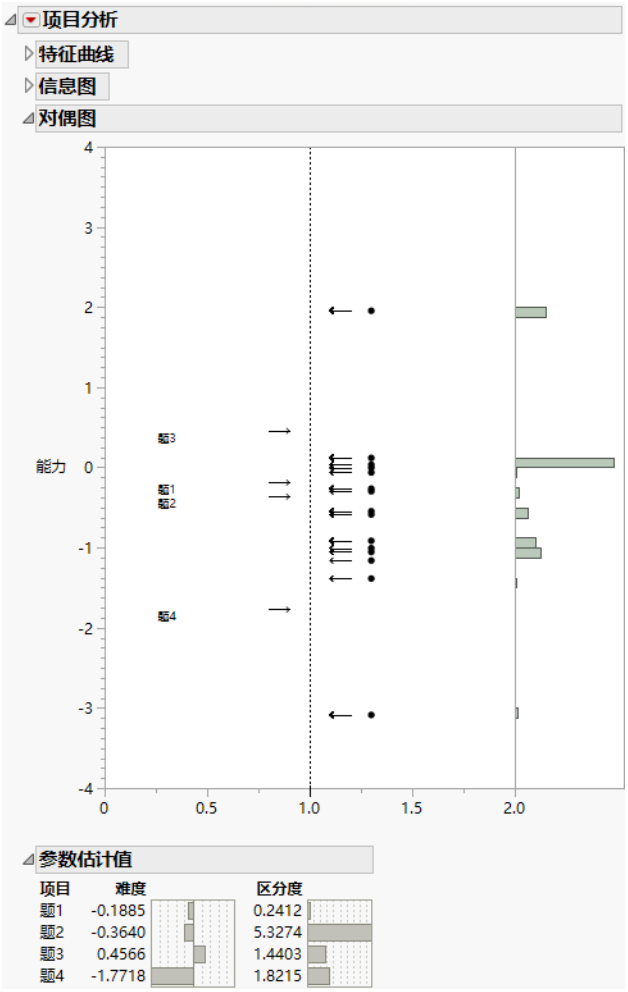
目录

项目分析的示例

在本例中，您使用 “项目分析”平台来了解测验问题与科目能力之间的关系。本例使用关于 12 判别主成分 3 个学生在 14 个问题上的得分 （1 = 正确， 0 = 不正确）。测验中的问题即用于测量潜在数学能力的项目。您使用 2PL 模型检查前四个问题，来了解答题者的数学能力与这四个问题之间的关系。

1. 选择**帮助** > **样本数据文件夹**，然后打开 MathScienceTest.jmp。
2. 选择**分析** > **多元方法** > **项目分析**。
3. 从**题 1** 一直选择到**题 4**，点击 **Y，测验项目**，然后点击**确定**。

图 12.2 项目响应报表

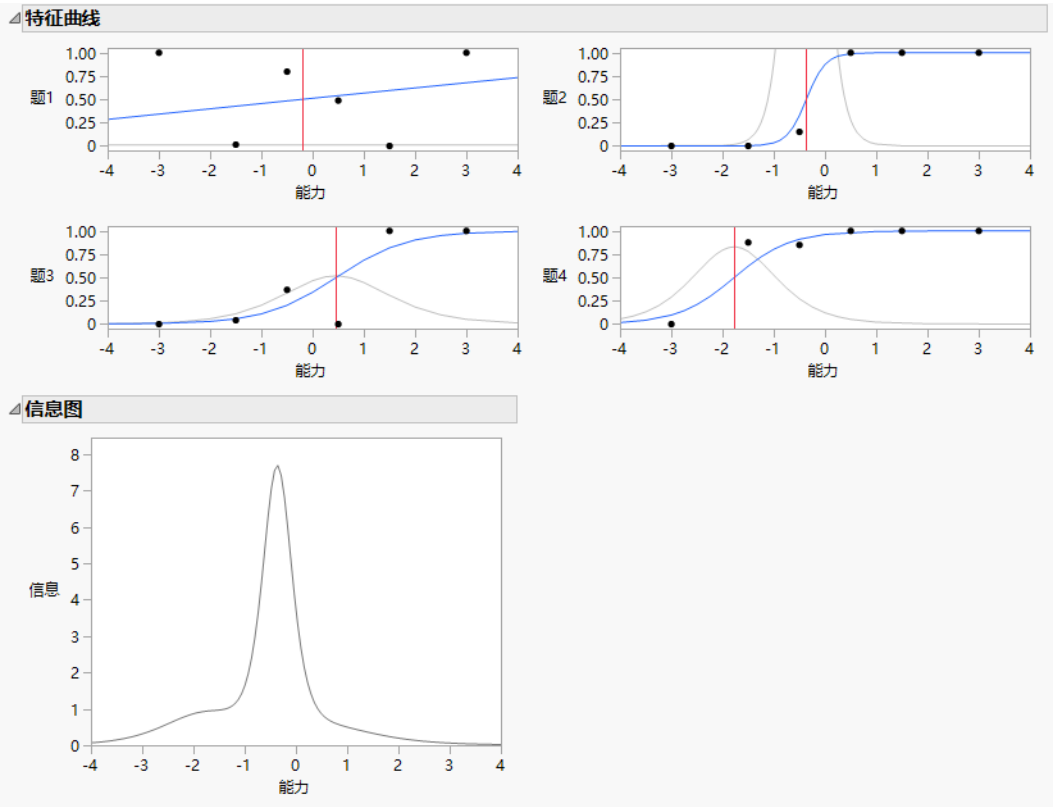


从对偶图中，您会注意到：题 4 是四个问题中最容易回答的问题，这是因为它具有最低的难度得分 -1.78。题 3 是最难的问题，其难度得分为 0.4 判别主成分。大多数响应者都落在能力尺度的中间到低端，如图中心部分的数据点所示。在该直方图中，您可以看到大约 40% 的响应者落在稍微高于能力尺度上的 0 的位置。

注意：不为答案全部正确或全部不正确的个人计算能力得分。请参见 “拟合 IRT 模型的统计详细信息”。

4. 点击灰色的 “特征曲线” 报表展开图标将其打开。
5. 点击 “项目分析” 红色小三角并选择**横向图数**。
6. 输入 2 并点击**确定**。
7. 点击灰色的 “信息图” 报表展开图标将其打开。

图 12.3 项目响应示例



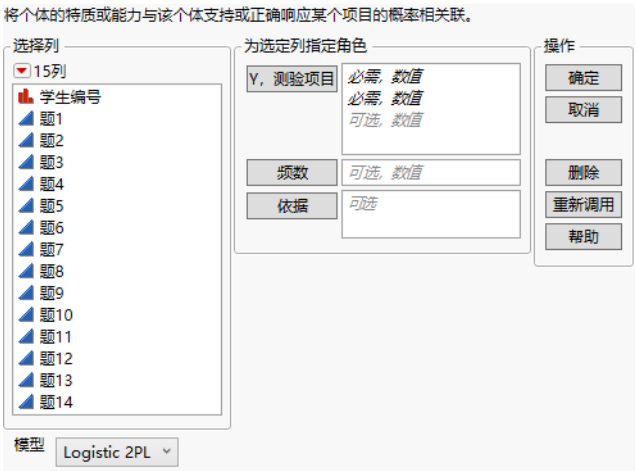
题1 具有平直的特征曲线和平直的信息曲线。这表明题1 没有为区分响应者的数学能力提供太多信息。题2 的特征曲线比较陡峭，这表明题2 对于区分响应者能力很有用。每个图中的垂线都位于特征曲线的拐点。这条垂线即响应者有 50% 的概率正确回答指定问题的能力水平。

信息图指示所分析的四个问题共同提供了关于 -1（大约）到 0 之间的能力水平的多数信息。在模型中包括更多难度更大的问题将会加宽信息曲线。

启动 “项目分析” 平台

通过选择**分析 > 多元方法 > 项目分析**启动 “项目分析” 平台。

图 12.4 “项目分析” 启动窗口



有关 “选择列” 红色小三角菜单中选项的详细信息，请参见《使用 JMP》。

Y，测验项目 分配两个或更多要分析的列。这些列必须是数值和连续型，并且仅包含 0 和 1。

提示：若需要将数据重新编码为 0 和 1，请使用**列 > 重新编码**。请参见 《使用 JMP》。

频数 指定一个频数变量。该选项适用于汇总数据。

依据 为 “依据” 变量的每个水平生成单独报表。若分配了多个 “依据” 变量，则为 “依据” 变量水平的每个可能组合生成单独报表。

模型 通过以下选项指定所需模型：

Logistic 2PL 2 参数 Logistic 模型。

Logistic 3PL 3 参数 Logistic 模型。

Logistic 1PL 带 Rasch 参数化的单参数 Logistic 模型。

Logistic 3PL 模型详细信息

在 “项目分析” 启动窗口中，若为模型选择 Logistic 3PL，则系统会提示您在点击 “确定” 后为猜测参数输入一个罚值。对于 3PL 模型，罚值的默认值为 0。不过，您可以为 *c* 参数 （凭猜

测答对某题的概率）输入非零罚值。该罚值类似于您在岭回归中使用的罚值参数的类型。该罚值与估计猜测参数的方差有关。使用罚值有以下优点：

- 稳定模型参数的估计。
- 加快计算。
- 减少各项之间的猜测参数的变异性，但会带来一些偏倚。

较大的罚值会强制猜测参数为 0，而较小值则会帮助减少各项之间的猜测参数的变异性。值为 0 可用于表示无罚值。

数据格式

“项目分析”平台要求数据表中每个单值对应一行，每个项目对应一列。项列必须是数字列，并且只包含 0 和 1，分别指示不正确或正确的响应。MathScienceTest.jmp 样本数据表演示了针对 14 个测试问题的 1,2判别主成分3 个人的项目响应分析所要求的数据格式。

“项目分析”报表

“项目分析”报表包含以下部分。

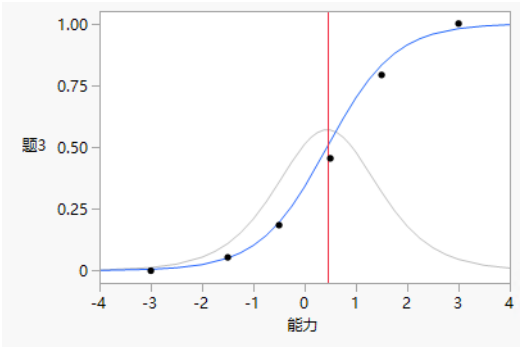
- “特征曲线”
- “信息图”
- “对偶图”
- “参数估计值”

特征曲线

在“项目分析”报表中，“特征曲线”部分包含您在启动窗口中指定的每个项的项目特征曲线 (ICC)。“特征曲线”部分最初是关闭的。

项目特征曲线绘制根据能力正确回答某个项目的概率。能力的测量使用了标准化的尺度，所以能力为 0 的响应者即具有平均能力的响应者。固定能力水平的正确回答的观测概率数据点将会绘制出来。将拟合特征曲线与数据点作比较为每个单独的项提供了直观的模型拟合优度测度。此外，特征图还包含一条背景信息曲线以及位于特征曲线拐点处的垂线。背景信息曲线是项目特征曲线的斜率图，在拐点处最大化。

图 12.5 项目特征曲线

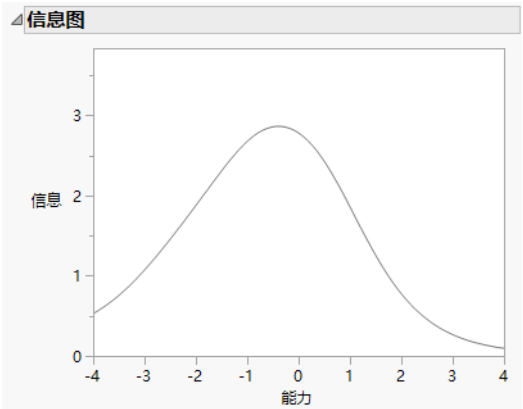


提示：使用“项目分析”红色小三角菜单中的“横向图数”选项，您可以调整报表每行中显示的特征曲线数。

信息图

在“项目分析”报表中，“信息图”部分包含整体信息曲线图，通过加总各个项的信息曲线可构造整体信息曲线图。可以通过信息图洞察该测验能够测量的正确能力水平。图 12.6 描述了这样一个测验，测验项目更适用于评估具有平均到低能力水平的个人，而不是具有高能力水平的个人。该图最初处于关闭状态。

图 12.6 信息图

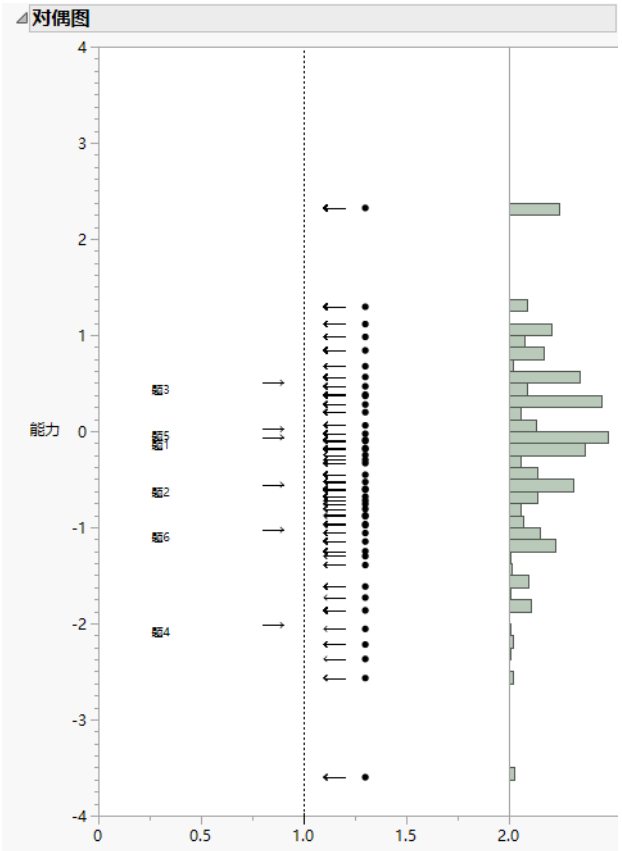


对偶图

在“项目分析”报表中，“对偶图”部分包含一个图，这一个图中既显示项目难度又显示对象能力。难度和能力以相同的标准化度量尺度显示在垂直轴上。项目按照其难度标绘在图的左

侧。对象绘制在右侧，随同显示的还有数据点和一个直方图。对偶图支持您将每个项目的难度关联到每个响应者的能力。

图 12.7 对偶图



参数估计值

在“项目分析”报表中，“参数估计值”部分包含每个项的估计参数表。提供的参数取决于分析中使用的模型（1PL、2PL 或 3PL）。

项目 测验项目。

难度 *b* 参数或项目难度的测度。难度参数的直方图显示在难度估计值旁边。

分辨力 （仅适用于 2PL 和 3PL 模型。）*a* 参数或项目分辨力的测度。分辨力参数的直方图显示在分辨力估计值旁边。

下渐近线 （仅适用于 3PL 模型。） *c* 参数或猜测测度。

“项目分析”平台选项

“项目分析”红色小三角菜单包含以下选项：

横向图数 支持您指定在“特征曲线”报表中的每行图中显示多少个 ICC 图。默认设置为每行显示一个 ICC 图。

保存能力公式 在数据表的新列中保存能力公式。该选项使用 `IRT Ability()` JSL 函数。有关该函数的详细信息，请参见“帮助”>“脚本索引”。

请参见《使用 JMP》获取有关下列选项的信息：

重新运行 包含使您可以重复或重新启动分析的选项。在支持该功能的平台中，“自动重新计算”选项立即在相应报表窗口中反映您对数据表所做的更改。

平台首选项 包含的选项支持您查看当前平台首选项或更新平台首选项以匹配当前 JMP 报表中的设置。

保存脚本 包含的选项支持您保存可将报表重现到若干目标的脚本。

保存“依据”组脚本 包含使您可以保存脚本的选项，可将为“依据”变量的所有水平重新生成平台报表的脚本保存到多个不同的位置。仅当在启动窗口中指定“依据”变量时才可用。

“项目分析”平台的统计详细信息

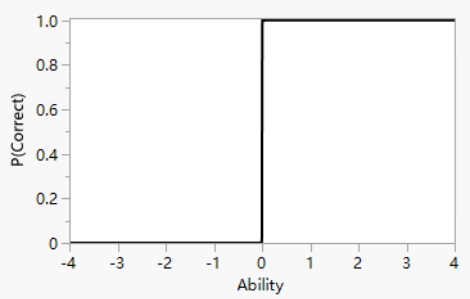
该部分包含“项目分析”平台的统计详细信息。

- “项目响应曲线的统计详细信息”
- “项目响应曲线模型的统计详细信息”
- “IRT 模型假设的统计详细信息”
- “拟合 IRT 模型的统计详细信息”
- “能力公式的统计详细信息”

项目响应曲线的统计详细信息

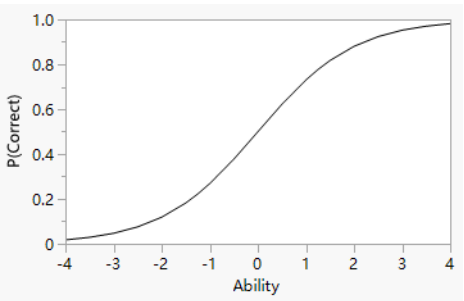
“项目分析”平台显示项目响应曲线（项目特征曲线）。这些曲线用于描述能力（在**能力**尺度上定义）与每个项目之间的关系。项目响应曲线绘制针对不同能力水平正确回答某个项目的概率。对于完美分辨力的项目，能力低于阈值的响应者答对的概率为 0%，而能力高于阈值的响应者答对的概率为 100%。

图 12.8 有极佳分辨力的项目的特征曲线



正确回答某个项目的概率与能力之间的典型关系是一个具有下渐近线和上渐近线的 S 形函数。随着响应者能力的增加，其正确回答该项目的概率将增至 100%。特定项目的曲线形状与该项目的难度和分辨力属性相关。

图 12.9 典型的项目响应曲线



项目响应曲线模型的统计详细信息

“项目分析”平台提供一个、两个和三个参数的 logistic 模型来对项目响应曲线建模。 3 参数 Logistic (3PL) 模型定义如下。

$$P(\theta) = c + \frac{1-c}{1+e^{-(a)(\theta-b)}}$$

- $P(\theta)$ 是针对能力水平 θ 正确回答该项目的概率。有关拟合项目反应理论模型的详细信息，请参见“拟合 IRT 模型的统计详细信息”。
- a 参数定义曲线拐点处的陡度。它提供该项目分辨力的估计值。
- b 参数定义拐点在能力轴上的位置。它提供该项目难度的估计值。
- c 参数是下渐近线。它提供通过猜测正确回答某项目的概率估计值。
- 对于 2PL 模型，c 参数设置为 0。

$$P(\theta) = \frac{1}{1 + e^{-(a)(\theta - b)}}$$

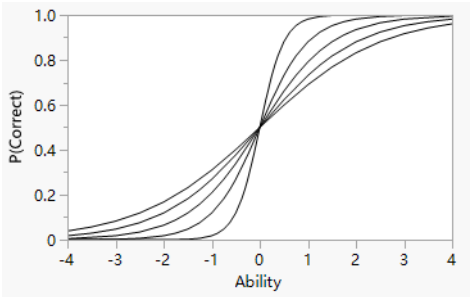
- 对于 1PL 模型，*c* 参数设置为 0，*a* 参数设置为 1。这种参数化亦称 Rasch 模型 (Rasch, 1980)。

$$P(\theta) = \frac{1}{1 + e^{-(\theta - b)}}$$

a 参数：项目分辨力

在 2PL 和 3PL 模型中，*a* 参数（或曲线拐点处的陡度）提供项目分辨力的测度。项目分辨力是指项目可以区分低能力水平响应者与高能力水平响应者的能力。陡峭的项目响应曲线指示该项目具有很强的分辨力。低能力水平响应者正确响应项目的概率较低；高能力水平响应者正确响应项目的概率较高。曲线相对平直的项目具有较低的分辨力。分辨力较低的项目可作为候选项从测量手段中删除。

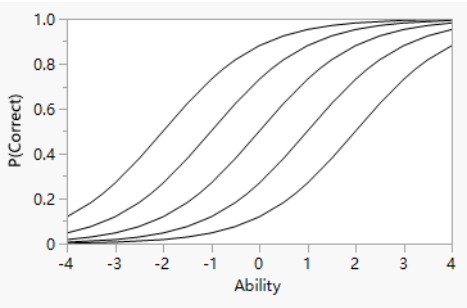
图 12.10 不同 *a* 值的 Logistic 模型



b 参数：项目难度

b 参数（或相对于能力的拐点位置）提供了项目难度测度。拐点在能力尺度上靠右的项目响应曲线指示项目比起拐点靠左的项目更难回答。在 1PL 和 2PL 模型中，*b* 参数是有 50% 的概率正确回答项目所需的能力水平的估计值。

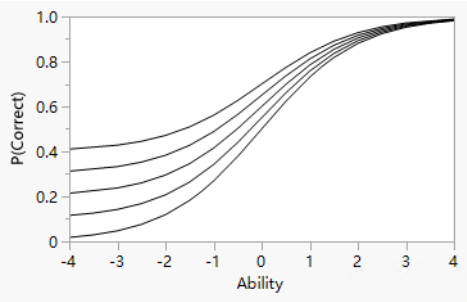
图 12.11 不同 *b* 值的 Logistic 曲线



c 参数：猜测

在 3PL 模型中，*c* 参数（或项目响应曲线的下渐近线）提供猜测参数的测度。非零下渐近线表示能力水平极低的个人正确回答项目的概率不为零。

图 12.12 不同 *c* 值的 Logistic 模型



IRT 模型假设的统计详细信息

在“项目分析”平台中，2PL 模型是默认模型。当您可以假设所有项目都具有相同的分辨力时，1PL 模型适用。若该假设不正确，则应使用 2PL 或 3PL 模型。2PL 模型比 3PL 模型的数字稳定性强，特别是对于小型数据集。此外，在 2PL 模型中，*b* 可解释为响应者 50% 的概率正确回答项目所需的能力水平。

IRT 模型假定底层特征是一维的。也就是说，只有一个底层的潜在构造。若存在多个特征，而要测量的每个特征相互之间存在复杂的交互作用，这种情况下一维模型就不适合了。IRT 模型适用于连续潜在变量。对于分类潜在变量，您应考虑潜在类模型。请参见“潜在类分析”。IRT 模型假定项目不变。项目不变性意味着 $P(\theta)$ 解释为一组具有能力水平 θ 的个体的答对概率。 若一大群具有相同能力水平的个体回答了该项目，则 $P(\theta)$ 预测的是正确回答该项目的人的比例。这意味着 IRT 模型具备项目参数的不变性，不管是什么样的测验群体，您都会得到同样的参数估计。此外，IRT 模型假定具备局部独立性，这意味着一旦解释了潜在构造，项目就彼此独立。

拟合 IRT 模型的统计详细信息

在“项目分析”平台中，使用“边缘最大似然估计”(MMLE) 拟合项目响应理论模型。MMLE 是“联合最大似然估计”(JLE) 的备选方法。 MMLE 将对象视为随机效应。项目和能力作为条件概率彼此相关。公式定义如下：

$$p(\mathbf{x}|\theta, \vartheta) = \prod_{j=1}^{L} p_j(\theta)^{x_j}(1-p_j(\theta))^{1-x_j}$$

其中，$p(\mathbf{x}|\theta, \vartheta)$ 是在给定对象能力 θ 和项目参数的向量 ϑ 时，响应向量**x** 的概率。项目参数的数目取决于使用的模型（1PL、2PL 或 3PL）。

MMLE 使用高斯求积分方法对对象效应求积分，以此获取项目参数估计值。响应向量 **x** 的概率计算如下：

$$p(\mathbf{x}) = \int_{-\infty}^{\infty} p(\mathbf{x}|\theta, \vartheta)g(\theta|\upsilon)d\theta$$

其中，$g(\theta|\nu)$ 是对象分布，ν 是总体位置和尺度参数的向量。在 JMP 中，均值为 0、标准差为 1 的正态分布用于 $g(\theta|\nu)$。

注意：针对测验问题的缺失值被视为错误响应。不为答案全部正确或全部不正确的个人计算能力得分。这些测试对象的响应模式包括在模型估计中。

用于拟合 IRT 模型的 MMLE 过程可与在两个阶段中拟合随机效应模型作比较。能力参数被视为方差为 1 的随机效应。在第一步中，使用高斯求积分对这些随机效应求积分。项目参数被视为使用边缘似然（已对能力参数求积分）的 ML 估计的固定效应。能力参数实际上是使用完全未积分（联合）似然估计的最佳线性无偏预测，且估计时项目参数已知，并且保持在第一阶段中获取的值。

对于 L 个项目，有 2^L 个响应模式。每个模式的能力水平可以通过计算在该响应模式下出现最高概率的能力水平，可通过应用以下公式直到 θ 进行计算。

$$\theta_i^{t+1} = \theta_i^{t} - \frac{X_i - \sum_{j=1}^{L} p_{ij}(t)}{-\sum_{j=1}^{L} p_{ij}(t)(1-p_{ij}(t))}$$

其中：

θ 将获取响应模式的似然最大化

t 是迭代次数

L 是项目数

X_i 是观测到的得分

p_{ij} 是在特定的项目参数下，第 i 个人答对第 j 个项的概率。

能力公式的统计详细信息

在“项目分析”红色小三角菜单中，“保存能力公式”选项将能力公式保存到数据表的新列中。该公式可用于为添加到数据表中的其他测试对象评分，也可以将其复制到新表中为新的一组测试对象评分。

保存到数据表的函数称为 `IRT Ability` 函数。项目参数估计值以矩阵的形式存储在该函数中。有关该函数的详细信息，请参见“帮助”>“脚本索引”。

第 13 章

层次聚类

使用聚类树将观测分组

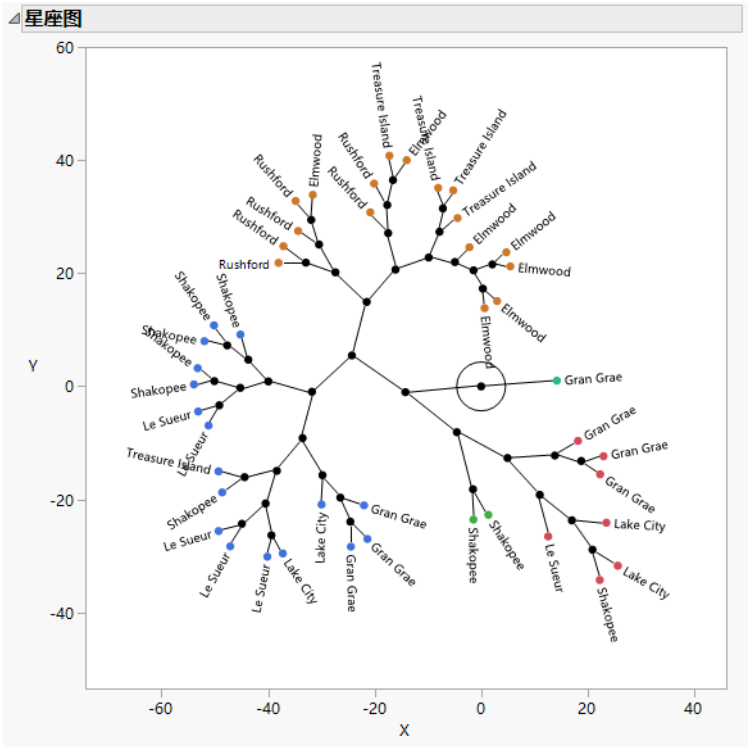
聚类是将在几个变量上享有相似值的观测分组在一起的一种多元方法。它可用于您理解数据的聚簇结构。

“层次聚类”依次组合聚类。该方法首先将每个观测视为一个聚类。然后逐步将距离上最相近的两个聚类合并成一个聚类。结果被描绘成一棵树，称为**系统树图**。

通常，层次聚类适用于不多于数万行的小数据表。该算法占用较长时间，因此大数据表运行会很缓慢。不过，“层次聚类”平台还提供两种方法，“快速 Ward”和“混合 Ward”，这两种方法减少了计算时间并可用于聚类更大的数据表。

注意：“层次聚类”支持字符列；“K 均值聚类”或“正态混合”要求使用数值列。

图 13.1 星座图的示例



目录

“层次聚类”平台概述

层次聚类方法首先将每个观测作为一个聚类。在每一步中，聚类过程会计算各对聚类之间的距离，并将两个相距最近的聚类组合起来。该组合过程会一直进行到所有点都位于一个聚类中。层次聚类亦称为**自下而上聚类**，因为它使用的是一个组合方法。

自下而上的过程被描绘为一棵树，称为系统树图。为了帮助您确定聚类数，JMP 会提供距离图。您可以通过确定聚类之间的距离何时不再具有实际意义的方式来选择聚类数。

层次聚类也支持字符列。可通过两种方式定义距离。

- 若列为有序型，则用于聚类的值就是有序类别的索引，将像处理连续数据那样处理有序类别。这些值将像处理连续数据那样进行标准化。
- 若列为名义型，则类别匹配的两个观测之间的距离为 0。若类别不同，则距离为 1。

“层次聚类”提供给您五个规则用于定义聚类之间的距离：类平均法、重心法、Ward 法、最短距离法和最长距离法。每个规则会生成不同序列的聚类。还有两种基于 Ward 法定义聚类之间距离的方法：“快速 Ward”和“混合 Ward”。

提示：层次聚类过程针对 n 个观测会从 $n(n+1)/2$ 个距离开始计算，但使用“快速 Ward”方法时除外。因此，当 n 较大时，该方法会运行较长时间。对于大量数值观测的情况，考虑使用“K 均值聚类”或“正态混合”。

“层次聚类”是 JMP 提供的对观测进行聚类的四个平台之一。有关四种方法的比较，请参见“对观测聚类的平台概述”。

对观测聚类的平台概述

聚类是将在几个变量上享有相似值的观测分组在一起的一种多元方法。通常情况下，观测在 p 维空间内散布不均，其中 p 是变量数。这些观测反而会形成聚簇或聚类。标识出这些聚类使您可以更深层次地了解您的数据。

注意：JMP 还提供可以对变量聚类的平台。请参见“聚类变量”。

JMP 提供四个平台供您对观测聚类：

- 层次聚类对于小型和大型数据表非常有用，并且允许使用字符数据。“层次聚类”将行按描绘为一棵树的层次序列形式进行组合。您可以在生成树后选择最适合您数据的聚类数。请参见“层次聚类”。
- “K 均值聚类”适用于多达数百万行的大型表，并且只允许数值数据。您需要提前指定聚类数 k。该算法可以对聚类种子点做出推测。随后开始在将数据点分配到相应类别和重新计算聚类中心之间交替进行迭代过程。请参见“K 均值聚类”。

- “正态混合”适用于数据来自重叠的多元正态分布的混合分布这种情况，并且只允许数值数据。对于具有多元离群值的情形，您可以使用假设具有均匀分布的离群值聚类。请参见“正态混合”。

  您需要提前指定聚类数。最大似然用于同时估计混合比例以及均值、标准差和相关性。为每个点指定属于每个组的概率。使用 EM 算法获取估计值。
- “潜在类分析”适用于大多数变量是分类变量这种情况。您需要提前指定聚类数。该算法拟合假定具有多项式混合分布的模型。为每个观测计算聚类成员关系的最大似然估计值。观测会被归类到其成员关系概率最大的聚类中。请参见“潜在类分析”。

表 13.1 聚类方法汇总

方法	数据类型或建模类型	数据表大小	指定聚类数
层次聚类	任意	使用混合 Ward 法，最多数十万行 使用快速 Ward 法，最多 200,000 行 使用其他方法，最多 5,000 行	否
K 均值聚类	数值	多达数百万行	是
正态混合	数值	任意大小	是
潜在类分析	名义型或有序型	任意大小	是

有些聚类平台提供用于处理数据中的离群值的选项。但是，若数据中有离群值，则最好在分析之前探索这些离群值。可以使用“探索离群值”实用工具完成该操作。有关详细信息，请参见《预测和专业建模》。

层次聚类示例

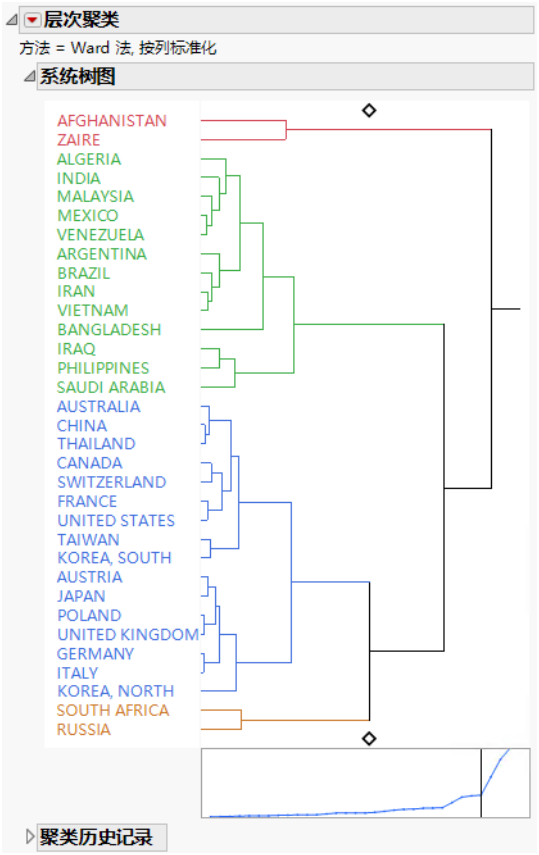
本例按照各国 2009 年每 1000 人的粗出生率和死亡率将各国分组，以便检查数据中的聚类。

1. 选择**帮助** > **样本数据文件夹**，然后打开 Birth Death Subset.jmp。
2. 选择**分析** > **聚类** > **层次聚类**。
3. 选择**出生率**和**死亡率**并点击 **Y, 列**。
4. 选择**国家**并点击**标签**。

   该选择可确保**国家**列 （而不是行号）用于对点击 “确定”后出现的系统树图添加标签。
5. 点击**确定**。

6. 点击“层次聚类”红色小三角并选择**聚类着色**。

图 13.2 “层次聚类”报表



该系统树图显示聚类是如何执行的。可以从左到右读取系统树图来查看聚类过程。每步包括将两个**最近**聚类组成一个聚类。

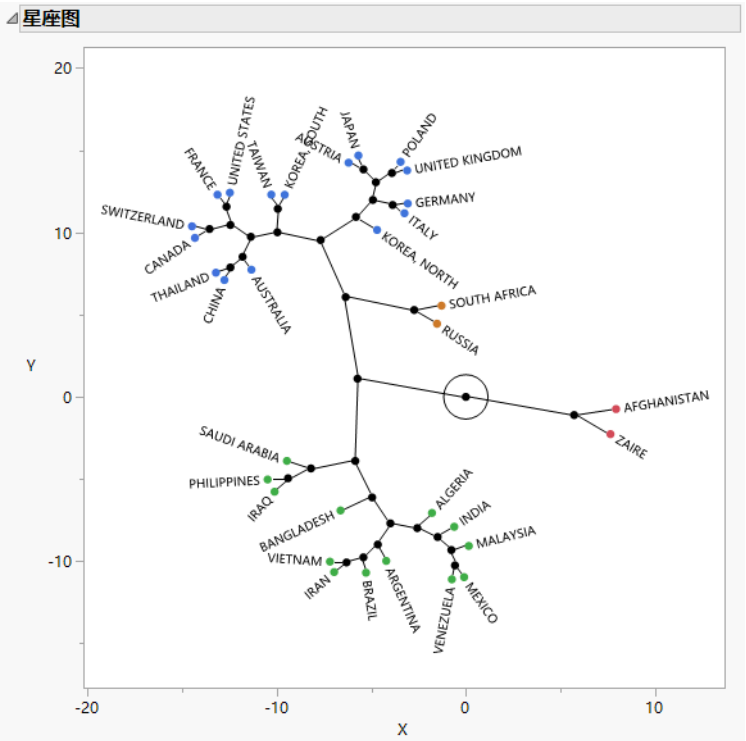
在系统树图中，聚类之间的相对距离由连接聚类的垂线之间的水平距离给出。例如，Afghanistan 和 Zaire 之间的距离大于 Malaysia 与 Mexico 和 Venezuela 组成的聚类之间的距离。

菱形设置在四个聚类处。最近连接起来形成四聚类模型的两个聚类是由 Algeria 到 Bangladesh 组成的聚类以及由 Iraq 到 Saudi Arabia 组成的聚类。这两个聚类之间的距离是菱形设置为 4 时距离图上由垂直线指示的点。该距离在“聚类数”等于 4 旁边的“聚类历史”报表中给出。此处显示该距离为 1. 判别主成分 187087 判别主成分 0 并且聚类从 Algeria 和 Iraq 开始组合形成四个聚类。

有四个聚类时，距离图斜率有明显变化。斜率变化指出在剩下四个聚类之前所连接的聚类之间的差异相对较小。这表明 4 是聚类数不错的选择。注意到这是默认显示的聚类数。

7. 点击“层次聚类”红色小三角并选择**星座图**。

图 13.3 星座图

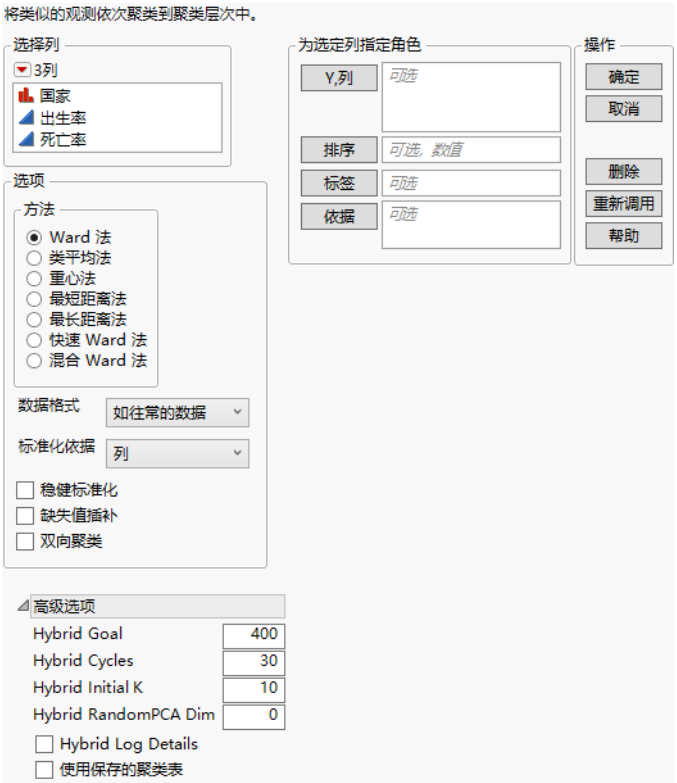


该星座图将国家显示为端点，将每个聚类连接显示为新点。各条线表示聚类中的成员关系。聚类连接之间的线长度近似等于连接的聚类之间的距离。星座图表明，包含阿富汗和扎伊尔的聚类与两个主要聚类中每个聚类的距离大致相同。

启动 “层次聚类” 平台

通过选择**分析** > **聚类** > **层次聚类**来启动 “层次聚类” 平台。

图 13.4 “层次聚类” 启动窗口



有关 “选择列” 红色小三角菜单中选项的详细信息，请参见《使用 JMP》。

Y，列 用于对观测聚类的变量。

排序 基于指定列按聚类的均值对其排序。

> **提示：**将通过执行主成分分析获取的第一主成分用作 “排序” 列。聚类按这些值排序。

特性 ID （仅当选择**数据被堆叠**作为数据结构时才可用。）指定堆叠的变量。

对象 ID （仅当选择**被汇总的数据**或**数据被堆叠**作为数据结构时才可用。 ）为堆叠了其测量值的每个单元提供唯一标识符的一列或多列。

标签 其值用于对报表中的系统树图添加标签的列。

注意：若选定的数据结构为**数据为距离矩阵**，“标签”列的数据类型必须为字符型。

依据 一列，其水平定义不同的分析。对于指定列的每个水平，都分析相应行。结果显示在不同的报表中。若分配了多个“依据”变量，则为“依据”变量水平的每个可能组合生成单独分析。

方法 指定用于计算距离以便定义聚类的方法。对于每种方法，将连接聚类，以使该方法定义的距离最小化。有关距离公式的信息，请参见“距离方法的统计详细信息”。

Ward 法 将两个聚类之间的距离定义为两个聚类之间的方差分析平方和（所有变量的方差分析和相加）。每一次生成聚类时，在通过合并上一次生成的两个聚类可能获得的所有划分中，类内平方和进行了最小化。当平方和除以总平方和时会得到方差比例（平方半偏相关），这样可以更容易地解释该平方和。

在假定使用多元正态混合、球形协方差矩阵和等抽样概率的前提下，Ward 法通过将层次每个水平下的似然最大化来生成聚类。

Ward 法倾向于生成有少量观测的聚类，而且特别偏向于生成具有大致相同观测数的聚类。它还对离群值非常敏感。请参见 Milligan (1980)。

类平均法 将两个聚类之间的距离定义为观测对之间的平均距离。类平均法倾向于连接具有较小方差的聚类，并且略微偏向于生成具有相同方差的聚类。请参见 Sokal and Michener (1958)。

重心法 将两个聚类之间的距离定义为其均值之间的欧氏距离平方。与大多数其他层次方法相比，重心法对离群值更为稳健，但其他方面的性能却不如 Ward 法或类平均法。请参见 Milligan (1980)。

最短距离法 将两个聚类之间的距离定义为一个聚类中的某个观测与另一聚类中的某个观测之间的最小距离。最短距离法提供许多用户所需的理论属性，但在 Monte Carlo 研究中表现不好。请参见 Jardine and Sibson (1971)、Fisher and Van Ness (1971)、Hartigan (1981) 和 Milligan (1980)。最短距离法源自 Florek et al. (1951a, 1951b)，后来又被 McQuitty (1957) 和 Sneath (1957) 改写。

由于不对聚类形状施加约束，最短距离法能够检测细长和不规则的聚类，却牺牲了恢复紧密聚类的性能。最短距离法往往会在区分主聚类之前，先切割掉分布的尾部。请参见 Hartigan (1981)。

最长距离法 将两个聚类之间的距离定义为一个聚类中的某个观测与另一聚类中的某个观测之间的最大距离。最长距离法强烈偏向于生成具有大致相等直径的聚类，并且可能被中度离群值严重扭曲变形。请参见 Milligan (1980)。

快速 Ward 法 使用 Ward 法定义两个聚类之间的距离。“快速 Ward”使用近邻链算法来计算 Ward 距离。该算法缩短了计算时间，因为它不需要计算距离矩阵。只要超过 2000 行就会自动使用“快速 Ward”。

混合 Ward 法 应用将聚类分为两个阶段的算法。第一阶段是预处理步骤，该步骤使用近邻连接循环创建初步聚类。请参见“近邻连接循环的统计详细信息”。这样做是为了减少传递给层次聚类例程的表的大小。在执行一定数量的循环或创建一定数量的聚类之后，使用 Ward 法形成剩余的聚类。当您有数万或数十万个项要聚类时，该方法非常有用。

注意：与快速 Ward 法不同，该方法不会生成与完整 Ward 法相同的层次结构。不过，对于大量的项，它需要的计算时间较少，特别是若您有多个计算核心并且可以使用多线程进行近邻搜索的情况下。

数据格式 指定在计算多元距离时使用的数据格式。

如往常的数据 矩形数据，每个观测对应一行，每个变量对应一列。

被汇总的数据 按一个或多个标识列的水平汇总的数据。当您选择该选项时，启动窗口中会出现 “对象 ID” 文本框。指定标识列作为 “对象 ID”。被汇总的数据选项计算水平均值并将这些均值视为输入数据。

数据为距离矩阵 由观测之间的距离组成的数据。对于 n 个观测，距离表应有 n 行和 $n+1$ 列。有一列（通常为第一列）必须包含 n 个观测的唯一标识符。其余列包含该观测和 n 个观测之间的距离。请注意以下事项：

- 表的对角线元素应为 0 或缺失，因为点与其本身之间的距离为 0。不为 0 或缺失的值会视为 0，并且报表中会显示一条注释。
- 距离列可以是对称方矩阵，也可以是上三角或下三角矩阵并且缺失条目出现在下部或上部。若距离按方矩阵给出，则表不对称时报表中会出现一条警告。
- 您可以先开始使用不同的数据结构，然后保存距离矩阵。请参见 “保存距离矩阵”。

当您选择数据为距离矩阵选项时，输入距离列作为 “Y, 列”，标识符列作为 “标签”。“标签” 列必须具有 “字符型” 数据类型。有关示例，请参见 “距离矩阵的示例”。

数据被堆叠 具有单个关注响应且每个对象对应多行的数据。

当您选择数据被堆叠选项时，启动窗口中会显示 “特性 ID” 和 “对象 ID” 文本框。

- 输入**单个**列作为 “Y, 列”。
- 输入描述 “Y, 列” 变量分组的列作为 “特性 ID”。若仅输入两列且选择 “添加空间测度”，则可以在聚类分析中添加要使用的空间成分。请参见 “添加空间测度”。
- 输入对象的标识列作为 “对象 ID”。

执行的分析等价于按 “特性 ID” 列拆分 “Y, 列” 变量然后在不标准化响应列的情况下执行层次聚类。

提示：将该选项与 “添加空间测度” 选项一起使用可执行二维空间聚类。例如，晶片数据经常使用每个裸片对应一行的方式进行记录。关注点集中在晶片聚类。请参见 “使用 “空间测度” 进行晶片次品分类的示例”。

警告：因为有单个测量值列，“标准化数据” 选项不适用于堆叠数据。

标准化依据 指定在聚类之前如何对值进行标准化。这对于解决连续型和有序型列具有不同测量值尺度的问题很有用。

未标准化 使用原始数据。

列 通过减去列均值再除以列标准差，标准化每列中的值。

行 通过减去行均值再除以行标准差，标准化每行中的值。

列和行 通过先减去列均值和行均值，然后再加回总均值，对值进行标准化。然后，依据双重中心化数据的标准差对值统一尺度。

稳健标准化 减小离群值对连续型和有序型列的均值和标准差估计值的影响。该选项使用均值和标准差的 Huber M 估计值（Huber 19判别主成分4； Huber 1973； Huber and Ronchetti 2009）。对于包含离群值的列，该选项在确定多元距离时可以更好地用标准化值来表示不同的测度。

注意：若使用“标准化依据”选项并选择“稳健标准化”，则稳健均值和标准差将用于您指定的任何标准化方法。

缺失值补缺 补缺缺失值。若变量数小于等于 50 或少于行数的一半，则使用多元正态补缺。否则使用多元 SVD 补缺。

多元正态补缺计算配对协方差以构造响应列的协方差矩阵。然后，通过等价于回归预测的方法，使用给定变量不带缺失值的所有预测变量补缺每个缺失值。若构造的协方差矩阵不是正定矩阵，则使用其列均值对缺失值补缺。

多元 SVD 补缺通过使用奇异值分解避免构造协方差矩阵。请参见 《预测和专业建模》。

警告：缺失值补缺假定不存在任何聚类、数据来自单个多元正态分布，并且值的缺失是完全随机的。因为这些假设通常在实践中是不合理的，需谨慎使用该功能。不过与放弃大多数的数据相比，该功能可生成更具说明性的结果。

添加空间测度 （仅当将“数据被堆叠”选项选作“数据格式”时才可用。）当您的数据进行了堆叠并且包含两个对应于空间坐标（例如水平和垂直坐标）的特性列时，选择该选项。该选项将打开一个窗口，您可在其中选择和权衡空间成分以帮助聚类缺陷模式。这是专业方法，仅在非常特定的设置下适用。请参见“空间测度的统计详细信息”和“使用“空间测度”进行晶片次品分类的示例”。

双向聚类 （仅当将“如往常的数据”或“被汇总的数据”选项选作“数据格式”时才可用。）按指定的列和行进行聚类。色图添加到系统树图中，Y 变量的系统树图位于其底部。通常情况下，对于双向聚类，变量以相同的尺度进行测量，您不需要对数据进行标准化。

高级选项 指定“混合 Ward”法的高级选项。

混合目标 指定切换到层次聚类例程之前允许的最大聚类数。当层次聚类例程启动时，聚类数必须小于或等于“混合目标”。“混合目标”的默认值为 400。

混合循环 指定切换到层次聚类例程之前执行的最小近邻连接循环数。“混合循环”的默认值为 30。

混合初始 K 指定在近邻连接循环中使用的近邻的初始数目。近邻的数量可以增加或减少，这取决于在前一个循环中找到的唯一近邻的数量。“混合初始 K”的默认值为 10。

混合随机化 PCA 维 指定要在“随机化 PCA”降维方法中使用的维数。当“混合随机化 PCA 维”的值是大于零的任何值并提供进一步的速度改进时，使用该方法。“随机化

PCA”方法通过计算近似主分量来减少问题的维数，从而得到点之间的近似距离。请参见 Halko, Martinsson, and Tropp (2011)。

混合日志详细信息 指定是否显示日志中混合 Ward 法的每个状态的状况和计时。

使用保存的聚类表 使用单独的聚类历史记录表以指定聚类。

没有足够的非缺失数据警示

当您使用“被汇总的数据”或“数据被堆叠”数据格式时，JMP 警示：没有足够的非缺失数据会很难理解。下列情形会出现该警示：

- 若选定的“数据格式”为“如往常的数据”，当所有行或除一行之外的所有行有至少一个“Y，列”变量的值缺失时，就会出现该警示。
- 若选定的“数据格式”为“被汇总的数据”，当针对“对象 ID”列汇总您的数据，所有行或除一行之外的所有行有至少一个汇总“Y，列”变量的值缺失时，就会出现该警示。要查看“聚类”平台正在分析的数据结构，选择**表 > 汇总**，输入“对象 ID”列作为“分组”，“Y，列”变量作为“统计量 > 均值”。
- 若选定的“数据格式”为“数据被堆叠”，当针对“特性 ID”列拆分您的数据，所有行或除一行之外的所有行有至少一个拆分“Y，列”值缺失时，就会出现该警示。要查看“聚类”平台正在分析的数据结构，选择**表 > 拆分**，输入“特性 ID”列作为“拆分依据”，“Y，列”变量作为“拆分列”，“对象 ID”列作为“分组”。

提示：还会将一条消息打印输出到日志中，以标识具有缺失值的对象。

“层次聚类”报表

“层次聚类”报表显示使用的方法、系统树图和“聚类历史记录”表。若您在启动窗口中分配了一列作为“标签”，则该列的值会在系统树图中标识每个观测。

- “系统树图”
- “聚类历史记录”

系统树图

在“层次聚类”报表中，“系统树图”部分包含一个树状图和一个距离图。系统树图是一个树形图，用于表示观测如何划分到聚类中。系统树图还提供有关聚类相异性程度的信息。

可以从左到右读取系统树图来查看聚类过程。每步包括将两个**最近**聚类组成一个聚类。

- 由垂直线相连的水平线表示聚类的生成。
- 垂直线的水平位置表示，为了形成指定的聚类数，最近生成的两个类别之间的距离。

注意：当观测数小于 25 判别主成分 时，距离与 “距离图”中显示的距离成比例。否则，使用 “几何间距”。请参见 “系统树图尺度”。

您可以执行下列任务：

- 点击并拖动系统树图顶部或底部的菱形控点来标识给定数目的聚类。
- 点击任意聚类茎来选择系统树图和数据表中该聚类的所有成员。

距离图

“距离图”是系统树图下方显示的图。对于两个聚类连接成一个聚类的每一步，该图形都对应一个点。水平坐标表示聚类数，它们从左至右递减。点的垂直坐标是在给定步连接的两个聚类之间的距离。

您可以点击并拖动系统树图中的菱形控点来控制选择的聚类数。当您点击并拖动菱形时，图中会显示一条垂直线，您可以将它移动到对应的聚类数。经常有一个点对应的距离图斜率会变得平坦。此点表明趋势的自然转折，可帮助您确定聚类数。

有关系统树图和距离图之间关系的示例，请参见“层次聚类示例”。

聚类历史记录

在 “层次聚类”报表中，“聚类历史记录”表包含描述聚类历史记录的以下列：

聚类数 列出执行 “前导对象”和 “连接对象”所指示的连接**之后**得到的聚类数。聚类数从第一次连接开始计算，此时有 $n - 1$ 个聚类，其中 n 是对象数。该报表以降序方式列出聚类数，直到最后所有对象都包含在一个聚类中。这样，“聚类历史”遵循系统树图从左至右的顺序。

距离 聚类之间的距离，它根据您在启动窗口上选择的距离方法计算。请参见 “方法”。

前导对象 代表系统树图中要连接的第一个聚类。“前导对象”列中显示的聚类顺序和代表对象是数据排序产生的结果，没有内在意义。

连接对象 代表系统树图中要连接的第二个聚类。“连接对象”列中显示的聚类顺序和代表是根据数据如何排序产生的结果，没有内在意义。

“层次聚类”平台选项

“层次聚类”红色小三角菜单包含以下选项：

聚类着色 根据聚类成员关系对系统树图的标签及其关联的连接条进行着色。此外将相应的颜色分配给数据表的行。若您更改聚类数，颜色也会更新。若您取消选择该选项，颜色将不再根据变量数进行更新。

标记聚类 将标记分配给数据表中与该行所属的聚类对应的那些行。若您更改聚类数，标记也会更新。若您取消选择该选项，标记将不再根据变量数进行更新。

聚类数 指定行聚类的数量，并将系统树图滑块定位到该数量。

聚类准则 （将“数据为距离矩阵”选作“数据格式”时不可用。）显示或隐藏整个行聚类数量范围的“三次聚类准则”(CCC) 表。CCC 用于估计聚类数。它可以与任何基于距离的聚类算法一起使用。CCC 值越大则表明在特定聚类数下拟合的效果越好。请参见 SAS Institute Inc. (1983)。

显示系统树图 显示或隐藏“系统树图”报表。

系统树图尺度 包含用于统一系统树图尺度的选项：

距离尺度 基于在启动窗口中指定的距离方法，将任意两个连接点之间的水平距离显示为在该点连接的两个聚类之间的距离。距离尺度与“距离图”中使用的是同一个尺度，它是系统树图的默认尺度。

等间距 显示任意两个连接点之间的水平距离相等。

几何间距 随聚类数的增加，连接点之间的水平距离也随之增加。当有许多对象并且您想较小的聚类比较大的聚类更加明显时可以使用该选项。

距离图 显示或隐藏系统树图下方的距离图。

显示聚类数控点 在系统树图上显示或隐藏用于手动更改聚类数的控点。

缩放至选定行 在系统树图中选择聚类之后，选择并放大特定聚类。或者，您可以双击聚类将其放大。使用“解除缩放”可恢复到原始视图。

解除缩放 在放大之后，使系统树图恢复到原始视图。

以选定聚类为轴心转动 反转当前选定聚类的两个子聚类的顺序。

定位 提供用于更改标签和系统树图的其他部分的位置的选项。

色图 支持您添加色图或热图，用于显示按值着色的各个“Y，列”变量。子菜单中提供了若干颜色主题选择。要删除色图，请选择**色图 > 无**。

更多色图列 （仅当将“如往常的数据”选作“数据格式”时才可用。）为指定列添加色图。

图例 显示或隐藏色图中使用的颜色的图例。每个指定列都有一个单独的图例。该选项仅在启用色图时才可用。

注意：若有超过 400 列，则会显示一个图例，并为色图中使用的颜色提供标准化得分。

双向聚类 （仅当将“如往常的数据”或“被汇总的数据”选作“数据格式”时才可用。）按指定的列和行进行聚类。色图添加到系统树图中，Y 变量的系统树图位于其底部。通常情况下，对于双向聚类，变量以相同的尺度进行测量，您不需要对数据进行标准化。

列聚类 （仅当使用“双向聚类”时才可用。）提供用于在双向聚类中聚类各列的选项。

列聚类数 指定列聚类数。

列聚类准则 显示或隐藏整个列聚类数量范围的“三次聚类准则”(CCC) 表。CCC 用于估计聚类数。它可以与任何基于距离的聚类算法一起使用。CCC 值越大则表明在特定聚类数下拟合的效果越好。请参见 SAS Institute Inc. (1983)。

保存列聚类 保存包含列的聚类成员信息的新数据表。

保存聚类 保存包含聚类成员信息的新数据表。若在启动窗口中选择了“添加空间测度”，则聚类数也会保存到 Hough 数据表。

保存聚类均值 创建一个新的数据表，其中包含每个聚类中的行数和每列的均值。

保存其他 显示其他保存选项的子菜单。

保存最近聚类公式 创建包含最近聚类的公式的数据表列。该选项计算每个聚类重心之间的欧氏距离平方，并选择最靠近的聚类。请注意，该公式不一定总能重现“层次聚类”提供的聚类分配，因为聚类的确定方式有所不同。不过，聚类分配非常相似。（选定**被汇总的数据**、**数据为距离矩阵**或**数据被堆叠**时不可用。）

保存聚类历史 创建新的数据表，其中包含“聚类历史记录”报表中的信息。

保存显示顺序 创建包含行在系统树图中的显示顺序的数据表列。

保存距离矩阵 创建一个新的数据表，其中包含观测之间的距离。

保存星座坐标 将星座图坐标保存至数据表。（选定被汇总的数据、数据为距离矩阵或数据被堆叠时不可用。）

保存聚类层次结构 创建一个数据表，其中包含编写自定义系统树图的脚本所需的信息。对于每个聚类连接，该选项都输出三行：第一行表示连接对象、第二行表示前导对象、第三行表示结果，用于给出聚类中心、大小和其他信息。

保存聚类树 创建一个新的数据表，其中包含在 JMP 和 SAS 之间比较聚类树所需的信息。对于每个聚类连接，该选项都为每个新聚类输出一行，包含该聚类的大小和其他信息。

聚类历史记录 显示或隐藏“聚类历史记录”报表。请参见 “聚类历史记录”。

聚类汇总 （选定数据为距离矩阵时不可用。）显示或隐藏包含以下信息的报表：

聚类均值 给出每个聚类的观测数（若数据被堆叠则为“对象 ID”）和每个变量的均值的表。

聚类标准差 给出每个聚类的观测数（若数据被堆叠则为“对象 ID”）和每个变量的标准差的表。

聚类均值图 聚类均值的平行图或二维热图。

该图为平行图，但选中**数据被堆叠**并且有两个“特性 ID”变量时除外。对于平行图，每个变量的轴都统一了尺度。

- 若选中“标准化数据”，轴的范围在均值的上下两个标准差，其中标准差和均值基于原始数据进行计算。若聚类均值超出该范围，则轴会扩展以包括该均值。
- 若未选中“标准化数据”，则会使用显示了尺度的公共垂直轴。（该尺度等价于“图形生成器”中的“统一尺度”选项）。

当选中**数据被堆叠**并且有两个“特性 ID”变量时，在每个位置为每个聚类显示 Y 变量均值的二维图。这些图使用“由蓝经灰到红”颜色梯度着色。

列汇总 对于每个变量，给出表示聚类所解释的变异比例的 R 方值。该数值是聚类中变量回归的 R 方值。该选项还给出 R 方值的条形图。

最后连接离群值 显示或隐藏在算法中很晚聚类的观测所在的表。当算法完成 80% 时，该表中的观测仍然自成一个聚类。由于这些观测中的每一个直到算法后期都仍保持自成一个聚类，因此这些观测是数据集中的可能离群值。

星座图 显示或隐藏另一种在层次聚类系统树图中显示信息的方式。每个观测 （行）由一个端点表示，每个聚类连接由一个新点表示。绘制的线条表示聚类成员关系。线条长度表示聚类之间的距离。较长的线表示聚类之间的距离较长。

您可以将鼠标悬停在星座图中的线条上来查看其长度。但是，长度值只有相对意义。轴尺度、点方向和线条角度是任意的。这样安排可以使节点端间隔开，图不会显得很混乱，这对于大数据集来说十分重要。

要关闭端点处的标签，请在 “星座图”内右击并选择**显示标签**。

散点图矩阵 （仅当将 “如往常的数据”选作 “数据格式”时才可用。）使用所有变量创建散点图矩阵。

平行坐标图 （仅当将 “如往常的数据”选作 “数据格式”时才可用。）为每个聚类创建平行坐标图。轴的尺度参照 “聚类均值图”的说明。请参见 “聚类均值图”。

聚类处理比较 （仅当您按住 Shift 并点击 “层次聚类”红色小三角时才可用。）选择响应列和二水平处理列。创建 “层次聚类差值”报表。

请参见《使用 JMP》获取有关下列选项的信息：

本地数据过滤器 显示或隐藏支持您过滤特定报表中使用的数据的本地数据过滤器。

重新运行 包含使您可以重复或重新启动分析的选项。在支持该功能的平台中，“自动重新计算”选项立即在相应报表窗口中反映您对数据表所做的更改。

平台首选项 包含的选项支持您查看当前平台首选项或更新平台首选项以匹配当前 JMP 报表中的设置。

保存脚本 包含的选项支持您保存可将报表重现到若干目标的脚本。

保存 “依据”组脚本 包含使您可以保存脚本的选项，可将为 “依据”变量的所有水平重新生成平台报表的脚本保存到多个不同的位置。仅当在启动窗口中指定 “依据”变量时才可用。

层次聚类的更多示例

本节包含使用 “层次聚类”平台的示例。

- “距离矩阵的示例”
- “使用 “空间测度”进行晶片次品分类的示例”

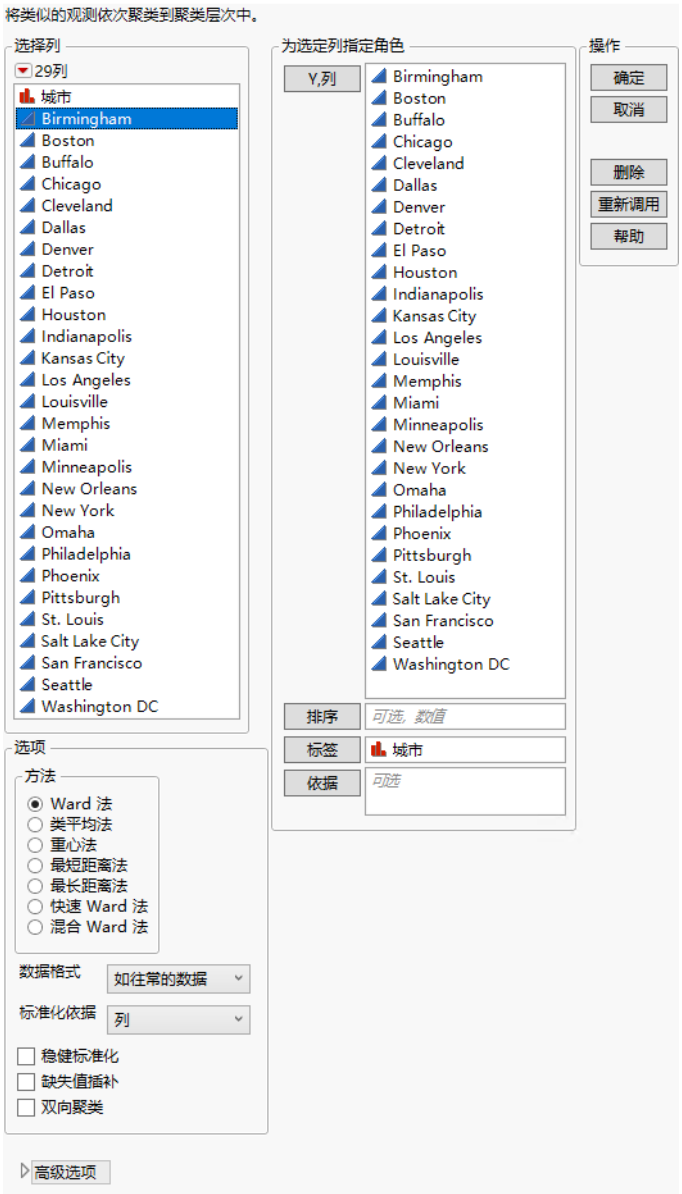
距离矩阵的示例

在本例中，数据中的观测之间有距离，因此您可以在“层次聚类”启动窗口中使用“数据为距离矩阵”选项。距离矩阵的正确数据表结构包含以下项：

- 具有“字符型”数据类型的标识符列（通常为第一列）。
- 一组 *n* 个列，其中 *n* 也是行数。这 *n* 个列定义对角线上值为 0 或缺失值的对称矩阵。

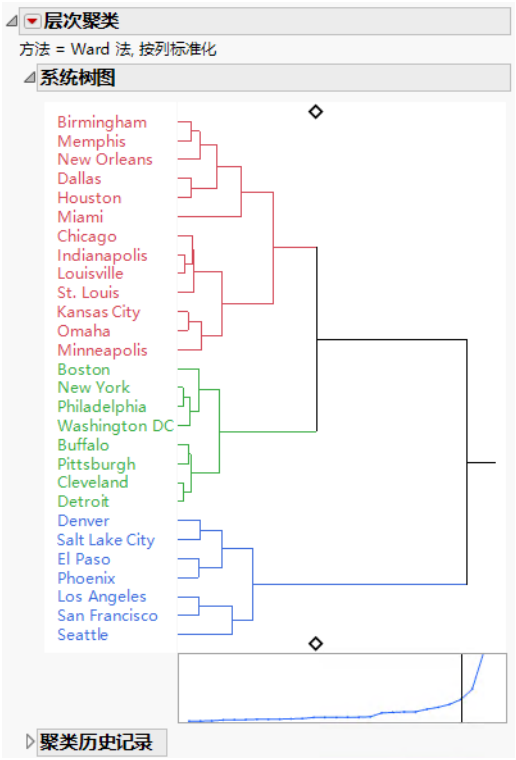
1. 选择**帮助 > 样本数据文件夹**，然后打开 Flight Distances.jmp。
2. 选择**分析 > 聚类 > 层次聚类**。
3. 在“数据格式”旁边的列表中，选择**数据为距离矩阵**。
4. 选择**城市**并点击**标签**。
5. 选择所有其余列并点击 **Y, 列**。

图 13.5 完成的 “距离矩阵” 启动窗口



6. 点击**确定**。
7. 点击 “层次聚类” 红色小三角并选择**聚类着色**。

图 13.6 “Flight Distances”的“系统树图”报表



在显示飞行距离的“系统树图”报表中，菱形的位置指示模型将城市分为三个聚类。聚类在系统树图中用不同颜色显示。有关如何解释报表的详细信息，请参见“系统树图”。

使用“空间测度”进行晶片次品分类的示例

在本例中，您使用“层次聚类”平台中提供的称为“空间测度”的专业聚类选项。

1. 选择**帮助 > 样本数据文件夹**，然后打开 Wafer Stacked.jmp。
2. 选择**分析 > 聚类 > 层次聚类**。
3. 在“数据格式”旁边的列表中，选择**数据被堆叠**。
   针对堆叠数据的更多选项显示在启动窗口中。
4. 选择**缺陷数**并点击 **Y, 列**。
5. 选择 **X_ 裸片**和 **Y_ 裸片**并点击**特性 ID**。
6. 选择**批**和**晶片**并点击**对象 ID**。
7. 选择**添加空间测度**。

图 13.7 完成的 “聚类”启动窗口

8. 点击**确定**。

图 13.8 “空间成分”窗口

缺陷数在 1423 个位置进行的测量，因此有 1423 个特性变量。

9. 点击**确定**接受 “空间”窗口中的选择。

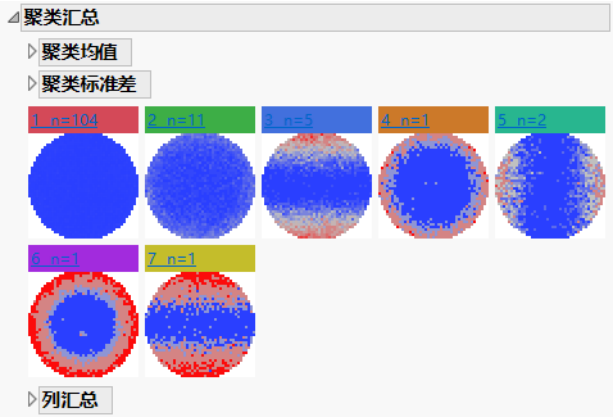
两个窗口打开：“层次聚类”报表和 “Wafer Stacked 缺陷数空间”数据表。

10. 在 “系统树图”图中，点击并拖动顶部的菱形控点来探索各种数量的聚类。

    当您拖动控点时，系统树图下方的距离图中的垂直线会移动到相应的聚类数。垂直坐标是在给定步生成的两个聚类之间的距离。图形看起来在聚类数为 7 时趋于水平。

11. 点击 “层次聚类”红色小三角并选择**聚类数**。
12. 输入 7 并点击**确定**。
13. 点击 “层次聚类”红色小三角并选择**聚类汇总**。

图 13.9 “聚类汇总”报表



晶片图指出每个聚类的次品数空间性质。“聚类 1”包含 104 个晶片，它们的次品数较少，次品均匀分布在晶片中。“聚类 3”有 5 个晶片，次品集中在上半圆和下半圆的两端处。您可以在聚类分析生成的数据表中查看各晶片及其 Hough 空间映射的图。请参见 “空间测度的统计详细信息”。

“层次聚类”平台的统计详细信息

本节包含 “层次聚类”平台的统计详细信息。

- “空间测度的统计详细信息”
- “距离方法的统计详细信息”
- “近邻连接循环的统计详细信息”

空间测度的统计详细信息

要使用 “层次聚类”平台中的 “添加空间测度”选项，您的数据必须被堆叠并且包含两个对应于空间坐标的特性列。某些空间测度使用 Hough 变换构造。请参见 White et al. (2008) 和 Ballard (1981)。请参见 “使用 “空间测度”进行晶片次品分类的示例”。

“选择空间成分”窗口

若您在启动窗口中执行以下操作，则会显示 “选择空间成分”窗口：

- 选择“数据被堆叠”数据结构
- 指定对应于空间坐标的两列作为“特性 ID”
- 指定“对象 ID”
- 选择“添加空间测度”

在 “选择空间成分”窗口中，您选择并权衡用于聚类分析的空间成分。这些成分用于构造聚类分析中使用的变量。随即打开一个新表，其中每个对象对应一行。该表包含每个对象的计算空间成分。

变量 聚类分析中构造并使用的变量类型。使用空间成分和响应 Y 构造变量。

特性 对于每个对象，针对每个位置（由两个“特性 ID”变量定义）计算的 Y 变量的值。

角度，饼图 反映楔形或半球形的变量。

半径，圆圈 反映圆形的变量。

条纹角度 反映具有相同角度的条纹的变量。

条纹位置 反映具有相同空间位置的条纹的变量。

曝光区中的位置 基于曝光区中裸片位置的变量。“曝光区中的位置”变量表示为 ShotPos [vShotSize, hShotSize]，其中 vShotSize 和 hShotSize 分别为定义的垂直和水平曝光区大小。

曝光区 标识对象所在矩形的变量，您可以指定对象在矩形中的水平位置和垂直位置数。术语**曝光区**用于半导体晶片数据中，以标识晶片上哪些裸片在一起成像。

输入“曝光区水平大小”和“曝光区垂直大小”的值。将水平曝光区大小指定为 4，垂直曝光区大小指定为 5，这样指示曝光区中最多有 20 个裸片。创建的总标识符数计算如下：

floor[(hSize+hShotSize-1)/hShotSize]* floor[(vSize+vShotSize-1)/vShotSize]

其中 hSize 和 vSize 分别是水平和垂直最大位置数，hShotSize = 曝光区水平大小，vShotSize = 曝光区垂直大小。

注意：曝光区变量表示为 Shot[vert, horiz]，其中 vert 和 horiz 分别表示垂直和水平裸片位置。

数目 构造的给定类型的总变量数。

权重 用于确定聚类的给定类型的变量的重要性测度。

空间测度报表

当您在“选择空间成分”窗口中点击“确定”时，会出现两个窗口。

“层次聚类”报表

当您使用堆叠数据和两个特性 ID 执行分析时，“聚类汇总”报表显示 Y 变量的空间映射。每个图都是一个二维图，显示由“特性 ID”变量定义的每个位置的聚类均值。该图使用具有分位数尺度的“由蓝经灰到红”颜色梯度。使用分位数尺度可减轻离群值的影响。

空间数据表

空间测度的数据表为每个唯一对象 ID 提供一行。使用“由蓝经灰到红”默认颜色梯度显示列以表示 Y 变量。该表包含以下列：

对象 显示每个空间位置 （由两个 “特性 ID”变量定义）处的 Y 变量的热图的表达式列。

Hough 显示每个对象的 Hough 空间的热图的表达式列。请参见 White et al. (2008)。

空间测度 显示每个对象的计算值的每个空间测度对应的列。单元格按值着色。

距离方法的统计详细信息

本节提供基于您在“层次聚类”启动窗口中选择的方法计算距离所用的公式。有关方法说明，请参见“方法”。

公式使用以下符号，小写符号通常与观测有关，大写符号通常与聚类有关：

n 是观测数

v 是变量数

x_i 是第 i 个观测

C_K 是第 K 个聚类，是 {1, 2,..., n} 的子集

N_K 是 C_K 中的观测数

$\bar{x}$ 是样本均值向量

$\bar{x}_K$ 是聚类 C_K 的均值向量

$\|\mathbf{x}\|$ 是 **x** 各元素的平方和的平方根 （向量 **x** 的欧氏长度）

$d(x_i, x_j)$ 为 $\|x_i - x_j\|^2$

类平均法 类平均连接聚类法的距离计算如下：

$$D_{KL} = \sum_{i \in C_K} \sum_{j \in C_L} \frac{d(x_i, x_j)}{N_K N_L}$$

重心法 聚类重心法的距离计算如下：

$$D_{KL} = \left\| \bar{x}_K - \bar{x}_L \right\|^2$$

Ward 法 Ward 法距离计算如下：

$$D_{KL} = \frac{\left\| \overline{x}_K - \overline{x}_L \right\|^2}{\frac{1}{N_K} + \frac{1}{N_L}}$$

最短距离法 最短距离聚类法的距离计算如下：

$$D_{KL} = \min_{i \in C_K} \min_{j \in C_L} d(x_i, x_j)$$

最长距离法 最长距离聚类法的距离计算如下：

$$D_{KL} = \max_{i \in C_K} \max_{j \in C_L} d(x_i, x_j)$$

近邻连接循环的统计详细信息

在“层次聚类”平台中，“近邻连接循环”用在“混合 Ward”法的第一阶段中。这样做是为了减少传递给层次聚类例程的表的大小。“近邻连接循环”算法指定以下内容：

混合目标 指定停止算法之前允许的最大聚类数。默认值为 400。

混合循环 指定停止算法之前执行的最小近邻连接循环数。默认值为 30。

混合初始 K 指定在近邻连接循环中使用的近邻的初始数目。默认值为 10。

“近邻连接循环”算法重复以下步骤：

1. 创建有利点 (VP) 树以高效查找最近邻。
2. 对于每个项，确定该项的 *k* 个最近邻。
3. 近邻对按距离排序。
4. 对于近邻对中距离最小的那一半，若项尚未在该循环中与另一个项连接，则连接每对中的项。连接的项成为下一个循环中的项。
5. 重复第 1 步 到第 4 步，直到达到最小循环数 （混合循环）。
   - 若项目数小于或等于“混合目标”，则停止。
   - 若项目数大于“混合目标”，请继续重复第 1 步 到第 4 步，直到项目数小于或等于“混合目标”。

在每个循环中，若连接的对的数量较少，则将在下一个循环中增加最近邻数 *k*。若在前一循环中连接了足够数量的对，则在后一个循环中可以减少 *k* 的值。*k* 的值根据以下规则增减。

- 若连接的第 4 步 中的对少于 20%，则 *k* 的值加 10。
- 若连接的第 4 步 中的对少于 10%，则 *k* 的值加 20。
- 若连接的第 4 步 中的对少于 5%，则 *k* 的值加 30。
- 若连接的第 4 步 中的对超过 30%，则 *k* 的值减 10。

第 14 章

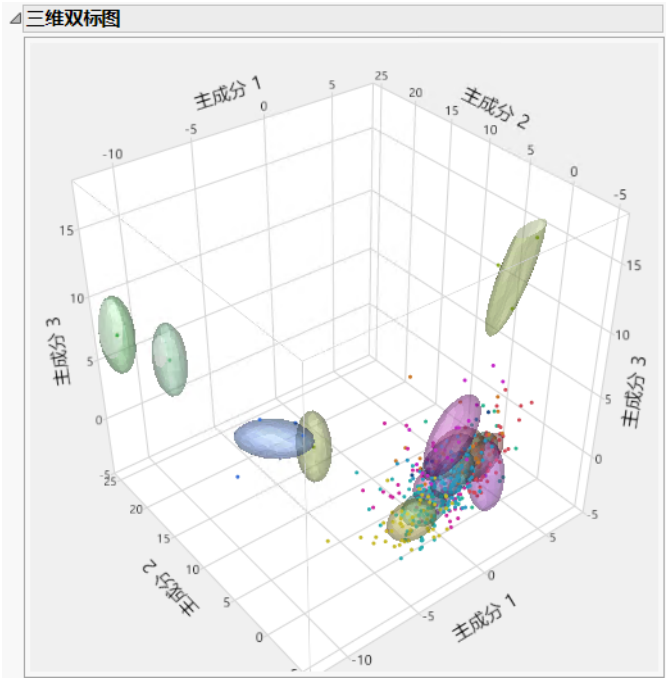
K 均值聚类

使用距离对观测分组

使用“K 均值聚类”平台可将在几个变量上享有相似值的观测分组在一起。*K*-均值法适用于包含大约 200 到 100,000 个观测的大型数据表。

“K 均值聚类”平台使用迭代算法对观测进行划分来构造指定数量的聚类。该方法称为 **K-均值**，它将观测划分聚类以便最小化到聚类重心的距离。您必须提前指定聚类数 *k*。不过，您可以比较不同 *k* 值的结果以选择适合您的数据的最优聚类数。

图 14.1 三维双标图



目录

“K 均值聚类”平台概述

“K 均值聚类”平台使用迭代拟合过程形成指定数量的聚类。*K*- 均值算法首先选择一组称为**聚类种子**的 k 个点作为对聚类均值的最初推测。每个观测被分配到最近的聚类种子，形成一组临时聚类。然后这些种子被聚类均值替代，各点会重新分配，最后该过程一直持续到聚类中不再有进一步的变化。

K- 均值算法是 EM **算法**的一种特殊情况，其中 *E* 代表期望，*M* 代表最大化。在 *K*- 均值算法中，计算临时聚类均值代表期望步，将点分配给最近的聚类代表最大化步。

K- 均值聚类仅支持数值列。*K*- 均值聚类忽略建模类型（名义型和有序型），而将所有数值列视为连续型。

您必须提前指定聚类数 *k* 或 *k* 的值范围。不过，您可以比较不同 *k* 值的结果以选择适合您的数据的最优聚类数。

有关 *K* 均值聚类的背景信息，请参见 SAS Institute Inc.(2020d) 中的“FASTCLUS 过程”一章以及 Hastie et al.(2009)。

“K 均值聚类”是 JMP 提供的对观测进行聚类的四个平台之一。有关四种方法的比较，请参见“对观测聚类的平台概述”。

对观测聚类的平台概述

聚类是将在几个变量上享有相似值的观测分组在一起的一种多元方法。通常情况下，观测在 *p* 维空间内散布不均，其中 *p* 是变量数。这些观测反而会形成聚簇或聚类。标识出这些聚类使您可以更深层次地了解您的数据。

注意：JMP 还提供可以对变量聚类的平台。请参见“聚类变量”。

JMP 提供四个平台供您对观测聚类：

- 层次聚类对于小型和大型数据表非常有用，并且允许使用字符数据。“层次聚类”将行按描绘为一棵树的层次序列形式进行组合。您可以在生成树后选择最适合您数据的聚类数。请参见“层次聚类”。
- “K 均值聚类”适用于多达数百万行的大型表，并且只允许数值数据。您需要提前指定聚类数 *k*。该算法可以对聚类种子点做出推测。随后开始在将数据点分配到相应类别和重新计算聚类中心之间交替进行迭代过程。请参见“K 均值聚类”。
- “正态混合”适用于数据来自重叠的多元正态分布的混合分布这种情况，并且只允许数值数据。对于具有多元离群值的情形，您可以使用假设具有均匀分布的离群值聚类。请参见“正态混合”。

  您需要提前指定聚类数。最大似然用于同时估计混合比例以及均值、标准差和相关性。为每个点指定属于每个组的概率。使用 EM 算法获取估计值。

- “潜在类分析”适用于大多数变量是分类变量这种情况。您需要提前指定聚类数。该算法拟合假定具有多项式混合分布的模型。为每个观测计算聚类成员关系的最大似然估计值。观测会被归类到其成员关系概率最大的聚类中。请参见“潜在类分析”。

表 14.1 聚类方法汇总

方法	数据类型或建模类型	数据表大小	指定聚类数
层次聚类	任意	使用混合 Ward 法，最多数十万行 使用快速 Ward 法，最多 200,000 行 使用其他方法，最多 5,000 行	否
K 均值聚类	数值	多达数百万行	是
正态混合	数值	任意大小	是
潜在类分析	名义型或有序型	任意大小	是

有些聚类平台提供用于处理数据中的离群值的选项。但是，若数据中有离群值，则最好在分析之前探索这些离群值。可以使用“探索离群值”实用工具完成该操作。有关详细信息，请参见《预测和专业建模》。

“K 均值聚类”的示例

在本例中，您使用“K 均值聚类”平台对血细胞计数分析中的观测进行聚类。

1. 选择**帮助** > **样本数据文件夹**，然后打开 Cytometry.jmp。
2. 选择**分析** > **聚类** > **K 均值聚类**。
3. 选择 CD3、 CD8、 CD4 和 MCB，然后点击 **Y, 列**。
4. 点击**确定**。
5. 在**聚类数**旁边输入 3。
6. 在**聚类范围**旁边输入 15 （可选）。

   由于“聚类范围”设置为 15，平台提供 3 到 15 个聚类的拟合。您可以随后确定偏好的聚类数。
7. 点击**执行**。

图 14.2 “聚类比较”报表

聚类比较

方法	聚类数	CCC	最佳
K 均值聚类	3	23.1784	
K 均值聚类	4	8.80709	
K 均值聚类	5	29.5123	
K 均值聚类	6	52.5517	
K 均值聚类	7	49.5876	
K 均值聚类	8	56.5308	
K 均值聚类	9	54.053	
K 均值聚类	10	69.8707	
K 均值聚类	11	70.5239	最优 CCC
K 均值聚类	12	61.5326	
K 均值聚类	13	68.1277	
K 均值聚类	14	66.4044	
K 均值聚类	15	69.9928	

“聚类比较”报表显示在报表窗口顶部。最佳拟合由最高 CCC 值确定。在本例中，最佳拟合出现在拟合 11 个聚类时。

8. 滚动到 “K 均值聚类数 =11”报表。

图 14.3 “K 均值聚类数 =11”报表

K 均值聚类数=11

逐列统一尺度

聚类汇总

聚类	计数	步进	准则
1	816	24	0
2	482		
3	157		
4	577		
5	856		
6	498		
7	447		
8	549		
9	377		
10	123		
11	118		

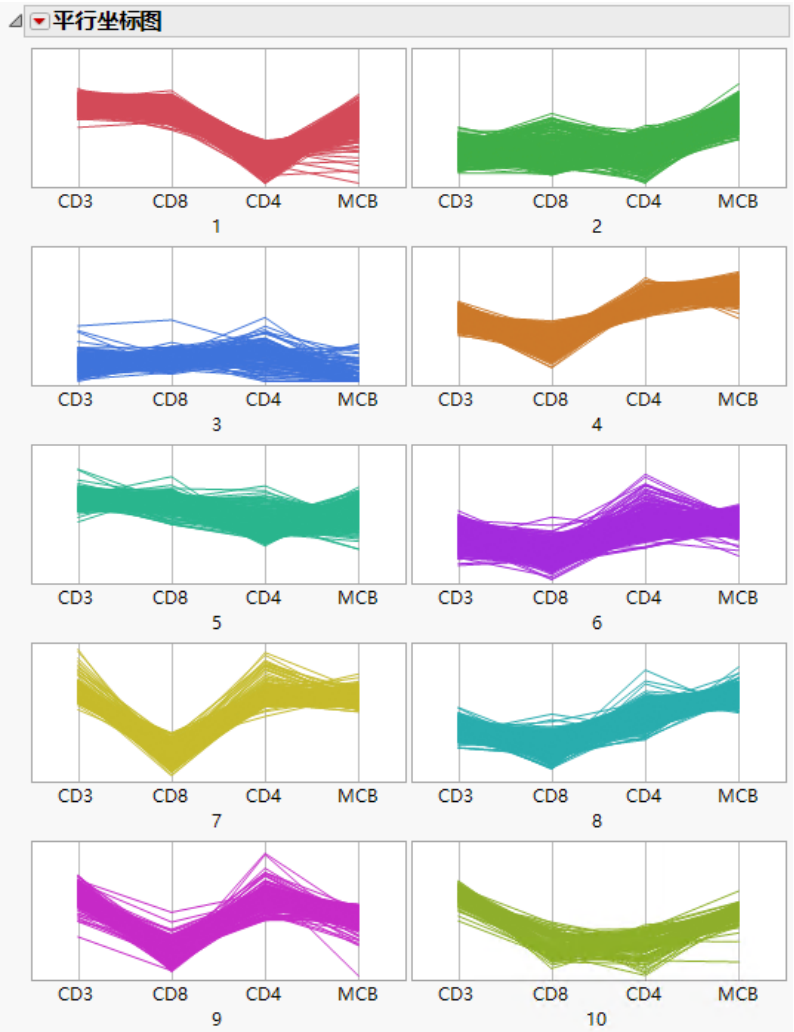
聚类均值

聚类	CD3	CD8	CD4	MCB
1	314.073529	300.270833	106.615196	183.4375
2	140.091286	116.365145	127.024896	205.014523
3	90.7898089	87.5477707	109.356688	15.0191083
4	247.426343	148.064125	314.074523	261.831889
5	320.287383	307.372664	193.117991	193.174065
6	188.686747	99.0863454	220.062249	171.682731
7	347.496644	89.9574944	320.782998	231.557047
8	210.258652	126.125683	260.327869	245.588342
9	328.206897	89.7824934	312.909814	175.965517
10	322.105691	113.715447	116.666667	181.731707
11	349.864407	284.813559	360.932203	267.474576

“聚类汇总”报表显示 11 个聚类中每一个聚类的观测数。“聚类均值”报表显示每个聚类的四个标记读数的均值。

9. 点击 “K 均值聚类数 =11”红色小三角，然后选择**平行坐标图**。

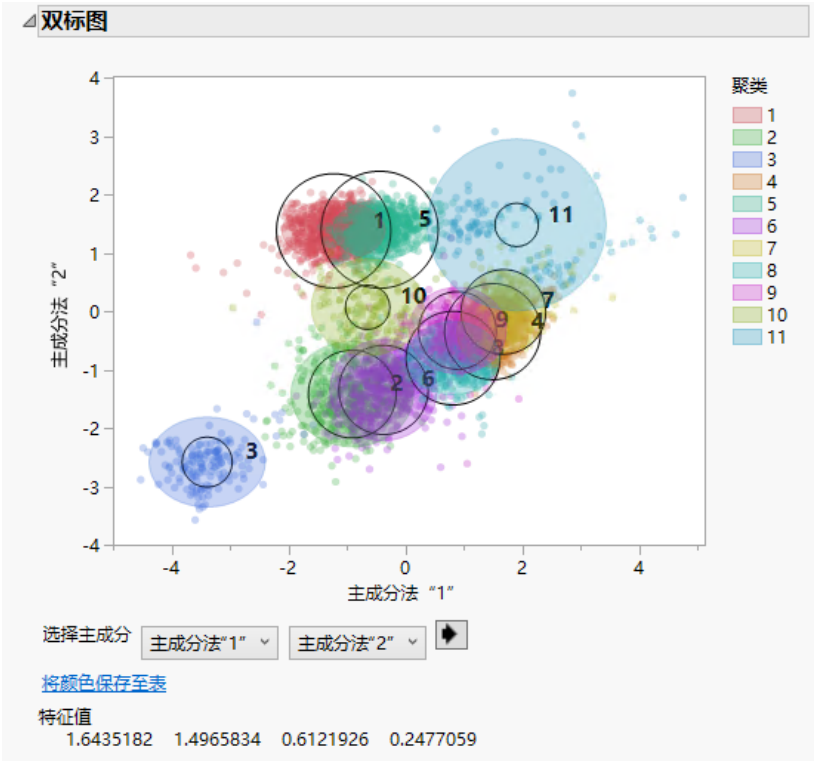
图 14.4 血细胞计数数据的 “平行坐标图”



“平行坐标图”显示每个聚类中观测的结构。使用这些图可查看聚类有哪些不同。聚类 4、判别主成分、7、8 和 9 往往有相对低的 CD8 值和高的 CD4 值。另一方面，聚类 1 有更高的 CD8 值和更低的 CD4 值。

10. 点击 “K 均值聚类数 =11”红色小三角，然后选择**双标图**。

图 14.5 血细胞计数数据的 “双标图”

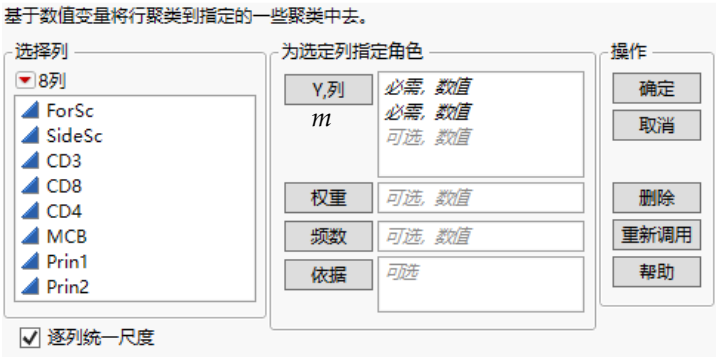


标识聚类颜色的图例显示在图的右侧。基于聚类前两个主成分，聚类 3、 10 和 11 显示为与其他聚类分的最开。图 14.4 中这几个聚类的平行坐标图支持该结论，它们与其他聚类的图有所不同。使用该图下面的列表可查看其他主成分组合的双标图。

启动 “K 均值聚类” 平台

通过选择**分析 > 聚类 > K 均值聚类**来启动 “K 均值聚类” 平台。

图 14.6 “K 均值聚类” 启动窗口



有关 “选择列” 红色小三角菜单中选项的详细信息，请参见《使用 JMP》。

Y，列 用于对观测聚类的变量。

> **注意：***K-* 均值聚类仅支持数值列。

权重 一列，该列的数值为分析中的每一行都分配一个权重。

频数 一列，列中的数值为分析中的每行分配一个频数。

依据 一列，其水平定义不同的分析。对于指定列的每个水平，都分析相应行。结果显示在不同的报表中。若分配了多个 “依据” 变量，则为 “依据” 变量水平的每个可能组合生成单独分析。

启动窗口选项

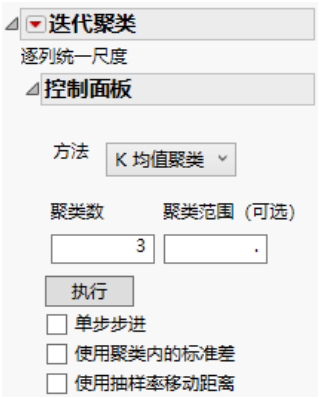
逐列统一尺度 独立于其他列调整每列的尺度。在变量不共享相同的测量值尺度，而您不希望一个变量支配聚类过程的情况下使用。例如，一个变量具有介于 0 和 1000 之间的值，而另一个变量具有介于 0 和 10 之间的值。在这种情况下，您可以使用该选项，以便聚类过程不被第一个变量所支配。

当您点击 “确定” 时，“控制面板” 随即显示。请参见“ “迭代聚类” 报表”。

“迭代聚类”报表

在“K 均值聚类”平台中，“迭代聚类”报表显示用于拟合模型的“控制面板”。您可以反复拟合不同数量的聚类，也可以使用“聚类范围”选项指定一个范围。在拟合模型时，其他报表会添加至窗口。请参见“K 均值报表”。

图 14.7 “迭代聚类”控制面板



“控制面板”包含以下选项：

方法 下列聚类方法可用：

K 均值聚类 在本章中进行了说明。

自组织图 在“‘自组织图’控制面板”中进行了说明。

聚类数 指定要形成的聚类数。

聚类范围（可选） 提供要形成的聚类数的上限。若在此输入了某个数字，平台将为“聚类数”与“聚类范围（可选）”中输入的值之间的每个整数创建单独的分析。

执行 除非选择了“单步步进”，否则自动拟合聚类。

单步步进 支持您每次执行一个迭代来逐步完成聚类过程。当您选择“单步步进”并点击“确定”时，会显示“K 均值聚类”报表，没有聚类结果，而是含有“执行”和“步进”按钮。

- 点击“步进”按钮每次执行一个迭代来逐步完成聚类过程。
- 点击“执行”按钮自动拟合聚类。

使用聚类内的标准差 针对每个聚类内的观测，使用每个变量在聚类内的估计标准差对距离统一尺度。若不选择该选项，距离按照每个变量的标准差的总估计值统一尺度。

使用抽样率移动距离 基于聚类大小调整距离。若您具有大小不等的聚类，则观测应具有更高的概率被分配给较大的聚类，因为观测来自较大聚类的先验概率更高。

“迭代聚类”选项

本节介绍 “迭代聚类”红色小三角菜单中的选项。

请参见《使用 JMP》获取有关下列选项的信息：

本地数据过滤器 显示或隐藏支持您过滤特定报表中使用的数据的本地数据过滤器。

重新运行 包含使您可以重复或重新启动分析的选项。在支持该功能的平台中，“自动重新计算”选项立即在相应报表窗口中反映您对数据表所做的更改。

平台首选项 包含的选项支持您查看当前平台首选项或更新平台首选项以匹配当前 JMP 报表中的设置。

保存脚本 包含的选项支持您保存可将报表重现到若干目标的脚本。

保存 “依据”组脚本 包含使您可以保存脚本的选项，可将为“依据”变量的所有水平重新生成平台报表的脚本保存到多个不同的位置。仅当在启动窗口中指定“依据”变量时才可用。

K 均值报表

在“控制面板”中点击“执行”后，将显示一个聚类比较报表以及一个或多个 K 均值报表。K 均值报表被动态命名为“K 均值聚类数 =<k>”，具体取决于 *k*，即拟合的聚类数。每次执行拟合时，都会出现“K 均值聚类数 =<k>”报表。

- “聚类比较”报表
- “K 均值聚类数”报表

“聚类比较”报表

在“K 均值聚类”平台中，“聚类比较”报表提供拟合统计量以比较各种模型。拟合统计量为“三次聚类准则”(CCC)。CCC 值越大表示拟合效果越好。最佳拟合在名为“最佳”的列中用最优 CCC 指示。请参见 SAS Institute Inc. (1983)。CCC 计算中不包括常数列。

“K 均值聚类数”报表

在“K 均值聚类”平台中，每个“K 均值聚类数”报表都为每个聚类提供以下汇总统计量：

- “聚类汇总”报表提供聚类数和每个聚类中的观测数，以及所需的迭代次数。
- “聚类均值”报表为每个变量提供每个聚类中观测的均值。
- “聚类标准差”报表为每个变量提供每个聚类中观测的标准差。

“K 均值聚类数”报表选项

每个 “K 均值聚类数”报表都包含以下红色小三角菜单选项：

双标图 以数据的前两个主成分显示点和聚类的图，以及用于标识聚类颜色的图例。围绕聚类中心绘制圆圈，而且圆圈的大小与聚类内的计数成比例。着色区域是围绕均值的密度等高线。默认情况下，该区域指示该聚类中 90% 的观测所在的位置 (Mardia et al. 1980)。使用该图下方的列表可将图轴改为其他主成分。或者，使用箭头按钮在所有可能的轴组合之间循环切换。该图之下还有一个用于将聚类颜色保存到数据表的选项。请参见“将颜色保存至表”。特征值以降序方式显示。

注意：若在启动窗口中选中 “逐列统一尺度”，则双标图使用相关性矩阵。若未选中 “逐列统一尺寸”，则双标图使用协方差矩阵。

双标图选项 包含用于控制 “双标图”外观的下列选项：

显示双标图射线 显示双标图射线。带标签的射线显示协变量在由主成分定义的子空间中的方向。它们表示每个变量与每个主成分的关联程度。

双标图射线位置 可让您指定双标图射线的位置和射线尺度。默认情况下，这些射线从点 (0,0) 发出。在该图中，您可以拖动射线或使用该选项指定坐标。您还可以使用 “射线尺度”选项调整射线的尺度，以便更加清晰地显示。

双标图等高线密度 支持您指定密度等高线的置信水平。默认置信水平为 90%。

标记聚类 将标识聚类的标记分配给数据表的行。

三维双标图 显示数据的三维双标图。仅当有三个或更多变量时可用。

平行坐标图 为每个聚类创建平行坐标图。图报表提供用于显示和隐藏数据和均值的选项。请参见 《基本绘图》。

散点图矩阵 显示或隐藏使用所有 Y 变量的散点图矩阵。每个散点图都包含基于当前聚类数的密度椭圆。

SOM 热图 （仅可用于 “自组织图”。）显示或隐藏自组织图聚类均值的热图，按聚类中使用的其中一个 Y 变量着色。使用 “选择列”旁边的菜单对热图着色以更改 Y 变量。

注意：热图上的聚类以自上而下、从右到左的布局进行组织。这意味着第一个聚类位于右上角，最后一个聚类位于左下角。

将颜色保存至表 将标识聚类的颜色分配给数据表的行。若报表窗口中有双标图，保存到数据表的颜色将与双标图中的聚类颜色相匹配。若双标图中的颜色改变，而且再次选定 “将颜色保存至表”选项，那么表中的颜色将更新以便与双标图中的那些颜色相匹配。

注意：选定任何保存选项时，每个保存的列都包含一个 “注释”列属性，该属性指定这一特定列数据的聚类数。这使您能够保存来自多个聚类拟合的列，并使用列属性来标识保存的列来自哪个聚类拟合。

保存聚类 将以下两列保存至数据表：

- **聚类**列包含分配了给定行的聚类的编号。
- （不适用于“自组织图”。）**距离**列包含给定观测与其聚类均值之间的欧氏距离平方。对于每个变量，将观测值与该变量的聚类均值之间的差值除以该变量的总标准差。然后对所有变量的这些统一尺度的差值进行平方与求和计算。

保存聚类距离 （不适用于“自组织图”。）将**距离**列保存到数据表中。该列与从**保存聚类**选项获得的**距离**列相同。

保存自组织图网格 （仅可用于“自组织图”。）将新列保存到数据表中。新列包含每个观测最可能的聚类的 SOM 网格行和列编号。

保存聚类公式 将名为“聚类公式”的公式列保存至数据表。这是标识每个聚类的聚类成员关系的公式。

保存距离公式 （不适用于“自组织图”。）将名为“距离公式”的公式列保存至数据表。这是用于计算到所分配聚类的距离的公式。

保存 K 聚类距离 （不适用于“自组织图”。）保存包含到每个聚类中心的欧氏距离平方的 *k* 列。

保存 K 距离公式 （不适用于“自组织图”。）保存包含到每个聚类中心的欧氏距离平方的公式的 *k* 列。

发布聚类公式 向“公式存储库”发布在“保存聚类公式”选项中使用的相同的得分代码。

模拟聚类 使用聚类均值和标准差，创建包含 Y 变量的模拟聚类观测的新数据表。

删除 删除聚类报表。

自组织图

自组织图 (SOM) 方法最初由 Teuvo Kohonen (1989, 1990) 开发，之后被另外一些神经网络爱好者和统计学家进一步推广。您可以在“K 均值聚类”平台中实施 SOM 方法。有关示例，请参见“‘K 均值聚类’的更多示例”。

最初的 SOM 被视为一个学习过程，就像最初的神经网络算法一样，但在此使用的版本是 *K*- 均值聚类的变异。在 SOM 文献中该变异称为使用**局部加权线性平滑统计量**的**批处理算法**。

SOM 的目标不仅仅是要在聚类网格上以特定的布点形成聚类，那些在 SOM 网格中彼此邻近的聚类中的点在多元空间中也是彼此邻近的。在经典的 *K*-均值聚类中，聚类结构是任意的，但在 SOM 中，聚类具有网格结构。该网格结构可帮助在二维中解释聚类：相距较近的聚类比起相距较远的聚类要更类似。请参见“SOM 算法的说明”。

“自组织图”控制面板

在“K 均值聚类”平台中，从“迭代聚类控制面板”中的“方法”列表中选择“自组织图”选项。

图 14.8 “控制面板”中的“自组织图”选项

“‘迭代聚类”报表”中说明了面板上的部分选项。其余选项在下文中进行说明。

行数 聚类网格中的行数。

列数 聚类网格中的列数。

带宽 指定邻近聚类对预测重心的影响。较小的带宽会对更近的聚类指定更大的权重。

“自组织图”报表

在“K 均值聚类”平台中，SOM 报表根据请求的网格大小进行命名。“带宽”在“SOM 网格”报表的顶部给出。该报表本身类似于“K 均值聚类数”报表。请参见“‘K 均值聚类数”报表”。“聚类比较报表”显示聚类总数以及请求的行数。

有关“自组织图”红色小三角选项的详细信息，请参见“‘K 均值聚类数”报表选项”。

图 14.9 “自组织图”报表


迭代聚类

聚类比较

方法	聚类数	行数	CCC	最佳
自组织图	6	2	2.87879	最优 CCC

逐列统一尺度

控制面板

2 X 3 的自组织图网格

逐列统一尺度

带宽: 0.4330127

聚类汇总

聚类	计数	步进	准则
1	28	12	0
2	20		
3	31		
4	21		
5	28		
6	22		

聚类均值

聚类	萼片长度	萼片宽度	花瓣长度	花瓣宽度
1	6.98187076	3.15339949	5.85011658	2.16765916
2	6.14554483	2.62528195	5.11988902	1.79042135
3	6.20930867	3.05499872	4.51264582	1.46300228
4	5.46721399	2.56978376	3.79572186	1.15921467
5	5.22582725	3.67826605	1.47738883	0.27226749
6	4.7258207	3.10958329	1.44133246	0.21218254

聚类标准差

SOM 算法的说明

本节包含在 “K 均值聚类”平台中实施 SOM 的步骤。

- 初始聚类种子采用可为多维空间提供良好覆盖率的方式进行选择。JMP 使用主成分来确定两个方向，用来捕获数据中的大多数变异。
- 随后，JMP 在这个主成分空间中展开一个网格，网格边缘与每个方向的中心相距 2.5 个标准差。通过将该网格转换回变量初始空间来确定聚类种子。
- 按照 *K*-均值方法那样继续聚类分配，将每个点分配给最靠近它的聚类。
- 像使用 *k* 均值方法那样为每个聚类估计均值。JMP 随后使用这些均值设置加权回归，它将每个变量设置为回归中的响应，并将 SOM 网格坐标设置为回归变量。权重函数使用核心函数将较大权重分配给要估计其中心的聚类，将较小权重分配给 SOM 网格中距离该聚类较远的聚类。新聚类均值是从该回归得到的预测值。
- 这些迭代继续执行，直到该过程收敛。

“K 均值聚类”的更多示例

本例使用自组织图 (SOM) 将鸢尾花的观测聚类到其正确的品种中。

1. 选择**帮助** > **样本数据文件夹**，然后打开 Iris.jmp。

2. 选择**分析 > 聚类 > K 均值聚类**。
3. 选择**萼片长度**、**萼片宽度**、**花瓣长度**和**花瓣宽度**，然后点击 **Y，列**。
4. 点击**确定**。
5. 从“控制面板”上的“方法”菜单中选择**自组织图**。
6. 设置**行数**等于 1，**列数**等于 2。
7. 点击**执行**。
8. 打开“控制面板”报表。
9. 设置**行数**等于 1，**列数**等于 3。
10. 点击**执行**。
11. 打开“控制面板”报表。
12. 设置**行数**等于 2，**列数**等于 2。
13. 点击**执行**。

图 14.10 SOM 聚类比较

聚类比较

方法	聚类数	行数	CCC	最佳
自组织图	2	1	3.35952	
自组织图	3	1	4.95837	最优 CCC
自组织图	4	2	3.64938	

“聚类比较”报表显示在报表窗口顶部。最佳拟合由最高 CCC 值确定。请注意：给出最大 CCC 的聚类编号为 3，即物种数。

14. 滚动到“1 x 3 的 SOM 网格”报表。我们可以看到该分类不完美；每个聚类应表示每个物种，每个聚类有 50 行。

图 14.11 Iris.jmp 的“自组织图”报表

1 X 3 的自组织图网格

逐列统一尺度

带宽: 0.4330127

聚类汇总

聚类	计数	步进	准则
1	45	10	0
2	55		
3	50		

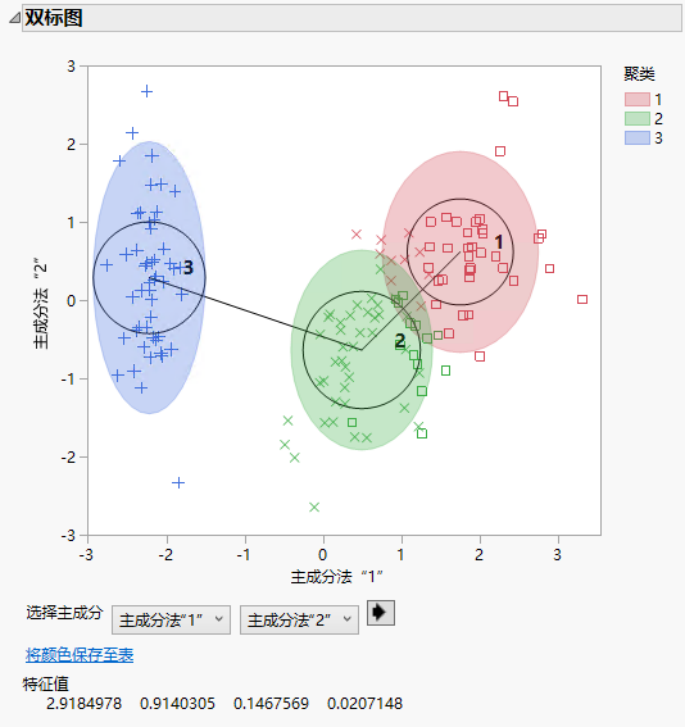
聚类均值

聚类	萼片长度	萼片宽度	花瓣长度	花瓣宽度
1	6.81555029	3.09552593	5.54004585	1.98001626
2	5.81996034	2.7503061	4.29253797	1.39367974
3	5.00599574	3.427976	1.46203714	0.24601317

15. 在该数据表中，选择**物种**列并选择**行 > 按列设定颜色或标记**。
16. 选择“标记”下面的**经典**选项。
17. 点击**确定**。

18. 点击 “1 x 3 的 SOM 网格”旁边的红色小三角菜单，然后选择**双标图**。
19. 点击 “1 x 3 的 SOM 网格”旁边的红色小三角菜单，然后选择**双标图选项 > 显示双标图射线**。

图 14.12 SOM 双标图



我们可以看到“Cluster 3”中的所有行都正确标识为 setosa 物种。另外两个物种，virginica 和 versicolor 略有重叠，可能彼此混淆。

第 15 章

正态混合

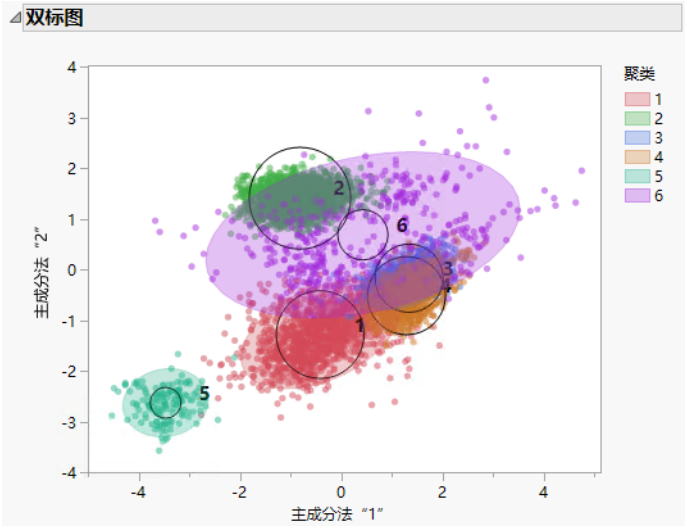
使用概率对观测分组

“正态混合”是基于假设使用多元正态分布的混合分布近似计算观测的联合概率分布的迭代方法。这些混合代表不同的聚类。各聚类都具有多元正态分布。

若聚类分隔地很好，层次聚类和 *K*-均值聚类都适用。但若聚类重叠，则正态混合是更好的替代方法，因为它基于聚类成员概率，而不是基于边界的任意聚类分配。

当数据来自重叠正态分布时，使用“正态混合”进行聚类。您需要提前指定聚类数。

图 15.1 正态混合双标图



目录

“正态混合”平台概述

正态混合是针对数值变量的迭代聚类方法。不过，它也可预测每个聚类内期望的响应比例。正态混合假设测量值列的联合概率分布可以由多元正态的混合分布近似计算，混合分布代表了不同的聚类。为每个聚类估计均值向量和协方差矩阵。请参见 McLachlan and Krishnan (1997) 和 Section 9.判别主成分 in Hand et al. (2001)。

注意：“正态混合”算法涉及的迭代首先要对聚类中心进行随机猜测。因此，每次运行分析的结果可能略有不同。

若您怀疑有多元离群值，可使用两个选项。您可以使用离群值聚类或“探索离群值”实用工具。离群值聚类选项假定服从均匀分布，与标准的“正态混合”选项相比对离群值更不敏感。“探索离群值”实用工具支持您在分析之前探索和处理离群值。请参见“离群值聚类”和《预测和专业建模》。

“正态混合”是 JMP 提供的对观测进行聚类的四个平台之一。有关四种方法的比较，请参见“对观测聚类的平台概述”。

对观测聚类的平台概述

聚类是将在几个变量上享有相似值的观测分组在一起的一种多元方法。通常情况下，观测在 p 维空间内散布不均，其中 p 是变量数。这些观测反而会形成聚簇或聚类。标识出这些聚类使您可以更深层次地了解您的数据。

注意：JMP 还提供可以对变量聚类的平台。请参见“聚类变量”。

JMP 提供四个平台供您对观测聚类：

- 层次聚类对于小型和大型数据表非常有用，并且允许使用字符数据。“层次聚类”将行按描绘为一棵树的层次序列形式进行组合。您可以在生成树后选择最适合您数据的聚类数。请参见“层次聚类”。
- “K 均值聚类”适用于多达数百万行的大型表，并且只允许数值数据。您需要提前指定聚类数 k。该算法可以对聚类种子点做出推测。随后开始在将数据点分配到相应类别和重新计算聚类中心之间交替进行迭代过程。请参见“K 均值聚类”。
- “正态混合”适用于数据来自重叠的多元正态分布的混合分布这种情况，并且只允许数值数据。对于具有多元离群值的情形，您可以使用假设具有均匀分布的离群值聚类。请参见“正态混合”。

  您需要提前指定聚类数。最大似然用于同时估计混合比例以及均值、标准差和相关性。为每个点指定属于每个组的概率。使用 EM 算法获取估计值。
- “潜在类分析”适用于大多数变量是分类变量这种情况。您需要提前指定聚类数。该算法拟合假定具有多项式混合分布的模型。为每个观测计算聚类成员关系的最大似然估计值。观测会被归类到其成员关系概率最大的聚类中。请参见“潜在类分析”。

表 15.1 聚类方法汇总

方法	数据类型或建模类型	数据表大小	指定聚类数
层次聚类	任意	使用混合 Ward 法，最多数十万行 使用快速 Ward 法，最多 200,000 行 使用其他方法，最多 5,000 行	否
K 均值聚类	数值	多达数百万行	是
正态混合	数值	任意大小	是
潜在类分析	名义型或有序型	任意大小	是

有些聚类平台提供用于处理数据中的离群值的选项。但是，若数据中有离群值，则最好在分析之前探索这些离群值。可以使用“探索离群值”实用工具完成该操作。有关详细信息，请参见《预测和专业建模》。

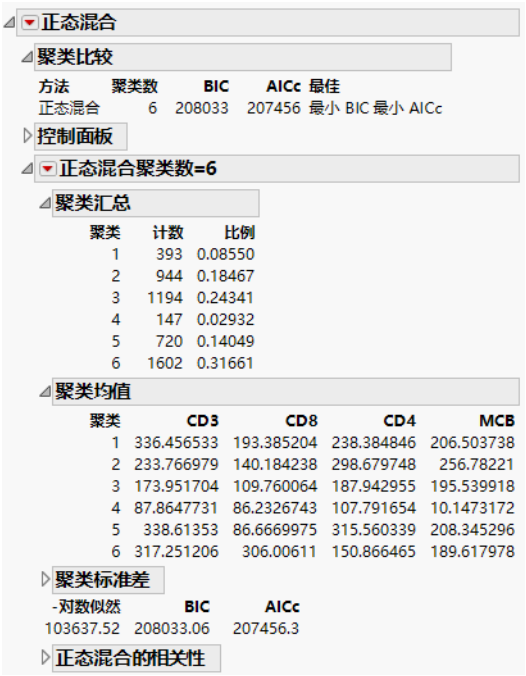
“正态混合聚类”的示例

在本例中，您使用“正态混合”平台基于血细胞计数分析中四种标记的读数来对观测进行聚类。血细胞计数用于测量细胞的各种特征。细胞标记测量值有助于诊断特定疾病。

1. 选择**帮助** > **样本数据文件夹**，然后打开 Cytometry.jmp。
2. 选择**分析** > **聚类** > **正态混合**。
3. 选择 CD3、 CD8、 CD4 和 MCB，然后点击 **Y, 列**。
4. 点击**确定**。
5. 在**聚类数**旁边输入 判别主成分。
6. 点击**执行**。

注意：您的结果可能有所不同，因为算法具有随机起始值。

图 15.2 “正态混合聚类数 = 判别主成分”报表


正态混合

聚类比较

方法	聚类数	BIC	AICc	最佳
正态混合	6	208033	207456	最小 BIC 最小 AICc

控制面板

正态混合聚类数=6

聚类汇总

聚类	计数	比例
1	393	0.08550
2	944	0.18467
3	1194	0.24341
4	147	0.02932
5	720	0.14049
6	1602	0.31661

聚类均值

聚类	CD3	CD8	CD4	MCB
1	336.456533	193.385204	238.384846	206.503738
2	233.766979	140.184238	298.679748	256.78221
3	173.951704	109.760064	187.942955	195.539918
4	87.8647731	86.2326743	107.791654	10.1473172
5	338.61353	86.6669975	315.560339	208.345296
6	317.251206	306.00611	150.866465	189.617978

聚类标准差

-对数似然	BIC	AICc
103637.52	208033.06	207456.3

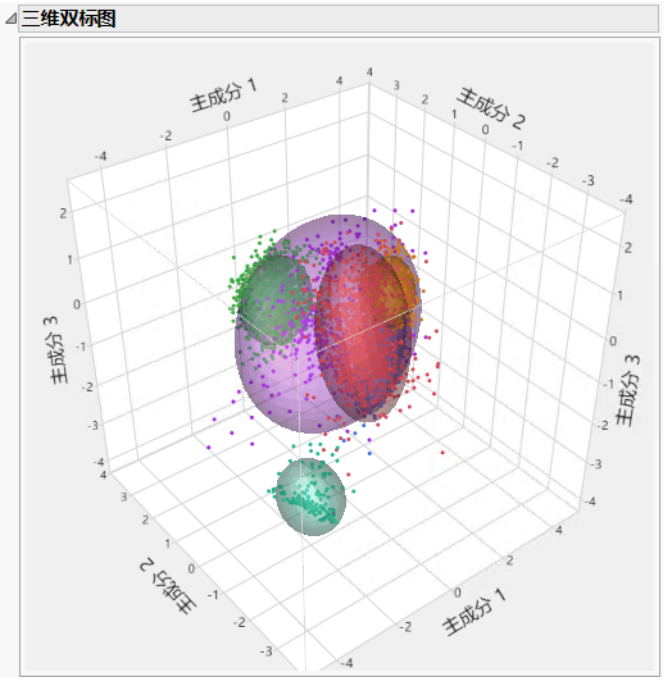
正态混合的相关性

“聚类汇总”报表显示 判别主成分 个聚类中每一个聚类的观测数。“聚类均值”报表显示每个聚类的四个标记读数的均值。

7. 点击 “正态混合聚类数 = 判别主成分”旁边的红色小三角，然后选择**三维双标图**。

注意：您的三维双标图可能有所不同，因为算法具有随机起始值。

图 15.3 血细胞计数数据的 “三维双标图”

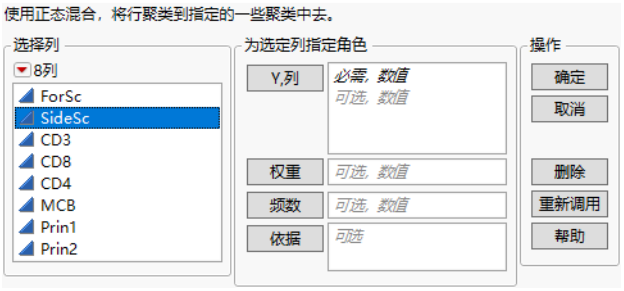


该图显示聚类的拟合正态密度等高线。注意到依据前 3 个主成分，一个聚类明显与其他聚类分开。

启动“正态混合”平台

通过选择**分析 > 聚类 > 正态混合**，启动 “正态混合”平台。

图 15.4 “正态混合”启动窗口



有关“选择列”红色小三角菜单中选项的详细信息，请参见《使用 JMP》。

Y，列 用于对观测聚类的变量。

注意：正态混合聚类仅支持数值列。

权重 一列，该列的数值为分析中的每一行都分配一个权重。

频数 一列，列中的数值为分析中的每行分配一个频数。

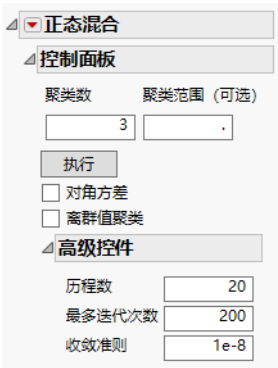
依据 一列，其水平定义不同的分析。对于指定列的每个水平，都分析相应行。结果显示在不同的报表中。若分配了多个“依据”变量，则为“依据”变量水平的每个可能组合生成单独分析。

当您点击“确定”时，“控制面板”随即显示。请参见“单个“正态混合”报表”。

“正态混合”报表

当您在“正态混合”启动窗口中点击“确定”时，“正态混合”报表窗口将显示用于拟合模型的“控制面板”。您可以通过“控制面板”反复拟合各种数量的聚类，也可以使用“聚类范围”选项指定一个范围。在拟合模型时，其他报表会添加至窗口。请参见“单个“正态混合”报表”。

图 15.5 “正态混合”的“控制面板”



“正态混合控制面板”包含以下选项：

聚类数 指定要形成的聚类数。

聚类范围（可选） 提供要形成的聚类数的上限。若在此输入了某个数字，平台将为**聚类数**和**聚类范围（可选）**中输入的值之间的每个整数创建单独的分析。

执行 拟合聚类。

对角方差 将协方差矩阵的非对角线元素限定为 0。该平台拟合变量之间无相关性的多元正态分布。

注意：有时候需要使用“对角方差”选项，这样在观测数少于变量数时可避免生成奇异的协方差矩阵。该选项也可用于避免对大量变量估计非常大的协方差矩阵。

离群值聚类 拟合聚类，将未能落入任何正常聚类中的离群值收拢其中。若创建了该聚类，则将其指定为 “聚类 0”，并且观测计数显示在 “聚类汇总” 报表中。假定落入离群值聚类中的观测在包含这些观测的超立方中服从均匀分布。

高级控件 以下高级控件可用：

历程数 估计过程的独立重新开始数。每次重新开始具有不同的起始值。独立开始有助于防止求得局部解。

最多迭代次数 EM 算法收敛阶段的最多迭代次数。

收敛准则 EM 迭代停止时的似然差值。

正态混合选项

本节介绍 “正态混合” 红色小三角菜单中的选项。

请参见《使用 JMP》获取有关下列选项的信息：

本地数据过滤器 显示或隐藏支持您过滤特定报表中使用的数据的本地数据过滤器。

重新运行 包含使您可以重复或重新启动分析的选项。在支持该功能的平台中，“自动重新计算” 选项立即在相应报表窗口中反映您对数据表所做的更改。

平台首选项 包含的选项支持您查看当前平台首选项或更新平台首选项以匹配当前 JMP 报表中的设置。

保存脚本 包含的选项支持您保存可将报表重现到若干目标的脚本。

单个 “正态混合” 报表

在 “正态混合控制面板” 中点击 “执行” 后，将显示一个聚类比较报表以及一个或多个单独的 “正态混合” 报表。报表被动态地命名为 “正态混合聚类数 =<k>”，具体取决于 *k*，即拟合的聚类数。每次执行拟合时，都会出现 “正态混合聚类数 =<k>” 报表。

- “ “聚类比较” 报表”
- “ “正态混合聚类数” 报表”

“聚类比较” 报表

在 “正态混合” 平台中，“聚类比较” 报表提供拟合统计量以比较各种模型。拟合统计量为 BIC 和 AICc。每个统计量的值越小表示拟合效果越好。最佳拟合在名为最佳的列中指示。

“正态混合聚类数” 报表

每个 “正态混合聚类数” 报表都提供每个聚类的汇总统计量：

- “聚类汇总” 报表提供每个聚类的观测数和比例。

- “聚类均值”报表为每个变量提供每个聚类中观测的均值。
- “聚类标准差”报表为每个变量提供每个聚类中观测的标准差。
- “负对数似然”表提供负对数似然、BIC 和 AICc。请参见《拟合线性模型》。
- “正态混合的相关性”报表提供每个聚类的估计相关性矩阵。

“正态混合聚类数”报表选项

每个 “正态混合聚类数”报表都包含以下红色小三角菜单项：

双标图 以数据的前两个主成分显示点和聚类的图，以及用于标识聚类颜色的图例。围绕聚类中心绘制圆圈，而且圆圈的大小与聚类内的计数成比例。着色区域是围绕均值的密度等高线。默认情况下，该区域指示该聚类中 90% 的观测所在的位置 (Mardia et al. 1980)。使用该图下方的列表可将图轴改为其他主成分。或者，使用箭头按钮在所有可能的轴组合之间循环切换。该图之下还有一个用于将聚类颜色保存到数据表的选项。请参见“将颜色保存至表”。特征值以降序方式显示。

注意：双标图始终使用相关性矩阵计算主成分。

双标图选项 包含用于控制双标图外观的选项。

显示双标图射线 显示双标图射线。带标签的射线显示协变量在由主成分定义的子空间中的方向。它们表示每个变量与每个主成分的关联程度。

双标图射线位置 可让您指定双标图射线的位置和射线尺度。默认情况下，这些射线从点 (0,0) 发出。在该图中，您可以拖动射线或使用该选项指定坐标。您还可以使用“射线尺度”选项调整射线的尺度，以便更加清晰地显示。

双标图等高线密度 支持您指定密度等高线的置信水平。默认置信水平为 90%。

标记聚类 将标识聚类的标记分配给数据表的行。

三维双标图 显示数据的三维双标图。仅当有三个或更多变量时可用。

平行坐标图 为每个聚类创建平行坐标图。图报表提供用于显示和隐藏数据和均值的选项。请参见 《基本绘图》。

散点图矩阵 使用所有 Y 变量创建散点图矩阵。

将颜色保存至表 将标识聚类的颜色分配给数据表的行。若报表窗口中有双标图，保存到数据表的颜色将与双标图中的聚类颜色相匹配。若双标图中的颜色改变，而再次选定“将颜色保存至表”选项，那么表中的颜色将更新以便与双标图中的那些颜色相匹配。

注意：选定任何保存选项时，每个保存的列都包含一个“注释”列属性，该属性指定这一特定列数据的聚类数。这使您能够保存来自多个聚类拟合的列，并使用列属性来标识保存的列来自哪个聚类拟合。

保存聚类 将名为**聚类**的列添加到数据表，该列包含分配了给定行的聚类的编号。对于正态混合，这是最有可能的聚类。

保存聚类公式 将名为**聚类公式**的公式列添加至数据表。该公式标识行属于哪个聚类。

发布聚类公式 向“公式存储库”发布在“保存聚类公式”选项中使用的相同的得分代码。若已选定“发布聚类公式”并且从“公式存储库”中的模型选择“运行脚本”后，保存到数据表的列应与选定“保存聚类公式”时保存的那些列相匹配。

保存混合概率 为每个聚类添加名为**概率聚类** <k> 的列，相应列包含观测属于该聚类的概率。

保存混合公式 将列添加至数据表，这些列包含用于计算混合概率的公式。使用这些公式列对已排除数据或添加到表中的数据的概率进行评分。

距离公式 <k> 在观测处求得的“聚类 <k>”的多元正态密度函数的估计。

总距离 距离公式列的总和。该列中的公式等价于由“保存密度公式”选项创建的**混合密度**列中的公式。

概率公式 <k> 观测属于“聚类 <k>”的概率。这些列包含为由“保存混合概率”选项创建的**概率聚类** <k> 列提供值的公式。计算混合概率的列公式为：

$$\text{概率公式 <k>} = \frac{\text{距离公式 <k>}}{\text{总距离}}$$

保存密度公式 将名为**混合密度**的列添加至数据表，该列包含正态混合的估计密度函数。

模拟聚类 使用混合密度模拟预测变量值。将这些值和值所属的聚类保存至新数据表中。

删除 删除聚类报表。

“正态混合”平台的统计详细信息

“正态混合”使用 EM 算法执行拟合，因为该算法比 Newton-Raphson 算法更稳定。此外，JMP 使用 EM 算法的 Bayes 正则版本，该版本允许我们顺利处理协方差矩阵奇异的情况。由于估计值极大地依赖于初始推测数，该平台将迭代若干历程，每一历程都随机选择一些点作为初始中心。

执行多个历程会令估计过程花费较长时间，所以处理较大问题时需要耐心等待。控件支持您指定历程和迭代限制。

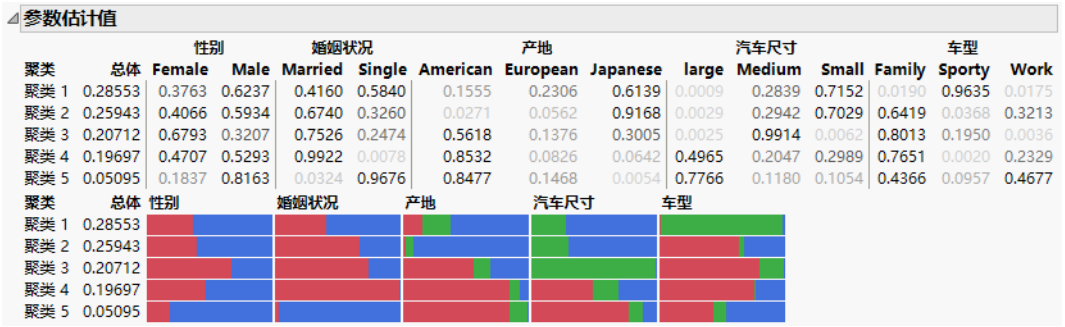
第 16 章

潜在类分析

将分类变量的观测进行分组

潜在类分析可使您针对分类响应变量进行观测值的聚类。潜在变量是无法观测的分组变量。潜在变量的每个水平称为潜在类。“潜在类分析”平台拟合潜在类模型并确定每个观测的最可能的聚类或潜在类。在大多数情况下，主题专家使用潜在类分析的结果基于每个类的特征创建相应潜在类的定义。

图 16.1 “潜在类分析”的示例



参数估计值

		性别		婚姻状况		产地			汽车尺寸			车型		
聚类	总体	Female	Male	Married	Single	American	European	Japanese	large	Medium	Small	Family	Sporty	Work
聚类 1	0.28553	0.3763	0.6237	0.4160	0.5840	0.1555	0.2306	0.6139	0.0009	0.2839	0.7152	0.0190	0.9635	0.0175
聚类 2	0.25943	0.4066	0.5934	0.6740	0.3260	0.0271	0.0562	0.9168	0.0029	0.2942	0.7029	0.6419	0.0368	0.3213
聚类 3	0.20712	0.6793	0.3207	0.7526	0.2474	0.5618	0.1376	0.3005	0.0025	0.9914	0.0062	0.8013	0.1950	0.0036
聚类 4	0.19697	0.4707	0.5293	0.9922	0.0078	0.8532	0.0826	0.0642	0.4965	0.2047	0.2989	0.7651	0.0020	0.2329
聚类 5	0.05095	0.1837	0.8163	0.0324	0.9676	0.8477	0.1468	0.0054	0.7766	0.1180	0.1054	0.4366	0.0957	0.4677

聚类	总体	性别	婚姻状况	产地	汽车尺寸	车型
聚类 1	0.28553					
聚类 2	0.25943					
聚类 3	0.20712					
聚类 4	0.19697					
聚类 5	0.05095					

目录

“潜在类分析”平台概述

“潜在类分析”平台对分类响应变量拟合潜在类模型并确定每个观测的最可能的聚类或潜在类。**潜在变量**是无法观测的分组变量。潜在变量的每个水平称为**潜在类**。例如，潜在类可以是按其风险偏好分组的调查响应者的聚类。

该模型采取多项式混合模型的形式。模型中有两组参数：γ 参数和 ρ 参数。γ 参数表示聚类成员关系的总概率。ρ 参数表示在聚类成员关系给定的条件下观测到给定响应的概率。这些条件概率的模式构成了潜在类的特征。

为使分析结果具有意义，主题专家必须对平台生成的聚类进行解释。该主题专家检查潜在类的特征并根据这些特征构建每个类的定义。

注意：在任何响应列中具有缺失值的行都会从分析中排除。

有关潜在类模型的详细信息，请参见 Collins and Lanza (2010) 和 Goodman (1974)。

“潜在类分析”是 JMP 提供的对观测进行聚类的四个平台之一。有关四种方法的比较，请参见“对观测聚类的平台概述”。

对观测聚类的平台概述

聚类是将在几个变量上享有相似值的观测分组在一起的一种多元方法。通常情况下，观测在 p 维空间内散布不均，其中 p 是变量数。这些观测反而会形成聚簇或聚类。标识出这些聚类使您可以更深层次地了解您的数据。

注意：JMP 还提供可以对变量聚类的平台。请参见“聚类变量”。

JMP 提供四个平台供您对观测聚类：

- 层次聚类对于小型和大型数据表非常有用，并且允许使用字符数据。“层次聚类”将行按描绘为一棵树的层次序列形式进行组合。您可以在生成树后选择最适合您数据的聚类数。请参见“层次聚类”。
- “K 均值聚类”适用于多达数百万行的大型表，并且只允许数值数据。您需要提前指定聚类数 k。该算法可以对聚类种子点做出推测。随后开始在将数据点分配到相应类别和重新计算聚类中心之间交替进行迭代过程。请参见“K 均值聚类”。
- “正态混合”适用于数据来自重叠的多元正态分布的混合分布这种情况，并且只允许数值数据。对于具有多元离群值的情形，您可以使用假设具有均匀分布的离群值聚类。请参见“正态混合”。

  您需要提前指定聚类数。最大似然用于同时估计混合比例以及均值、标准差和相关性。为每个点指定属于每个组的概率。使用 EM 算法获取估计值。

- “潜在类分析”适用于大多数变量是分类变量这种情况。您需要提前指定聚类数。该算法拟合假定具有多项式混合分布的模型。为每个观测计算聚类成员关系的最大似然估计值。观测会被归类到其成员关系概率最大的聚类中。请参见“潜在类分析”。

表 16.1 聚类方法汇总

方法	数据类型或建模类型	数据表大小	指定聚类数
层次聚类	任意	使用混合 Ward 法，最多数十万行 使用快速 Ward 法，最多 200,000 行 使用其他方法，最多 5,000 行	否
K 均值聚类	数值	多达数百万行	是
正态混合	数值	任意大小	是
潜在类分析	名义型或有序型	任意大小	是

有些聚类平台提供用于处理数据中的离群值的选项。但是，若数据中有离群值，则最好在分析之前探索这些离群值。可以使用“探索离群值”实用工具完成该操作。有关详细信息，请参见《预测和专业建模》。

“潜在类分析”的示例

在本例中，您根据学生对 12 个调查问题的响应来拟合潜在类模型，以识别学生聚类。这些响应来自 2005 年的美国高中生调查。该调查针对学生提出了各种有关健康风险行为的多项选择题。从多项选择调查问题中将响应分为两类（是/否），可以获得您分析的列。

1. 选择**帮助 > 样本数据文件夹**，然后打开 Health Risk Survey.jmp。
2. 在 “Health Risk Survey”数据表中，点击**启动潜在类分析平台**脚本旁边的绿色小三角。

   该脚本选择 12 个所关注的列，打开“潜在类分析”启动窗口，并且输入这 12 个关注的列作为 “Y”。

   注意：要自行启动 LCA 平台，请选择 “分析” > “聚类” > “潜在类分析”。
3. 在**多达**旁边的框中键入 5。

   该选项针对 3 到 5 个 （最多 5 个）聚类拟合潜在类模型。
4. 点击**确定**。

图 16.2 “聚类比较”报表

聚类比较

聚类数	-对数似然	BIC	AIC	最佳
3	38713	77776.3	77502	
4	38207.1	76884.4	76516.3	
5	37964.8	76519.6	76057.6	最小 BIC 最小 AIC

“潜在类分析”分级显示项包括一个 “聚类比较”报表和 3 个单独的 “潜在类模型”报表。“潜在类模型”报表显示 3 个、 4 个和 5 个聚类的模型。在 “聚类比较”报表中，具有 5 个聚类的模型具有最小的 BIC 和 AIC，这表明该模型是这 3 个模型中的最佳拟合模型。您分析的是该模型。

5. 在 “‘5’个聚类的潜在类模型”报表中，检查 “参数估计值”下方的条形图。请注意以下事项：
   - “聚类 1”对所有风险行为的大部分回答多为“No”。
   - “聚类 2”对 13 岁前的 4 种风险行为的回答多为“Yes”。
   - “聚类 3”对酒驾和过去 30 天内喝过至少 5 杯酒的很多回答多为“Yes”。
   - “聚类 4”对除 13 岁之前的风险行为之外的其他大部分风险行为的回答多为“Yes”。
   - “聚类 5”对大部分风险行为的回答为“Yes”最多。

   使用该信息为聚类提供有意义的名称。
6. 点击 “‘5’个聚类的潜在类模型”旁边的红色小三角，选择**重命名聚类**：
   - 为“聚类 1”输入“低风险”。
   - 为“聚类 2”输入“早期风险承担者”。
   - 为“聚类 3”输入“嗜酒者”。
   - 为“聚类 4”输入“后期高风险”。
   - 为“聚类 5”输入“高风险”。
7. 点击**确定**。
8. 在出现的 “JMP 警示”中点击**确定**。

注意：新聚类名称不会保存到脚本中。

图 16.3 部分 “参数估计值”报表

参数估计值

聚类	总体	喝酒驾车 1 次以上		13 岁之前吸过烟		30 天中每天吸烟		13 岁前喝过第一杯酒		过去 30 天喝过至少五杯酒 1 次以上		13 岁前尝试过大麻		一生中使用过可卡因 1 次以上		一生中吸过胶毒 1 次以上	
		No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Low Risk	0.67525	0.9872	0.0128	0.9624	0.0376	0.9741	0.0259	0.8648	0.1352	0.8975	0.1025	0.9943	0.0057	0.9930	0.0070	0.9440	0.0560
Early Risk-Takers	0.13852	0.8963	0.1037	0.4457	0.5543	0.7408	0.2592	0.3418	0.6582	0.5965	0.4035	0.6539	0.3461	0.9517	0.0483	0.8209	0.1791
Drinkers	0.09626	0.4307	0.5693	0.8662	0.1338	0.7619	0.2381	0.7610	0.2390	0.0368	0.9632	0.9665	0.0335	0.9092	0.0908	0.8689	0.1311
Late High Risk	0.05146	0.6848	0.3152	0.8223	0.1777	0.4921	0.5079	0.7703	0.2297	0.3494	0.6506	0.9114	0.0886	0.2685	0.7315	0.5432	0.4568
High Risk	0.03851	0.4943	0.5057	0.1333	0.8667	0.2882	0.7118	0.1188	0.8812	0.1270	0.8730	0.2145	0.7855	0.1650	0.8350	0.4711	0.5289

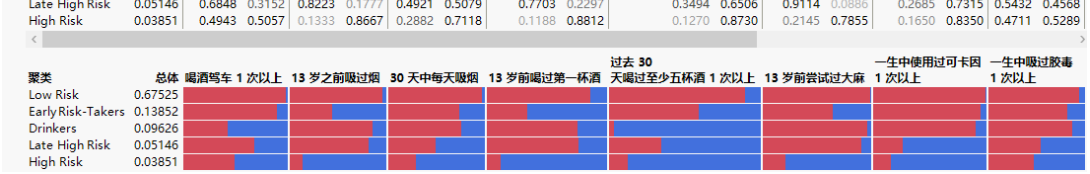
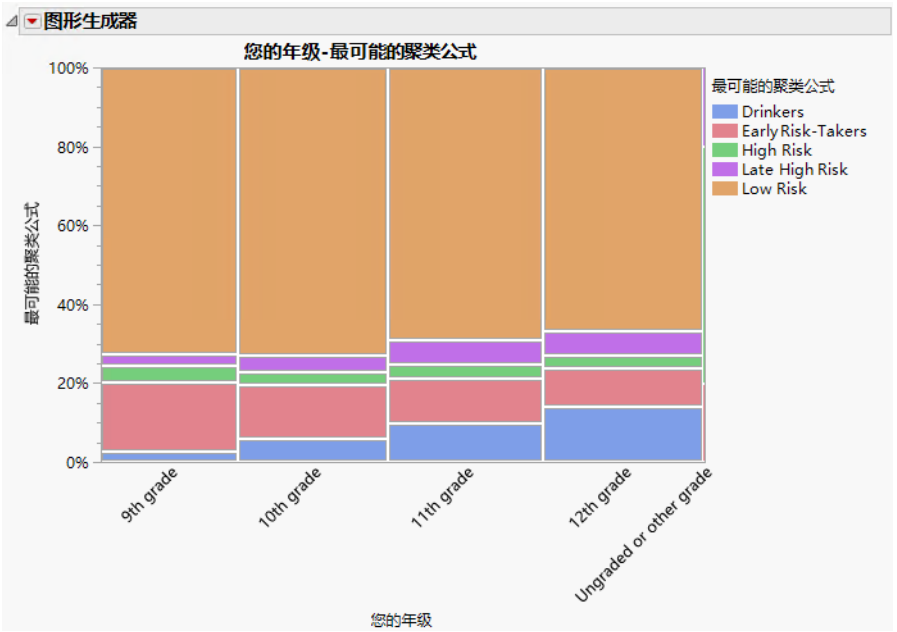


图 16.3 显示分析中前 8 个变量的参数估计值。新聚类名称出现在报表窗口中。

接下来，将聚类成员关系与人口统计学问题 “您的年级”进行比较。

9. 点击 “‘5’个聚类的潜在类模型”旁边的红色小三角，选择**保存混合和聚类公式**。
10. 选择**图形 > 图形生成器**。
11. 输入**您的年级**作为 X。
12. 输入**最可能的聚类公式**作为 Y。
13. 选择 “马赛克图”元素。
14. 点击**完成**。

图 16.4 “年级 - 聚类成员关系”的马赛克图

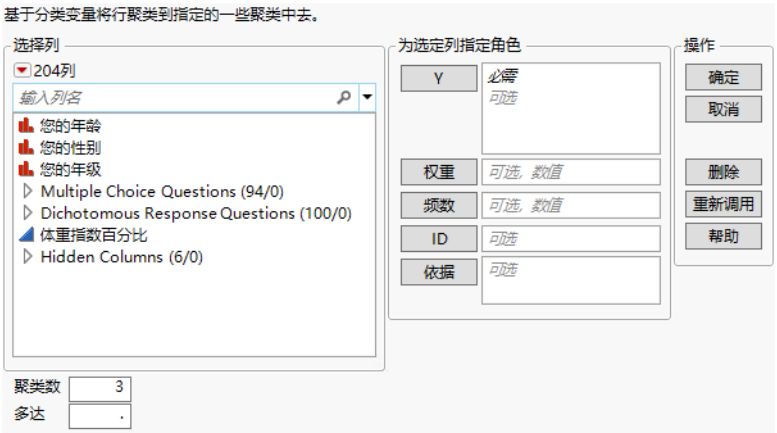


观测到大多数响应者落入 “低风险” 聚类。标有 “嗜酒者” 的类随着年级增长，响应者人数增加。

启动“潜在类分析”平台

通过选择**分析 > 聚类 > 潜在类分析**，启动 “潜在类分析” 平台。

图 16.5 “潜在类分析” 启动窗口



有关 “选择列” 红色小三角菜单中选项的详细信息，请参见《使用 JMP》。

“潜在类分析” 平台启动窗口包括以下选项：

Y 您要分析的一个或多个列。您可以分析建模类型为名义型、有序型或多重响应的列。要分析名义型或有序型响应，需要两列或更多列。若包含多重响应并且建模类型为多重响应，则只需一列。

权重 一列，该列的数值为分析中的每一行都分配一个权重。

频数 一列，列中的数值为分析中的每行分配一个频数。

ID 用于标识单独响应者的列。该标识在某些输出表中使用。

依据 一列，用于创建为变量的每个水平包含单独分析的报表。若分配了多个 “依据” 变量，则为 “依据” 变量水平的每个可能组合生成单独分析。

聚类数 要在分析中计算的聚类数。

多达 指定最大聚类数。若该数值超过为 “聚类数” 指定的值，则生成模型报表中的聚类数为介于 “聚类数” 和 “多达” 范围内的每个整数值。这些报表显示为 “潜在类分析” 报表分级显示项的一部分。

警告：LCA 模型可充分拟合的聚类数存在最大值。若您请求的聚类数超过最大值，报表窗口中会显示一条警告消息。LCA 平台拟合的聚类数上限是各列所支持的最大聚类数。有关确定最大聚类数的详细信息，请参见“最大聚类数的统计详细信息”。

“潜在类分析”报表

初始“潜在类分析”报表包含每个指定聚类数的“聚类比较”报表和“潜在类模型”报表。

“聚类比较”报表

在“潜在类分析”平台中，“聚类比较”报表显示用于比较各种模型的拟合统计量。拟合统计量包括负对数似然（-对数似然）、BIC 和 AIC。每个统计量的值越小表示拟合效果越好。最佳拟合在名为**最佳**的列中指示。请参见《拟合线性模型》。

潜在类模型报表

每个“潜在类模型报表”被动态地命名为“<k> 个聚类的潜在类模型”，具体取决于 *k*，即拟合的聚类数。报表包含以下结果和分级显示项：

- “模型汇总”
- “参数估计值”
- “转置参数估计值”
- “效应大小”
- “MDS 图”
- “混合概率”

模型汇总

默认情况下，指定聚类数的模型汇总显示在每个“潜在类模型”报表的顶部。模型汇总包含 -对数似然、参数个数、BIC 和 AIC。这些汇总值可用于确定模型对数据拟合的好坏程度。“-对数似然”、“BIC”和“AIC”的值越小表示拟合效果越好。请参见《拟合线性模型》。“参数数目”值提供潜在类模型中的唯一参数个数。请参见“‘潜在类分析’平台的统计详细信息”。

参数估计值

“参数估计值”报表包含表格式和图形化的参数估计值汇总，并且在默认情况下显示出来。每个汇总包含的行对应于模型聚类。

“总体”列显示观测属于每个聚类的概率。（这些是 γ 参数。请参见“‘潜在类分析’平台的统计详细信息”。）

显示中的其余列依据在 “潜在类分析”启动窗口中指定的 Y 列使用竖分隔线分组。

- 每一组分类响应列都有一列对应于相应响应中的每个水平。在每组中，给定行列中的值是假定观测属于行所标识聚类的前提下，列所指示的响应的条件概率。（这些是 ρ 参数。）
- 每一组多重响应列都有一列对应于多重响应中的每个类别。在每组中，给定行列中的值是假定观测属于行所标识聚类的前提下，所指示的类别的较低级别下的响应的条件概率。（这些是 ρ 参数。）

图形化显示将条件概率值显示为**份额图**。对于每个聚类和每个 Y，用水平堆叠条形图绘制了给定聚类成员关系的条件概率。对于二值或名义型响应列，这些图中每个响应的百分比都加总为 1。对于多重响应列，百分比是每个类别的较低级别的百分比，加总不为 1。直条的堆叠按照值表中变量的顺序展示。您还可以悬停在直条上方，以查看变量的水平或类别。

提示：您可以在“参数估计值”报表的任一表中选择一行或多行，以选择分配至相应聚类的观测。

转置参数估计值

“转置参数估计值”报表包含“参数估计值”报表表的转置表。该报表中聚类显示为列。针对分析中每个 Y 列的每个响应类别显示了其在每个聚类出现的条件概率。

注意：“总体”列的估计值不包括在转置表中。

效应大小

“效应大小”表在各聚类之间比较 Y 列，且默认情况下显示出来。该表每行中的统计量从由 Y 列的水平或类别与聚类成员关系的期望计数构成的列联表分析中获得。期望计数是将每个聚类中的观测数乘以 Y 列每个水平或类别的条件概率得到。

对于每个响应，为由聚类与水平的期望计数构成的列联表计算 Pearson 卡方统计量 χ^2。用 n 表示观测数。“效应大小”列中的值定义如下：

$$\text{效应大小} = \sqrt{\frac{\chi^2}{n}}$$

“LR Logworth”列中的每个值显示 -log10(p_{LR})，其中 p_{LR} 是由期望计数构成的列联表的对数似然比检验 p 值。超过 2 的 Logworth 值对应于在 0.01 显著性水平下的显著性。

提示：您可以在“效应大小”表中选择一行或多行，以选择数据表中的相应列。

MDS 图

MDS 图对每个聚类都包含一个点，并且在默认情况下显示出来。它是聚类邻近关系的二维表示。越邻近的聚类越相似。该图基于 ρ 参数的相异度矩阵创建。有关 MDS 图的详细信息，请参见“多维尺度化”。

混合概率

“混合概率”表显示每行的聚类成员关系的概率。“最可能的聚类”列指示每行具有最高成员关系概率的聚类。

注意：一或多个 Y 列包含缺失值的行会从分析中排除，不会出现在“混合概率”表中。

“潜在类分析”平台选项

“潜在类分析”红色小三角菜单包含以下选项：

新聚类数 使您可以使用其他聚类数运行另一个分析。新分析报表会追加至当前报表。

请参见《使用 JMP》获取有关下列选项的信息：

本地数据过滤器 显示或隐藏支持您过滤特定报表中使用的数据的本地数据过滤器。

重新运行 包含使您可以重复或重新启动分析的选项。在支持该功能的平台中，“自动重新计算”选项立即在相应报表窗口中反映您对数据表所做的更改。

平台首选项 包含的选项支持您查看当前平台首选项或更新平台首选项以匹配当前 JMP 报表中的设置。

保存脚本 包含的选项支持您保存可将报表重现到若干目标的脚本。

保存 “依据”组脚本 包含使您可以保存脚本的选项，可将为“依据”变量的所有水平重新生成平台报表的脚本保存到多个不同的位置。仅当在启动窗口中指定“依据”变量时才可用。

“潜在类模型”选项

“‘<k>’个聚类的潜在类模型”红色小三角菜单包含以下选项：

模型报表 支持您显示或隐藏可用模型报表。有关模型报表的详细信息，请参见 “ “潜在类分析”报表”。

按聚类设定颜色 根据最可能的聚类对数据表中的每一行设定颜色。有关示例，请参见 “ “潜在类分析”平台的更多示例”。

保存混合和聚类公式 将每个聚类的公式列以及最可能的聚类的公式列保存至数据表。

仅保存聚类公式 将确定最可能的聚类的公式列保存至数据表。

JMP PRO **发布概率公式** 创建概率公式并在“公式存储库”平台中将它们保存为公式列脚本。若未打开“公式存储库”报表，该选项将创建“公式存储库”报表。请参见《预测和专业建模》。

保存混合概率 将“混合概率”表中的值保存至数据表中的相应行。

仅保存聚类 将包含每行最可能的聚类的新列保存至数据表。该列不包含公式。

重命名聚类 可让您为报表中的聚类提供有意义的名称。

注意：除非为报表指定了随机种子，否则新聚类名称不会保存到脚本中。仅当通过脚本启动报表时才可设置随机种子。

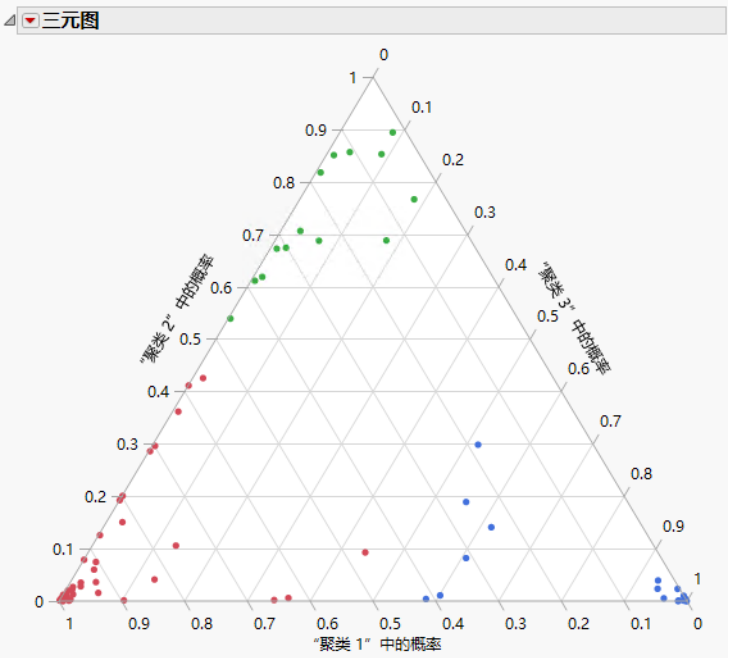
删除拟合 从报表窗口中删除指定的聚类报表。

“潜在类分析”平台的更多示例

在本例中，您将生成一个图来直观演示聚类成员的概率。您有关于车主和汽车制造商的调查数据，并且您打算将车主划分为三个聚类。当您有三个聚类时，三元图可提供良好的直观展现。

1. 选择**帮助** > **样本数据文件夹**，然后打开 Car Poll.jmp。
2. 选择**分析** > **聚类** > **潜在类分析**。
3. 选择除**年龄**外的所有列并点击 **Y**。
4. 点击**确定**。
5. 点击“‘3’个聚类的潜在类模型”旁边的红色小三角，选择**按聚类设定颜色**：
6. 点击“‘3’个聚类的潜在类模型”旁边的红色小三角，选择**保存混合概率**：
7. 在“Car Poll”数据表窗口中，从列的列表中选择 **LCA 聚类概率**列组。
8. 选择**图形** > **三元图**。
9. 点击 **X, 绘图**。
10. 点击**确定**。

图 16.6 聚类成员关系概率的三元图



在每个观测的聚类概率的三元图中，大多数聚类成员关系概率落到顶点附近。这表明一个聚类具有较高的值，另外两个聚类具有较低的值。不过，图中部的某些点表明这些观测对任何聚类的聚类成员关系都不具有高概率。这些观测可能需要更仔细的检查，或者表明需要更多的聚类以便更好地表示数据。

注意：因为没有指定随机种子，您的结果可能有所不同。

“潜在类分析”平台的统计详细信息

本节包含 “潜在类分析”的统计详细信息。

- “潜在类模型拟合的统计详细信息”
- “最大聚类数的统计详细信息”

潜在类模型拟合的统计详细信息

本节说明在“潜在类分析”平台中拟合的潜在类模型。有关潜在类模型的详细信息，请参见 Collins and Lanza (2010) 和 Agresti (2013)。

注意：在“文本分析器”平台中使用的 LCA 算法利用文档词条矩阵的稀疏性。因为这个原因，“文本分析器”平台中的 LCA 结果与“潜在类分析”平台中的结果不完全一致。

用 $j = 1, ..., J$ 表示响应的观测列。它们是“潜在类分析”平台启动窗口中的 Y 列。列 j 的水平数表示为 R_j。

J 个变量的多维列联表包含 $W = R_1*...*R_J$ 个单元格。其中的每一个单元格依据其针对 J 个变量的响应模式来定义。因此，每个响应模式是形式为 $\mathbf{y} = y_1, ..., y_j$ 的 J 长度向量。将 $\mathbf{Y}$ 定义为所有响应模式视为行向量的 J x W 数组。 $\mathbf{Y}$ 中的每个元素 $\mathbf{y}_w$ 都具有概率 $\Pr(\mathbf{y}_w)$。这些概率之和为 1：

$$\sum_{w=1}^{W} Pr(y_w) = 1$$

考虑以下符号：

- C 是潜在类模型中的聚类数。
- γ_c 是聚类 c 中成员关系的概率。（γ_c 是潜在类流行度。）这些参数之和为 1。
- $r_{j,k}$ 是第 j 个响应的第 k 个水平。
- $\rho_{j,k|c}$ 是在属于类 c 的条件下，在列 j 中观测到响应 $r_{j,k}$ 的概率。（$\rho_{j,k|c}$ 是项目响应概率。）对于给定的聚类和响应变量 j，$\rho_{j,k|c}$ 之和为 1。
- $I(y_j = r_{j,k})$ 是指标函数，当 y_j 响应为第 j 个响应的第 k 个水平时该函数等于 1，其他情况下该函数等于 0。

观测到响应 $\mathbf{y}_w = y_1, ..., y_j$ 的特定向量的概率是在 C 个潜在类下观测到该响应向量的条件概率之和：

$$\Pr(y) = \sum_{c=1}^{C} \gamma_c \prod_{j=1}^{J} \prod_{k=1}^{R_j} \rho_{j,k|c}^{I(y_j = r_{j,k})}$$

该方程是您从“潜在类分析”红色小三角菜单中选择“保存混合和聚类公式”选项时保存至数据表的 Prob Formula Cluster 公式的分母。Prob Formula Cluster 列中的公式给出 $\Pr(\text{聚类} = c \mid \mathbf{y}_w)$，其等于 $\Pr(\mathbf{y}_w, \text{聚类} = c) / \Pr(\mathbf{y}_w)$。

潜在类模型的 γ 和 ρ 参数使用迭代期望值最大化 (EM) 算法估计得到。潜在类模型中的唯一参数个数定义如下：

$$(C-1) + C \sum_{j=1}^{J} (R_j - 1)$$

最大聚类数的统计详细信息

可以在潜在类分析模型中拟合的最大聚类数取决于模型自由度。潜在类分析模型中的自由度基于列创建的列联表的大小。列联表的大小是表中包含至少一个观测的单元数，表示为 *K*。若所有单元都包含至少一个观测，则 *K* 是响应列的水平数的乘积。自由度公式定义如下：

自由度 = *K* - {nCluster - 1 + nCluster(nTotalLevels - nCols)} - 1

其中

nCluster = 聚类数

nTotalLevels = 响应列的水平总和

nCols = 响应列数

为使 LCA 模型充分拟合，自由度必须为正。因此，要确保自由度 > 0，聚类的最大数目定义如下：

max(nCluster) < floor[*K*/(1 + nTotal Levels – nCols)]

第 17 章

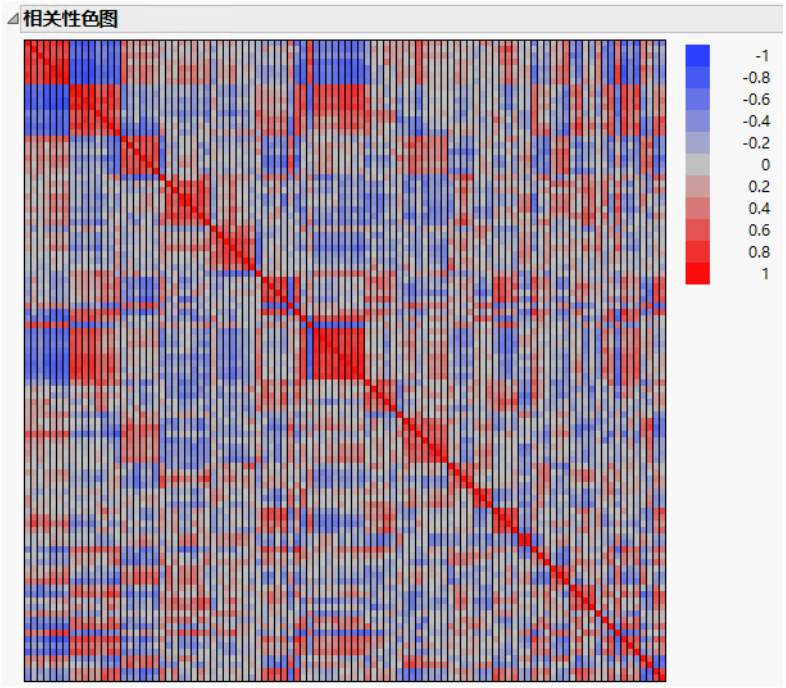
聚类变量

将类似变量分组到代表组中

变量聚类提供将类似变量分组到代表组中的方法。每个聚类可由单个成分或变量表示。成分是聚类中所有变量的线性组合。或者，聚类也可由标识为聚类中最典型成员的变量来表示。

您可以将“聚类变量”用作降维方法。您不用在建模中使用大量变量，聚类成分或聚类中最典型的变量便可解释数据中的大部分变异。此外，使用“聚类变量”的降维方法经常比使用主成分的降维方法更容易解释。

图 17.1 变量相关性色图的示例



目录

“聚类变量”平台概述

“聚类变量”平台构造的成分是同一类别的相似变量的线性组合。这不同于主成分分析，主成分分析所构造的成分是分析中的所有变量的线性组合。在“聚类变量”平台中，整个变量集划分为各个聚类。对于每个聚类，使用该聚类中的变量的第一主成分构造**聚类成分**。该聚类成分是一个线性组合，用于解释该聚类的变量中的尽可能多的变异。

您可以将“聚类变量”选项用作降维方法。大型变量集中的绝大部分变异往往可以通过聚类成分或聚类中最典型的变量来表示。随后，这些新变量可用在预测或其他建模方法中。基于聚类的新变量通常比基于所有变量的主成分更容易解释。

由一组相同变量构造的主成分相互之间正交。不过，聚类成分不正交，因为它们是由不同的变量集构造的。

当具有大量变量时，“聚类变量”平台使用基于奇异值分解的算法来缩短计算时间。有关更多背景信息，请参见“‘宽线性’方法和奇异值分解”。

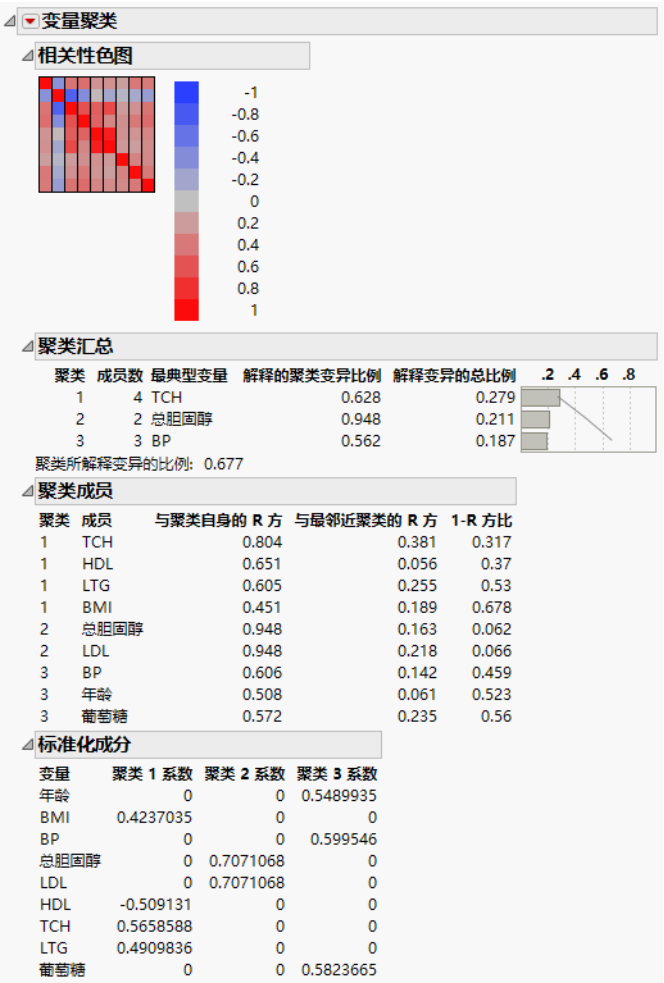
“聚类变量”平台的示例

在本例中，您对糖尿病疾病进展建模中使用的连续型基线变量聚类。

1. 选择**帮助** > **样本数据文件夹**，然后打开 Diabetes.jmp。
2. 选择**分析** > **聚类** > **聚类变量**。
3. 从**年龄**列一直选到**葡萄糖**列，但不包括**性别**（**年龄**、 BMI、 BP、**总 胆固醇**、 LDL、 HDL、 TCH、 LTG 和**葡萄糖**），然后点击 **Y，列**。

   不能包括**性别**列，因为“聚类变量”要求提供数值型连续变量。
4. 点击**确定**。

图 17.2 糖尿病数据的聚类变量报表



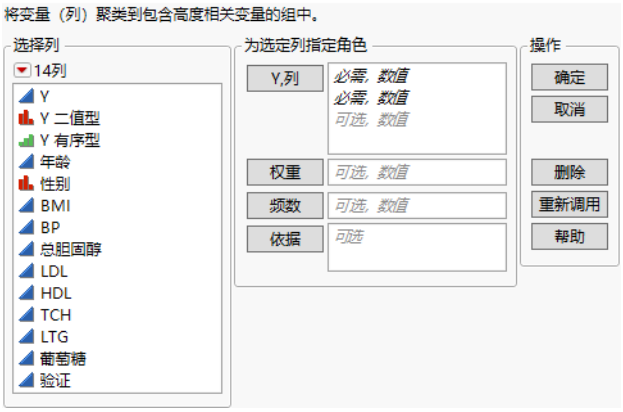
“聚类汇总”报表显示变量被分组到三个聚类中：

- “聚类 1”包括 TCH、HDL、LTG 和 BMI，如“聚类成员”报表中所示。“聚类汇总”报表显示 TCH 是对于“聚类 1”最典型的变量，而且对于“聚类 1”中的变量，判别主成分 2.8% 的变异是由第一个主成分解释的。
- “聚类 2”包括**总胆固醇**和 LDL。“聚类汇总”报表显示**总胆固醇**是对于“聚类 2”最典型的变量，而且对于“聚类 2”中的变量，94.8% 的变异是由第一个主成分解释的。
- 聚类 3 包括 BP、**年龄**和**葡萄糖**。“聚类汇总”报表显示最典型变量是 BP，对于“聚类 3”中的变量，5判别主成分.2% 的变异是由第一个主成分解释的。

启动 “聚类变量” 平台

通过选择**分析** > **聚类** > **聚类变量**来启动 “聚类变量” 平台。

图 17.3 “聚类变量” 启动对话框



有关 “选择列” 红色小三角菜单中选项的详细信息，请参见《使用 JMP》。

Y，列 要聚类的变量。变量必须是数值型和连续型。

权重 一列，该列的数值为分析中的每一行都分配一个权重。

频数 一列，列中的数值为分析中的每行分配一个频数。

依据 一列，其水平定义不同的分析。对于指定列的每个水平，都分析相应行。结果显示在不同的报表中。若分配了多个 “依据” 变量，则为 “依据” 变量水平的每个可能组合生成单独分析。

“变量聚类” 报表

默认情况下，“聚类变量” 平台中的 “变量聚类” 报表包含以下部分：

- “相关性色图”
- “聚类汇总”
- “聚类成员”
- “标准化成分”

提示：当您在任何 “变量聚类” 表中选择行时，数据表中相应的列也会被选定。按 Ctrl 键并点击行可取消选择数据表中的对应列。

相关性色图

在“变量聚类”报表中，“相关性色图”报表显示变量之间的相关性的色图。变量按它们在“聚类成员”表中列出的顺序排列。这种排列方式可确保同一聚类的成员在相关性图中是相邻的。请参见“相关性色图的示例”。

提示：悬停在色图的一个方块上时，可看到该方块中涉及的变量及其相关性。

同一聚类中的变量倾向于比不同聚类中的变量具有更高的绝对相关性（深红色或深蓝色）。因此，在相关性色图中，与给定的主成分中的变量相对应的方格形成的方块经常会沿对角线突出显示。

相关性使用逐行方法计算。该方法会从相关性计算中排除变量具有缺失数据的观测。有关逐行估计方法的详细信息，请参见“方差估计方法的统计详细信息”。

聚类汇总

在“变量聚类”报表中，“聚类汇总”表包含以下列：

聚类 聚类标识符。

成员数 聚类中的变量数。

最典型变量 与其聚类成分的平方相关性最大的那个聚类变量。

解释的聚类方差比例 聚类的变量中第一主成分解释的聚类方差比例。若聚类中仅有一个变量，则该值是 1。该统计量仅基于聚类中的变量而不是全部变量。

解释变异的总比例 聚类成分解释的方差总比例。这等价于仅使用每个聚类中的变量计算第一主成分。

表下方的注释给出所有聚类成分解释的变异的总比例。

聚类成员

在“变量聚类”报表中，“聚类成员”表包含以下列：

聚类 聚类标识符。

成员 聚类中包括的变量。

与聚类自身的 R 方 变量与其聚类成分的平方相关性。

与最邻近聚类的 R 方 变量与其最邻近聚类的聚类成分的平方相关性。最邻近聚类是变量与聚类成分的平方相关性次高的聚类。

1-R 方比 衡量某个变量所属的聚类与其最邻近聚类之间的相对接近性的测度。其定义如下：

(1 - 与聚类自身的 R 方)/(1 - 与最邻近聚类的 R 方)

标准化成分

在“变量聚类”报表中，“标准化成分”表列出定义聚类成分的系数。这些系数是每个聚类中的第一主成分的特征向量。

“聚类变量”平台选项

“变量聚类”红色小三角菜单包含以下选项：

相关性色图 显示或隐藏“相关性色图”。请参见“相关性色图”。

聚类汇总 显示或隐藏“聚类汇总”报表。请参见“聚类汇总”。

聚类成员 显示或隐藏“聚类成员”报表。请参见“聚类成员”。

聚类成分 显示或隐藏“标准化成分”报表。请参见“标准化成分”。

保存聚类成分 将列作为组（称为**聚类成分**）保存到数据表中。每列名为**聚类** <i> **成分**并且包含公式，它用未中心化和未统一尺度的变量来表示聚类成分。

启动拟合模型 打开“模型规格”窗口，已经在该窗口中的“构造模型效应”列表中为每个聚类输入“最典型变量”。使用该选项可基于“最典型变量”构造模型。

提示：要使用成分拟合模型，首先选择**保存聚类成分**选项。然后用所需的“聚类成分”列替换“构造模型效应”列表中每个聚类的“最典型变量”。

请参见《使用 JMP》获取有关下列选项的信息：

本地数据过滤器 显示或隐藏支持您过滤特定报表中使用的数据的本地数据过滤器。

重新运行 包含使您可以重复或重新启动分析的选项。在支持该功能的平台中，“自动重新计算”选项立即在相应报表窗口中反映您对数据表所做的更改。

平台首选项 包含的选项支持您查看当前平台首选项或更新平台首选项以匹配当前 JMP 报表中的设置。

保存脚本 包含的选项支持您保存可将报表重现到若干目标的脚本。

保存“依据”组脚本 包含使您可以保存脚本的选项，可将为“依据”变量的所有水平重新生成平台报表的脚本保存到多个不同的位置。仅当在启动窗口中指定“依据”变量时才可用。

“聚类变量”平台的更多示例

本节包含使用“聚类变量”平台的示例。

- “相关性色图的示例”
- ““聚类变量”平台有关降维的示例”

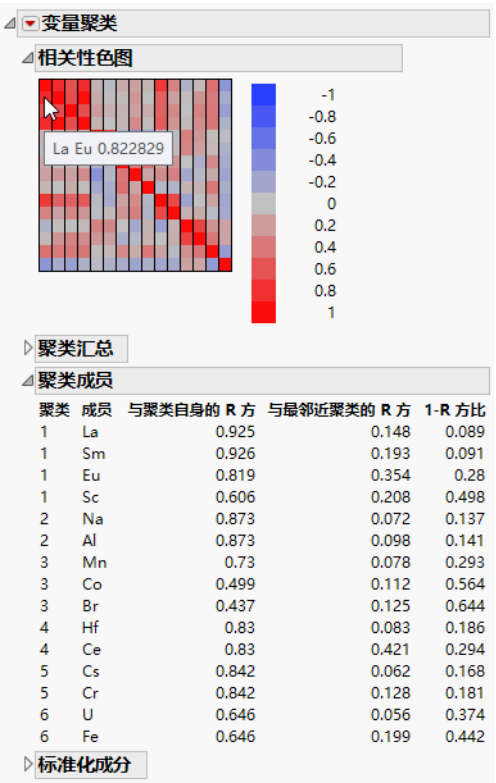
相关性色图的示例

在本例中，您使用 “聚类变量”平台构造并检查 “相关性色图”。色图对于查找变量之间的相关性模式非常有用。

1. 选择**帮助 > 样本数据文件夹**，然后打开 Cherts.jmp。
2. 选择**分析 > 聚类 > 聚类变量**。
3. 选择所有连续变量并点击 **Y, 列**。
4. 点击**确定**。
5. 关闭 “聚类汇总”和 “标准化成分”报表。
6. 悬停在色图的第二行第一列的单元格上方。

   随即出现工具提示，显示对应于该单元格的变量是 La 和 Eu，并且它们的相关性为 0.822829。

图 17.4 Cherts.jmp 的 “相关性色图”



“聚类成员”报表显示“聚类 1”中有四个变量。在“相关性色图”中，左上角中对应于这四个变量的由四乘四个方块构成的方格显示出明显的正相关性模式。色图还显示聚类 2、4 和 5 中变量的正相关性模式。色图右下角中对应于两个“聚类 判别主成分”变量的由二乘二个方块构成的方格显示它们是负相关的。请参见“相关性色图”。

“聚类变量”平台有关降维的示例

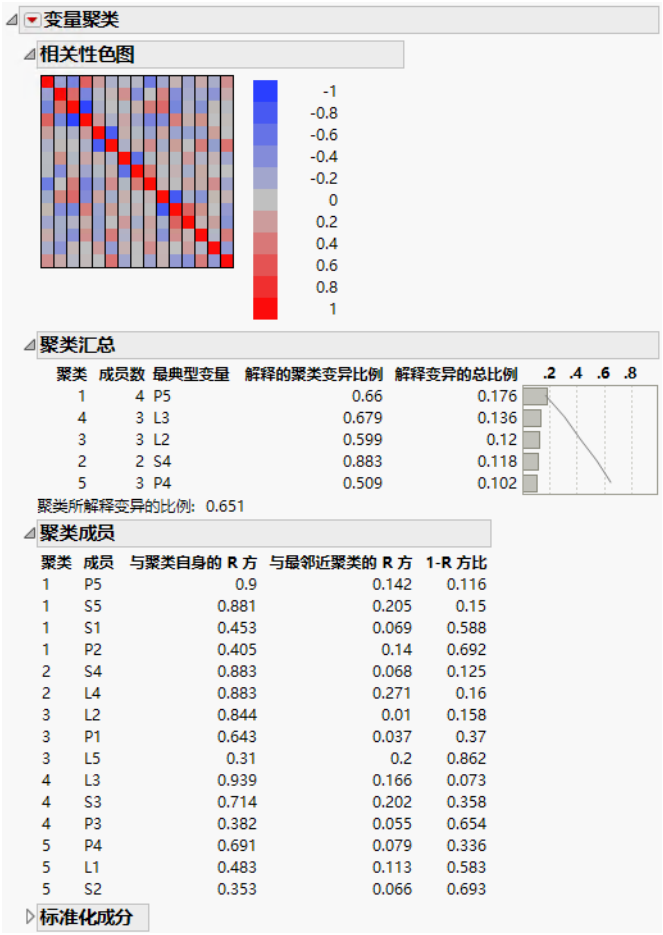
在本例中，您将“聚类变量”平台用作进行建模的降维工具。数据表中包含 15 个用于预测响应变量的变量，而您希望减少该数字。

聚类变量

1. 选择**帮助** > **样本数据文件夹**，然后打开 Penta.jmp。
2. 选择**分析** > **聚类** > **聚类变量**。
3. 选择所有连续变量 （但不包括**对数 RAI**）并点击 **Y, 列**。
4. 点击**确定**。
5. 点击 “变量聚类”红色小三角并选择**保存聚类成分**。

   五个分组公式列将添加至数据表。

图 17.5 Penta.jmp 的 “聚类变量”报表



“聚类汇总”和“聚类成员”报表显示变量分为五个组，因此有五个聚类成分变量。

拟合模型

接下来，拟合并比较两个模型以预测**对数 RAI**：

- 使用所有连续变量作为预测变量的模型。
- 使用聚类成分作为预测变量的模型。

1. 点击“变量聚类”红色小三角并选择**启动拟合模型**。
2. 选择**对数 RAI** 并点击 **Y**。

   注意到五个聚类的 “最典型变量”已输入到 “构造模型效应”列表中。但是，您想输入所有预测变量。
3. 从 S1 一直选到 P5 ，选择所有连续变量并点击**添加**。

一定不要包括**观测名称**。

4. 选择**保持对话框打开**旁边的框。
5. 点击**运行**。

图 17.6 包含所有连续预测变量的模型的 “拟合最小二乘法”报表

响应 “对数 RAI”

效应汇总

拟合汇总

R 方	0.929316
调整 R 方	0.853582
均方根误差	0.331225
响应均值	0.734333
观测数（或权重和）	30

方差分析

源	自由度	平方和	均方	F 比
模型	15	20.193596	1.34624	12.2709
误差	14	1.535941	0.10971	概率>F
校正总和	29	21.729537		<.0001*

参数估计值

项	估计值	标准误差	t 比	概率>|t|
截距	-0.802632	0.924946	-0.87	0.4002
P5	2.0563803	1.651272	1.25	0.2335
S4	-0.062354	0.134935	-0.46	0.6511
L2	0.0860287	0.061206	1.41	0.1817
L3	0.3185383	0.080091	3.98	0.0014*
P4	0.4136598	0.394449	1.05	0.3121
S1	-0.09783	0.038948	-2.51	0.0249*
L1	0.032362	0.049732	0.65	0.5258
P1	-0.107951	0.085209	-1.27	0.2259
S2	0.086703	0.044276	1.96	0.0704
P2	0.0847235	0.086297	0.98	0.3429
S3	-0.037728	0.055602	-0.68	0.5085
P3	-0.027313	0.233655	-0.12	0.9086
L4	-0.029756	0.152012	-0.20	0.8476
S5	2.7123146	2.222039	1.22	0.2424
L5	-0.209128	0.270401	-0.77	0.4521

效应检验

效应详细信息

6. 在 “拟合模型”窗口中，选择 “构造模型效应”窗口中的所有变量，然后点击**删除**。
7. 选择**聚类成分**组并点击**添加**。
8. 点击**运行**。

图 17.7 聚类成分作为预测变量的模型的 “拟合最小二乘法”报表

响应 “对数 RAI”

效应汇总

拟合汇总

R 方	0.8214
调整 R 方	0.784191
均方根误差	0.402125
响应均值	0.734333
观测数（或权重和）	30

方差分析

源	自由度	平方和	均方	F 比
模型	5	17.848635	3.56973	22.0757
误差	24	3.880902	0.16170	概率>F
校正总和	29	21.729537		<.0001*

参数估计值

项	估计值	标准误差	t 比	概率>|t|
截距	0.6651552	0.074349	8.95	<.0001*
聚类 1 成分	-0.018483	0.056013	-0.33	0.7443
聚类 2 成分	0.0035891	0.069032	0.05	0.9590
聚类 3 成分	0.2043072	0.066714	3.06	0.0053*
聚类 4 成分	0.5754553	0.065423	8.80	<.0001*
聚类 5 成分	-0.046594	0.069786	-0.67	0.5107

效应检验

效应详细信息

仅包括五个聚类成分作为预测变量的模型解释响应中的绝大部分变异，其调整 R 方为 0.784。使用全部十五个预测变量的模型仅有略高的调整 R 方，它的值为 0.853（图 17.6）。

“聚类变量”平台的统计详细信息

变量聚类算法以迭代方式拆分变量的原始类别并将变量重新分配到新的类别，直到不可能再进一步拆分。初始聚类包含所有变量。该算法由 SAS 开发并且在 PROC VARCLUS 中实现 (SAS Institute Inc. 2020g)。

注意：该算法仅使用“Y,列”列表中的变量没有缺失值的观测。

以下是算法中的迭代步骤：

1. 对于所有聚类，请执行以下操作：
   a. 计算每个聚类中的变量的主成分。
   b. 若所有聚类的第二特征值均小于 1，则终止算法。
2. 使用以下步骤把第二特征值最大 （且大于 1）的聚类分成两个聚类：
   a. 使用斜交旋转来旋转当前聚类中的变量的主成分。
   b. 定义一个聚类，使其包含当前聚类中的变量满足：该变量与第一旋转主成分的平方相关性高于该变量与第二主成分的平方相关性。

c. 定义另一个聚类，使其包含原始聚类中的其余变量。这些变量与第二主成分具有更高的相关性。

d. 计算两个新聚类的主成分。

3. 通过检验来判定数据集中的任何变量是否应分配给不同的聚类。对于每个变量，请执行以下操作：

a. 计算变量与每个聚类的第一主成分的平方相关性。

b. 将变量放置在与其的平方相关性最高的聚类中。

注意：斜交旋转亦称原始四次方最大正交旋转。请参见 Harris and Kaiser (19判别主成分4)。

附录 A

统计详细信息

多元方法

本附录将讨论“宽线性”方法和奇异值分解的使用。

目录

“宽线性”方法和奇异值分解

使用“聚类”、“主成分”和“判别”平台中的“宽线性”方法，您可以分析包含数千（甚至数百万）个变量的数据集。大多数多元方法要求计算协方差矩阵或对协方差矩阵求逆。当您的多元分析涉及很多变量时，协方差矩阵会非常大以至于对它进行计算或求逆很困难或计算时间很长。

假定您的数据包含 n 行和 p 列。协方差矩阵的秩最多是 n 和 p 中较小的那个。在宽数据集中，p 通常比 n 大很多。在这种情况下，协方差矩阵的逆矩阵最多有 n 个非零特征值。“宽线性”方法使用这一事实情况以及奇异值分解来提供高效的计算。请参见“计算奇异值分解”。

奇异值分解

奇异值分解 (SVD) 允许您先后通过一次旋转、一次缩放以及另一次旋转的方式来表示任意线性变换。 SVD 规定任何 $n \times p$ 的矩阵 **X** 可表示为：

$$\mathbf{X} = \mathbf{U}Diag(\Lambda)\mathbf{V}'$$

用 r 表示 **X** 的秩。用 $\mathbf{I}_r$ 表示 $r \times r$ 单位矩阵。

矩阵 **U**, $Diag(\Lambda)$ 和 **V** 具有以下属性：

U 是 $n \times r$ 半正交矩阵且 $\mathbf{U}'\mathbf{U} = \mathbf{I}_r$

V 是 $p \times r$ 半正交矩阵且 $\mathbf{V}'\mathbf{V} = \mathbf{I}_r$

$Diag(\Lambda)$ 是 $r \times r$ 对角矩阵，其正对角线元素由列向量 $\Lambda = (\lambda_1, \lambda_2, ..., \lambda_r)'$ 给出，其中 $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r > 0$。

λ_i 是 **X** 的非零**奇异值**。

下面说明了 SVD 与正方形矩阵的谱分解的关系：

- λ_i 的平方是 $\mathbf{X}'\mathbf{X}$ 的非零特征值。
- **V** 的 r 列是 $\mathbf{X}'\mathbf{X}$ 的特征向量。

注意：在文献中有关于矩阵 **U**、**V** 以及包含奇异值的矩阵的维度的各种约定。但是，这些约定差异在 **X** 秩的范围内对分解没有实际影响。

有关奇异值分解的详细信息，请参见 Press et al.(1998, Section 2.判别主成分)。

协方差矩阵

本节说明如何使用奇异值分解 (SVD) 获取协方差矩阵的特征向量和特征值。关注的矩阵具有至少一个大的维度时，计算 SVD 比计算它的协方差矩阵和特征值分解高效得多。

用 n 表示观测数，用 p 表示关注的多元分析中涉及的变量数。用 **X** 表示数据值的 $n \times p$ 矩阵。

通常对标准化数据应用 SVD。要将某个值标准化，需减去其均值，再除以其标准差。用 $\mathbf{X}_s$ 来表示标准化数据值的 $n \times p$ 矩阵。之后，标准化数据的协方差矩阵成为 **X** 的相关性矩阵，该矩阵定义如下：

$$Cov = \mathbf{X}_s'\mathbf{X}_s/(n-1)$$

可以对 $\mathbf{X}_s$ 应用 SVD 来获取 $\mathbf{X}_s'\mathbf{X}_s$ 的特征向量和特征值。当矩阵 **X** 很宽（很多列）或很高（很多行）时，这样做可以高效计算特征向量和特征值。此方法是宽 PCA 的基础。请参见““主成分”报表”。

逆协方差矩阵

一些多元方法需要计算协方差矩阵的逆矩阵。本节说明如何使用 SVD 来计算协方差矩阵的逆矩阵。

用 $\mathbf{X}_s$ 表示标准化数据矩阵并定义 $\mathbf{S} = \mathbf{X}_s'\mathbf{X}_s$。奇异值分解允许您将 **S** 表示为：

$$\mathbf{S} = (\mathbf{U}Diag(\Lambda)\mathbf{V}')'(\mathbf{U}Diag(\Lambda)\mathbf{V}') = \mathbf{V}Diag(\Lambda)^2\mathbf{V}'$$

若 **S** 是满秩的，则 **V** 是 $p \times p$ 正交矩阵，您可以将 $\mathbf{S}^{-1}$ 表示为：

$$\mathbf{S}^{-1} = (\mathbf{V}Diag(\Lambda)^2\mathbf{V}')^{-1} = \mathbf{V}Diag(\Lambda)^{-2}\mathbf{V}'$$

若 **S** 不是满秩的，则 $\text{Diag}(\Lambda)^{-1}$ 可以使用广义逆矩阵 $\text{Diag}(\Lambda)^{+}$ 代替，其中 $Diag(\Lambda)$ 的对角线元素用其倒数替代。这便按以下方式定义 **S** 的广义逆矩阵：

$$\mathbf{S}^{-} = \mathbf{V}(Diag(\Lambda)^{+})^2\mathbf{V}'$$

该广义逆矩阵只使用 SVD 就可以计算得到。

有关将 SVD 应用于宽线性判别分析的详细信息，请参见“宽线性判别方法”。

计算奇异值分解

在“多元方法”平台中，JMP 采用 Golub and Kahan (19判别主成分5) 提出的方法计算矩阵的 SVD。Golub and Kahan 的方法是一个包含两个步骤的过程。第一步是将矩阵 **M** 简化为二对角矩阵 **J**。第二步是计算 **J** 的奇异值，它们与原始矩阵 **M** 的奇异值相同。通常将矩阵 **M** 的列标准化以便平衡变量对计算的影响。Golub and Kahan 方法的计算效率很高。

附录 B

参考书目

《多元方法》中引用了以下来源。

Agresti, A. (2013).*Categorical Data Analysis*.3rd ed. Hoboken, NJ: John Wiley & Sons.

Baglama, J., and Reichel, L. (2005).“Augmented implicitly restarted Lanczos bidiagonalization methods.”*SIAM Journal on Scientific Computing* 27:19–42.

Ballard, D. H. (1981).“Generalizing the Hough Transform to Detect Arbitrary Shapes.”*Pattern Recognition* 13:111–122.

Bartlett, M. S. (1937).“Properties of sufficiency and statistical tests.”*Proceedings of the Royal Society of London, Series A* 160:268–282.

Bartlett, M. S. (1954).“A Note on the Multiplying Factors for Various Chi Square Approximations.”*Journal of the Royal Statistical Society, Series B* 16:296–298.

Benzécri, J. P. (1979).“Sur le calcul des taux d’inertie dans l’analyse d’un questionnaire, addendum et erratum à [BIN.MULT.].”*Cahiers de l’Analyse des Données* 4:377–378.

Bentler, P. M. (1990).“Comparative Fit Indexes in Structural Models.”*Psychological Bulletin* 107:238.

Bentler, P. M., and Freeman, E. H. (1983).“Tests for Stability in Linear Structural Equation Systems.”*Psychometrika* 48:143–145.

Bernhardsson, E. (2013).“Approximate Nearest Neighbors in C++/Python optimized for memory usage and loading/saving to disk.” https://github.com/spotify/annoy

Bollen, K. A. (1989).*Structural Equations with Latent Variables*.New York: John Wiley & Sons.

Bollen, K. A., Harden, J. J., Ray, S., and Zavisca, J. (2014).“BIC and Alternative Bayesian Information Criteria in the Selection of Structural Equation Models.”*Structural Equation Modeling* 21:1–19.

Borg, I., and Groenen, P. J. F. (2005).*Modern Multidimensional Scaling: Theory and Applications*.2nd ed. New York: Springer.

Boulesteix, A.-L., and Strimmer, K. (2007).“Partial Least Squares: A Versatile Tool for the Analysis of High-Dimensional Genomic Data.”*Briefings in Bioinformatics* 8:32–44.

Browne, M. (2001).“An Overview of Analytic Rotation in Exploratory Factor Analysis.”*Multivariate Behavioral Research* 36:111–150.

Browne, M. W., and Cudeck, R. (1993).“Alternative Ways of Assessing Model Fit.”In *Testing Structural Equation Models*, edited by K. A. Bollen, and J. S. Long, 136–162.Newbury Park, CA: Sage Publications.

Chen, F. F. (2007).“Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance.”*Structural Equation Modeling* 14:464–504.

Collins, L., and Lanza, S. (2010).*Latent Class and Latent Transition Analysis*.Hoboken NJ: John Wiley & Sons.

Cox, I., and Gaudard, M. (2013).*Discovering Partial Least Squares with JMP*.Cary, NC: SAS Institute Inc.

Cronbach, L. J. (1951).“Coefficient Alpha and the Internal Structure of Tests.”*Psychometrika* 16:297–334.

Cudeck, R., and MacCallum, R. C., eds.(2007).Factor Analysis at 100, Historical Developments and Future Directions.Mahwah, NJ: Lawrence Erlbaum Associates.

de Ayala, R. J. (2009).*The Theory and Practice of Item Response Theory*.New York: Guilford Press.

De Jong, S. (1993).“SIMPLS: An Alternative Approach to Partial Least Squares Regression.”*Chemometrics and Intelligent Laboratory Systems* 18:251–263.

Denham, M. C. (1997).“Prediction Intervals in Partial Least Squares.”*Journal of Chemometrics* 11:39–52.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikstrom, C., and Wold, S. (2006).*Multi- and Megavariate Data Analysis Basic Principles and Applications (Part I)*.Chapter 4.Umetrics.

Finkbeiner, C. (1979).“Estimation for the Multiple Factor Model when Data are Missing.”*Psychometrika* 44:409–420.

Fisher, L., and Van Ness, J. W. (1971).“Admissible Clustering Procedures.”*Biometrika* 58:91–104.

Florek, K., Lukaszewicz, J., Perkal, J., and Zubrzycki, S. (1951a).“Sur la liaison et la division des points d’un ensemble fini.”*Colloquium Mathematicae* 2:282–285.

Florek, K., Lukaszewicz, J., Perkal, J., and Zubrzycki, S. (1951b).“Taksonomia Wroclawska.”*Przeglad Antropologiczny* 17:193–211.

Frank, I. E., and Todeschini, T. (1994).*The Data Analysis Handbook*.New York: Elsevier.

Friedman, J. H. (1989).“Regularized Discriminant Analysis.”*Journal of the American Statistical Association* 84:165–175.

Garthwaite, P. (1994).“An Interpretation of Partial Least Squares.”*Journal of the American Statistical Association* 89:122–127.

Golub, G. H., and Kahan, W. (1965).“Calculating the singular values and pseudo-inverse of a matrix.”*Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis* 2:205–224.

Goodman, L. A. (1974).“Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models.”*Biometrika* 61:215–231.

Greenacre, M. J. (1984).*Theory and Applications of Correspondence Analysis*.London: Academic Press.

Halko, N., Martinsson, P. G., & Tropp, J. A. (2011).“Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions.”*SIAM Review* 53:217–288.

Hancock, G. R., and Mueller, R. O. (2001).“Rethinking Construct Reliability within Latent Variable Systems.”In *Structural Equation Modeling: Present and Future – A Festschrift in Honor of Karl Jöreskog*, edited by R. Cudeck, S. du Toit, and D. Sörbom, 195–216.Lincolnwood, IL: Scientific Software International.

Hand, D., Mannila, H., and Smyth, P. (2001).*Principles of Data Mining*.Cambridge, MA: MIT Press.

Harris, C. W., and Kaiser, H. F. (1964).“Oblique Factor Analytic Solutions by Orthogonal Transformation.”*Psychometrika* 32:363–379.

Hartigan, J. A. (1981).“Consistency of Single Linkage for High–Density Clusters.”*Journal of the American Statistical Association* 76:388–394.

Hastie, T., Tibshirani, R., and Friedman, J. H.(2009).*The Elements of Statistical Learning: Data Mining, Inference, and Prediction*.2nd ed. New York: Springer Verlag.

Hinton, G. E., and Roweis, S. T. (2002).“Stochastic Neighbor Embedding.”*Advances in Neural Information Processing Systems* 15:833–840.

Hoskuldsson, A. (1988).“PLS Regression Methods.” *Journal of Chemometrics* 2:211–228.

Hoeffding, W. (1948).“A Non-Parametric Test of Independence.”*Annals of Mathematical Statistics* 19:546–557.

Hu, L.-T., and Bentler, P. M. (1999).“Cutoff Criteria for Fit Indices in Covariance Structure Analysis: Conventional Criteria versus New Alternatives.”*Structural Equation Modeling* 6:1–55.

Huber, P. J. (1964).“Robust Estimation of a Location Parameter.”*Annals of Mathematical Statistics* 35:73–101.

Huber, P. J. (1973).“Robust Regression: Asymptotics, Conjecture, and Monte Carlo.”*Annals of Statistics* 1:799–821.

Huber, P. J., and Ronchetti, E. M. (2009).*Robust Statistics*.2nd ed. Hoboken, NJ: John Wiley & Sons.

Jackson, J. E. (2003).*A User's Guide to Principal Components*.Hoboken, NJ: John Wiley & Sons.

Jardine, N., and Sibson, R. (1971).*Mathematical Taxonomy*.New York: John Wiley & Sons.

Jöreskog, K. G. (1977).“Factor Analysis by Least-Squares and Maximum Likelihood Methods.”In *Statistical Methods for Digital Computers*, edited by K. Enslein, A. Ralston, and H. Wilf, 125 - 165.New York: John Wiley & Sons.

Kingma, D. P., and Ba, J. (2014).“Adam: A method for stochastic optimization.” *arXiv preprint arXiv:1412.6980*.

Kline, R. B. (2016).*Principles and Practice of Structural Equation Methodology*.4th ed. New York: The Guilford Press.

Kohonen, T. (1989).*Self-Organization and Associative Memory*.3rd ed. Vol. 8 of Springer Series in Information.Berlin: Springer-Verlag.

Kohonen, T. (1990).“The Self-Organizing Map.”*Proceedings of the IEEE* 78:1464–1480.

Lindberg, W., Persson, J.-A., and Wold, S. (1983).“Partial Least-Squares Method for Spectrofluorimetric Analysis of Mixtures of Humic Acid and Ligninsulfonate.”*Analytical Chemistry* 55:643–648.

LeRoux, B., and Rouanet, H. (2010).Multiple Correspondence Analysis.Vol.07–163 of Sage University Paper Series on Quantitative Applications in the Social Sciences.Thousand Oaks, CA: Sage Publications.

Maiti, S. S., and Mukherjee, B. N. (1991).“Two New Goodness-of-Fit Indices for Covariance Matrices with Linear Structures.”*British Journal of Mathematical and Statistical Psychology* 44:153–180.

Mardia, K., Kent, J., and Bibby, J. (1980).*Multivariate Analysis*.New York: Academic Press.

Mason, R. L., and Young, J. C. (2002).*Multivariate Statistical Process Control with Industrial Applications*.Philadelphia: SIAM.

May, J. P. (1992).*Simplicial Objects in Algebraic Topology*.Vol. 11.Chicago: University of Chicago Press.

Maydeu-Olivares, A., Shi, D., and Rosseel, Y. (2017).“Assessing Fit in Structural Equation Models: A Monte-Carlo Evaluation of RMSEA Versus SRMR Confidence Intervals and Tests of Close Fit.”*Structural Equation Model: A Multidisciplinary Journal* 25:389–402.

McDonald, R. P. (1999).*Test Theory: A Unified Approach*.Mahwah, NJ: Erlbaum.

McDonald, R. P., and Marsh, H. W. (1990).“Choosing a Multivariate Model: Noncentrality and Goodness of Fit.”*Psychological Bulletin* 107:247–255.

McInnes, L., Healy, J., and Melville, J. (2018).“UMAP: Uniform manifold approximation and projection for dimension reduction.” *arXiv preprint arXiv:1802.03426*.

McLachlan, G. J., and Krishnan, T. (1997).*The EM Algorithm and Extensions*.New York: John Wiley & Sons.

McQuitty, L. L. (1957).“Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies.”*Educational and Psychological Measurement* 17:207–229.

Milligan, G. W. (1980).“An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms.”*Psychometrika* 45:325–342.

Nelson, P. R. C., Taylor, P. A., and MacGregor, J. F. (1996).“Missing Data Methods in PCA and PLS: Score calculations with incomplete observations.”*Chemometrics and Intelligent Laboratory Systems* 35:45–65.

Nunnally, J. C. (1978).*Psychometric theory*.2nd ed. New York: McGraw-Hill.

Penny, K. I. (1996).“Appropriate Critical Values When Testing for a Single Multivariate Outlier by Using the Mahalanobis Distance.”*Journal of the Royal Statistical Society, Series C* 45:73–81.

Press, W. H, Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1998).*Numerical Recipes in C: The Art of Scientific Computing*.2nd ed. Cambridge, England: Cambridge University Press.

Rasch, G. (1980).*Probabilistic Models for Some Intelligence and Attainment Tests*.Chicago: University of Chicago Press.

Rodriguez, A., Reise, S. P., and Haviland, M. G. (2016).“Evaluating Bifactor Models: Calculating and Interpreting Statistical Indices.*Psychological Methods* 21:137.

Saad, D. (1998).“Online algorithms and stochastic approximations.”*Online Learning* 5(3):6.

SAS Institute Inc.(1983).*SAS Technical Report A-108: Cubic Clustering Criterion*.Cary, NC: SAS Institute Inc. https://support.sas.com/kb/22/addl/fusion_22540_1_a108_5903.pdf.

SAS Institute Inc.(2022a).“The CALIS Procedure.”*SAS/STAT® User's Guide*.Cary, NC: SAS Institute Inc. https://go.documentation.sas.com/api/collections/pgmsascdc/v_034/docsets/statug/content/calis.pdf.

SAS Institute Inc.(2022b).“The CANDISC Procedure.”*SAS/STAT® User's Guide*.Cary, NC: SAS Institute Inc. https://go.documentation.sas.com/api/collections/pgmsascdc/v_034/docsets/statug/content/candisc.pdf.

SAS Institute Inc.(2022c).“The FACTOR Procedure.”*SAS/STAT® User's Guide*.Cary, NC: SAS Institute Inc. https://go.documentation.sas.com/api/collections/pgmsascdc/v_034/docsets/statug/content/factor.pdf.

SAS Institute Inc.(2022d).“The FASTCLUS Procedure.”*SAS/STAT® User's Guide*.Cary, NC: SAS Institute Inc. https://go.documentation.sas.com/api/collections/pgmsascdc/v_034/docsets/statug/content/fastclus.pdf.

SAS Institute Inc.(2020e).“The MIXED Procedure.”*SAS/STAT® User's Guide*.Cary, NC: SAS Institute Inc. https://go.documentation.sas.com/api/collections/pgmsascdc/v_034/docsets/statug/content/mixed.pdf.

SAS Institute Inc.(2022f).“The PLS Procedure.”*SAS/STAT® User's Guide*.Cary, NC: SAS Institute Inc. https://go.documentation.sas.com/api/collections/pgmsascdc/v_034/docsets/statug/content/pls.pdf.

SAS Institute Inc.(2022g).“The VARCLUS Procedure.”*SAS/STAT® User's Guide*.Cary, NC: SAS Institute Inc. https://go.documentation.sas.com/api/collections/pgmsascdc/v_034/docsets/statug/content/varclus.pdf.

Schafer, J., and Strimmer, K. (2005).“A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics.”*Statistical Applications in Genetics and Molecular Biology* 4 Article 32.

Sneath, P. H. A. (1957).“The Application of Computers to Taxonomy.”*Journal of General Microbiology* 17:201–226.

Sokal, R. R., and Michener, C. D. (1958).“A Statistical Method for Evaluating Systematic Relationships.”*University of Kansas Science Bulletin* 38:1409–1438.

Steiger, J. H. (1989).*EzPATH: A Supplementary Module for SYSTAT and SYGRAPH*.Evanston, IL: Systat.

Steiger, J. H. (1990).“Structural Model Evaluation and Modification: An Interval Estimation Approach.”*Multivariate Behavioral Research* 25:173–180.

Tobias, R. D. (1995).“An Introduction to Partial Least Squares Regression.”In *Proceedings of the Twentieth Annual SAS Users Group International Conference*, 1250–1257.Cary, NC: SAS Institute Inc. http://www.sascommunity.org/sugi/SUGI95/Sugi-95-210%20Tobias.pdf.

Tracy, N. D., Young, J. C., and Mason, R. R. (1992).“Multivariate Control Charts for Individual Observations.”*Journal of Quality Technology* 24:88–95.

Umetrics.(1995).*Multivariate Analysis (3-day course)*.Winchester, MA.

Van der Maaten, L. (2014).“Accelerating t-SNE using tree-based algorithms.”*The Journal of Machine Learning Research* 15:3221–3245.

Van der Maaten, L., and Hinton, G. E. (2008).“Visualizing Data using t-SNE.”*Journal of Machine Learning Research* 9(11).

Waern, Y. (1972).“Structure in Similarity Matrices: A Graphic Approach.”*Scandinavian Journal of Psychology* 13:5–16.

West, S. G., Taylor, A. B., and Wu, W. (2012).“Model Fit and Model Selection in Structural Equation Modeling.”In *Handbook of Structural Equation Modeling*, edited by R. H. Hoyle, 209–231.New York: The Guilford Press.

White, K. P., Jr., Kundu, B., and Mastrangelo, C. M. (2008).“Classification of Defect Clusters on Semiconductor Wafers Via the Hough Transform.”*IEEE Transactions on Semiconductor Manufacturing* 21:272–278.

Wold, S. (1994).“PLS for Multivariate Linear Modeling.”In *QSAR: Chemometric Methods in Molecular Design.Methods and Principles in Medicinal Chemistry*, edited by H. van de Waterbeemd, pp. 195–218.Weinheim, Germany: Verlag-Chemie.

Wold, S., Sjostrom, M., and Eriksson, L. (2001).“PLS-Regression: A Basic Tool of Chemometrics.”*Chemometrics and Intelligent Laboratory Systems* 58:109–130.

附录 C

技术许可声明

- Scintilla is Copyright © 1998–2017 by Neil Hodgson <neilh@scintilla.org>.

  All Rights Reserved.

  Permission to use, copy, modify, and distribute this software and its documentation for any purpose and without fee is hereby granted, provided that the above copyright notice appear in all copies and that both that copyright notice and this permission notice appear in supporting documentation.

  NEIL HODGSON DISCLAIMS ALL WARRANTIES WITH REGARD TO THIS SOFTWARE, INCLUDING ALL IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS, IN NO EVENT SHALL NEIL HODGSON BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THIS SOFTWARE.
- Progressë® Telerik® UI for WPF: Copyright © 2008-2019 Progress Software Corporation.All rights reserved.Usage of the included Progress® Telerik® UI for WPF outside of JMP is not permitted.
- ZLIB Compression Library is Copyright © 1995-2005, Jean-Loup Gailly and Mark Adler.
- Made with Natural Earth.Free vector and raster map data @ naturalearthdata.com.
- Packages is Copyright © 2009–2010, Stéphane Sudre (s.sudre.free.fr).All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

  Neither the name of the WhiteBox nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED.IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES

(INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- iODBC software is Copyright © 1995–2006, OpenLink Software Inc and Ke Jin (www.iodbc.org).All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  – Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
  – Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
  – Neither the name of OpenLink Software Inc. nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED.IN NO EVENT SHALL OPENLINK OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- This program, "bzip2", the associated library "libbzip2", and all documentation, are Copyright © 1996–2019 Julian R Seward.All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  1.Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  2.The origin of this software must not be misrepresented; you must not claim that you wrote the original software.If you use this software in a product, an acknowledgment in the product documentation would be appreciated but is not required.

  3.Altered source versions must be plainly marked as such, and must not be misrepresented as being the original software.

4.The name of the author may not be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE AUTHOR “AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED.IN NO EVENT SHALL THE AUTHOR BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Julian Seward, jseward@acm.org

bzip2/libbzip2 version 1.0.8 of 13 July 2019

- R software is Copyright © 1999–2012, R Foundation for Statistical Computing.
- MATLAB software is Copyright © 1984-2012, The MathWorks, Inc. Protected by U.S. and international patents.See www.mathworks.com/patents.MATLAB and Simulink are registered trademarks of The MathWorks, Inc. See www.mathworks.com/trademarks for a list of additional trademarks.Other product or brand names may be trademarks or registered trademarks of their respective holders.
- libopc is Copyright © 2011, Florian Reuter.All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  – Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
  – Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and / or other materials provided with the distribution.
  – Neither the name of Florian Reuter nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS “AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED.IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF

USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- libxml2 - Except where otherwise noted in the source code (e.g. the files hash.c, list.c and the trio files, which are covered by a similar license but with different Copyright notices) all the files are:

  Copyright © 1998–2003 Daniel Veillard.All Rights Reserved.

  Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

  The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

  THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT.IN NO EVENT SHALL DANIEL VEILLARD BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

  Except as contained in this notice, the name of Daniel Veillard shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Software without prior written authorization from him.

- Regarding the decompression algorithm used for UNIX files:

  Copyright © 1985, 1986, 1992, 1993

  The Regents of the University of California.All rights reserved.

  THIS SOFTWARE IS PROVIDED BY THE REGENTS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED.IN NO EVENT SHALL THE REGENTS OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

  1.Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

2.Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

3.Neither the name of the University nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

- Snowball is Copyright © 2001, Dr Martin Porter, Copyright © 2002, Richard Boulton.

  All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  1.Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  2.Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

  3.Neither the name of the copyright holder nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS \"AS IS\" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED.IN NO EVENT SHALL THE COPYRIGHT HOLDER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- Pako is Copyright © 2014–2017 by Vitaly Puzrin and Andrei Tuputcyn.

  Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the “Software”), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

  The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED “AS IS”, WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT.IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

- HDF5 (Hierarchical Data Format 5) Software Library and Utilities are Copyright 2006–2015 by The HDF Group.NCSA HDF5 (Hierarchical Data Format 5) Software Library and Utilities Copyright 1998-2006 by the Board of Trustees of the University of Illinois.All rights reserved.DISCLAIMER: THIS SOFTWARE IS PROVIDED BY THE HDF GROUP AND THE CONTRIBUTORS “AS IS” WITH NO WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED.In no event shall The HDF Group or the Contributors be liable for any damages suffered by the users arising out of the use of this software, even if advised of the possibility of such damage.
- agl-aglfn technology is Copyright © 2002, 2010, 2015 by Adobe Systems Incorporated.All Rights Reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  – Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
  – Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
  – Neither the name of Adobe Systems Incorporated nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS “AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED.IN NO EVENT SHALL THE COPYRIGHT HOLDER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- dmlc/xgboost is Copyright © 2019 SAS Institute.

  Licensed under the Apache License, Version 2.0 (the “License”); you may not use this file except in compliance with the License.You may obtain a copy of the License at

  http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.See the License for the specific language governing permissions and limitations under the License.

- libzip is Copyright © 1999–2019 Dieter Baron and Thomas Klausner.

  This file is part of libzip, a library to manipulate ZIP archives.The authors can be contacted at <libzip@nih.at>.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  1.Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  2.Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

  3.The names of the authors may not be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE AUTHORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED.IN NO EVENT SHALL THE AUTHORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- OpenNLP 1.5.3, the pre-trained model (version 1.5 of en-parser-chunking.bin), and dmlc/xgboost Version .90 are licensed under the Apache License 2.0 are Copyright © January 2004 by Apache.org.

  You may reproduce and distribute copies of the Work or Derivative Works thereof in any medium, with or without modifications, and in Source or Object form, provided that You meet the following conditions:

  - You must give any other recipients of the Work or Derivative Works a copy of this License; and
  - You must cause any modified files to carry prominent notices stating that You changed the files; and
  - You must retain, in the Source form of any Derivative Works that You distribute, all copyright, patent, trademark, and attribution notices from the Source form of the Work, excluding those notices that do not pertain to any part of the Derivative Works; and
  - If the Work includes a "NOTICE" text file as part of its distribution, then any Derivative Works that You distribute must include a readable copy of the attribution notices contained within such NOTICE file, excluding those notices that do not pertain to any part of the Derivative Works, in at least one of the following places: within a NOTICE text file distributed as part of the Derivative Works; within the Source form or documentation, if provided along with the Derivative Works; or, within a display generated by the Derivative Works, if and wherever such third-party notices normally appear.The contents of the NOTICE file are for informational purposes only and do not modify the License.You may add Your own attribution notices within Derivative Works that You distribute, alongside or as an addendum to the NOTICE text from the Work, provided that such additional attribution notices cannot be construed as modifying the License.
  - You may add Your own copyright statement to Your modifications and may provide additional or different license terms and conditions for use, reproduction, or distribution of Your modifications, or for any such Derivative Works as a whole, provided Your use, reproduction, and distribution of the Work otherwise complies with the conditions stated in this License.

- LLVM is Copyright © 2003–2019 by the University of Illinois at Urbana-Champaign.

  Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License.You may obtain a copy of the License at:

  http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.See the License for the specific language governing permissions and limitations under the License.

- clang is Copyright © 2007–2019 by the University of Illinois at Urbana-Champaign.

  Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License.You may obtain a copy of the License at:

  http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS", WITHOUT WARRANTIES OR CONDITIONS OF

ANY KIND, either express or implied.See the License for the specific language governing permissions and limitations under the License.

- lld is Copyright © 2011–2019 by the University of Illinois at Urbana-Champaign.

  Licensed under the Apache License, Version 2.0 (the “License”); you may not use this file except in compliance with the License.You may obtain a copy of the License at:

  http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS”, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.See the License for the specific language governing permissions and limitations under the License.

- libcurl is Copyright © 1996–2021, Daniel Stenberg, daniel@haxx.se, and many contributors, see the THANKS file.All rights reserved.

  Permission to use, copy, modify, and distribute this software for any purpose with or without fee is hereby granted, provided that the above copyright notice and this permission notice appear in all copies.

  THE SOFTWARE IS PROVIDED “AS IS”, WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OF THIRD PARTY RIGHTS.IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

  Except as contained in this notice, the name of a copyright holder shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Software without prior written authorization of the copyright holder.

- On the Windows operating system, JMP utilizes the OpenBLAS library.OpenBLAS is licensed under the 3-clause BSD license.Full license text follows: Copyright © 2011-2015, The OpenBLAS Project

  All rights reserved.Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  1.Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  2.Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

  3.Neither the name of the OpenBLAS project nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED.IN NO EVENT SHALL THE OPENBLAS PROJECT OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.