魔 力™

(英文简称: **C**hem**D**ata**S**olution)

一复杂数据智慧分析处理软件系统

用户使用手册

大连达硕信息技术有限公司

Dalian ChemDataSolution Information Technology Co. Ltd

2014 年 12 月

目　　录

第一章　欢迎

欢迎使用本手册。

1.1 关于本软件

复杂数据智慧分析处理软件系统(英文名称: ChemDataSolution)，以下简称“本软件”或“该软件”，是一款先进的数据分析软件，由大连达硕信息技术有限公司独立开发，是公司相关领域科学家长期研究积累的智慧结晶。基于行业领先的复杂数据处理算法流框架，表现优异的数据处理方法和良好的用户体验，可智慧地解决“三高”数据分析中的信息提取与挖掘问题，从而辅助科学决策(“三高”是指高维、高通量和高复杂度)。本软件涵括丰富的数据处理方法，提供从数据预处理到特征选择，探索性分析到模式识别，定性定量模型构建到未知样本验证与预测的整体解决方案。软件功能全面，使用智能便捷，结果准确可靠，用户体验优越，应用范围广泛(如下表)。

本软件仅是公司的主打产品之一。我们亦研发并提供其他数据处理软件产品及服务，特别是不同行业领域的数据处理与信息挖掘个性化整体解决方案。

联系方式

地址: 大连市高新区礼贤街 32 号 B 座 505 (1-2)室

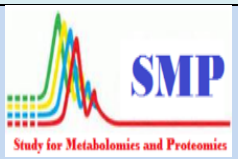
电话/传真: 0411-84753876；移动电话: 13842635729

电邮: contact@chemdatasolution.com

网址: www.ChemDataSolution.com

其他更多联系方式可浏览公司网址: www.chemdatasolution.com。

本软件所能解决的部分问题示例。

序号	行业描述	图示	具体应用点示例
1	仪器制造，仪器软、硬件结合与二次开发		色谱、质谱和光谱等仪器所产生原始数据分析处理的整体解决方案；硬件部件的二次开发，以及在线分析仪器所产生数据的分析处理等。
2	代谢组学与蛋白组学		代谢小分子与蛋白大分子标志物的发现及验证；不同疾病阶段与健康状态的分析及判别。
3	食品、油品与农产品的分析检测及安全		食品掺伪与判别，成份与添加剂分析和生产过程控制；油品品质分析，及农产品产地溯源等。
4	药物分析与中药质控		药物代谢组学分析，及复杂多组份定性定量分析；中药指纹图谱与质量控制；生产过程控制。
5	烟草与烟用香精香料质控，卷烟在线分析		烟草与烟用香精香料成份分析；香精香料智能勾兑、配比与辅助调香；烟草(在线)近红外分析；科学感官评吸，定性定量模型，以及模型转换。
6	环境的监测与监控		土壤、水质和空气中毒害物定性定量；环境监测数据多元分析；环境影响因素与监控。
7	红、白、啤酒分析		酿酒葡萄成熟度分析；酒品品种分析、真假识别，及产地与品质鉴定；香气成份等重要组份的定性定量分析；勾兑与指纹图谱技术分析。
8	检验检疫与政府质监		移动与便携式检测数据分析；检测数据处理与日常分析智慧化管理；质监“大数据”分析与政府决策支持。
9	科学研究 (高校与科研院所等)		数据质量提高；变量选择；探索性分析；模式识别；定量模型；QSAR 与 QSPR 研究；不同数据处理方法的结果比较，发表学术研究论文。

10	其他		石油化工、珠宝鉴定、司法鉴定、考古鉴定等。
	本软件功能之外的个性化数据处理与信息挖掘解决方案，请直接与我们联系。		

1.2. 关于达硕

大连达硕信息技术有限公司(Dalian ChemDataSolution Information Technology Co. Ltd)，以下简称“达硕”或“本公司”，座落在大连高新技术产业园区内。公司专注复杂多变量数据的智慧分析处理与信息提取挖掘，是目前国内外极少数具备创新性发展和融合数学、统计学、化学计量学、化学与生物信息学，人工智能以及计算机技术，提供数据处理整体解决方案，尤其是科学仪器数据(如色谱、质谱和光谱等)，服务智慧决策的高新技术企业。

公司团队由具有多年海外背景的交叉学科领域教授和科学家，以及多位具有硕士、博士和博士后以上学历或经历的专业人士构成，技术力量非常雄厚。公司拥有团队独创的系列复杂“三高”数据智能挖掘算法及相关知识产权，众多优越数据处理方法的软件实现与创新能力，为客户提供系列智慧高端软件产品，数据处理服务与咨询业务，以及个性化数据处理需求的整体解决方案。

值得信赖的产品与服务:

- 色谱、质谱与光谱数据处理和信息挖掘。
- 智能化软件开发、数据处理服务与咨询。
- 仪器软、硬件的结合与嵌入式数据处理。
- 化学及生物大数据处理与智能决策支持。
- 复杂体系分析解决方案与数据解决方案。

1.3. 版权声明

本软件由大连达硕信息技术有限公司独家持有，并拥有完整的自主知识产权，包括专利、

专利申请、软件著作权、商标、或其它知识产权，受相关法律保护，未经达硕公司的书面许可，任何单位、组织或个人均不得以任何形式或理由对该产品及其商标的全部或部分进行使用、复制、破解、修改、传播、抄录、与其它产品捆绑式销售，或其它损害达硕公司利益的行为。

凡有侵犯本公司权益的行为，我们必将依法追究其法律全部责任。

举报联系方式

电邮: contact@chemdatasolution.com；电话: 0411-84753876。

其他更多联系方式可浏览公司网址: www.chemdatasolution.com。

1.4. 如何使用本手册

本手册详细介绍产品功能和操作使用方法，必要时亦介绍相关预备知识及知识延伸，以求让用户可更好、更深刻地理解本软件的使用，并能解读数据处理所产生的结果和意义，从而得心应手地解决实际问题。

依照内容介绍的先后顺序，如下表概述本手册的主要内容。

章节	标题	图示	内容概述
1	欢迎		**欢迎信息:** 关于本产品和达硕公司。
2	走近 ChemDataSolution		**快速了解本软件:** 包括功能简介、软件特色、典型应用领域与实例。
3	技术术语与名词解释		**软件的基础和基本点:** 数据结构、数据类型、数学坐标与化学坐标、数据等长处理、绘图的优先性、内部绘图与外部绘图、自变量与因变量、工程、工程栏、文件夹、节点、算法流(亦称“批”)、

			批处理。
4	用户界面		**软件的主界面概述:** 主窗口、功能菜单区、工程导航栏区、其他辅助功能区。
5	软件功能及其他		**逐个介绍:** 文件、主页、图形与图形操作、预处理、变量选择、建模、预测、窗口、帮助；数据处理方法概述、参考文献等。

用户既可通过完整阅读本手册，以通晓产品的使用，亦可以在使用本软件的过程中，针对性地浏览或搜索任意感兴趣的内容，以便获得所需要的信息。特别地，可针对软件使用过程中出现的不同问题，有针对性地搜索相应关键词，快速获得需要的结果。

使用过程出现的任何疑问或问题，亦可即时与大连达硕信息技术有限公司联系。

第二章　走近 ChemDataSolution

本章内容旨在帮助用户快速入门，在没有详尽阅读本手册，以及使用本软件产品的情况下，了解本软件的主要功能、特色和应用领域等。用户亦可通过系统中“主页”或“帮助”菜单下“用户向导”快速了解本产品的使用。

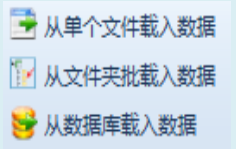
基于此，本章主要简述如下四个方面的内容:

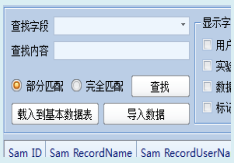
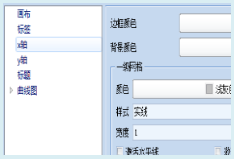
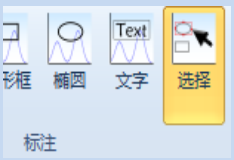
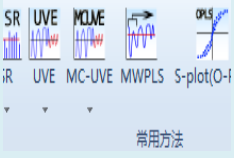
- **产品功能。**
- **重要特色。**
- **快速入门。**
- **典型应用领域与实例。**

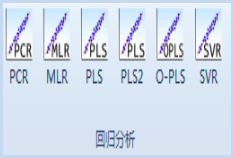
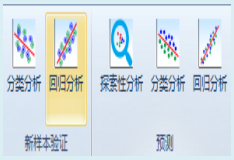
2.1. 产品功能

本产品提供复杂多变量数据处理的整体解决方案：从数据到图形，从数据预处理到变量选择，从探索性分析到分类，从定量模型构建到新样本验证与预测，从工程文件管理到报表。本产品旨在以极佳的用户体验，智慧地解决“三高”数据的分析处理与信息挖掘问题：基于领先的数据处理算法与算法流机制，减少用户对算法理解的要求和对方法使用中的频繁干预，并以直观方式呈现所有计算得到的中间结果和最终结果(包括表格和图形等)。

产品的主要功能见下表。

章号	功能名称	系统截图	功能概述
1	数据载入		本软件包括三种数据载入方式，即从单个文件、文件夹和数据库导入数据；适应复杂结构数据的智慧载入(如文件夹内全部数据自动批载入)；数据类型包括 txt, CSV, excel, Mat, SPC, JCAMP-DX 等。

2	数据操作		原始数据加载到数据表后，用户可对数据进行各种操作和管理，以便初步了解数据的整体情形，或对数据进行分析前的必要处理。
3	数据库与数据管理		本软件集成数据库的功能，管理被分析处理的数据。数据表征样本，数据库同时实现对数据和样本信息管理，以及数据保存和载入。
4	图形		图形是数据的可视化直观表达。本软件实现主要二维和三维图形绘制，并可实现自定义的绘图方式。
5	图形属性		图形属性修改包括二大类：一是画布、标签、x 轴、y 轴和标题等通用的属性，另一类则是不同类型图形的个性化属性，如线条、符号，以及填充方式等。
6	图形操作		除提供上述完整的属性修改功能外，本软件亦可对图形进行丰富的操作，以便更好地获取图形信息，包括标注、缩放和标记等。
7	数据预处理		数据预处理是了解数据或提高质量数据信息提取的关键步骤。本软件完整提供包括各种常用的预处理功能，可实现绝大部分情况下的多变量数据分析需求，包括插值、转置、加入噪声、标准化、标度化和数据运算等一般的操作，亦包括平滑、求导、背景和漂移校正、多元散射校正、正交信号校正和去趋势化等高级数据预处理方法。
8	变量选择		本软件涵括主要表现优异的变量选择方法，包括经典方法、常用方法和基于模型集群分析的方法；不同方法可分别用于分类和回归分析中的变量选择。

9	探索性分析		探索性分析在只有数据(自变量)矩阵 **X**，且没有其他数据先验信息的情况下，对数据进行初步探究，了解数据样本(或变量)间的相关关系、相似性与差异性，或变量重要性等。
10	分类建模		分类建模是在同时提供数据(自变量)矩阵 **X** 和样本类别(2 类或更多类)属性 **y** 后，通过算法构建样本间的分类(判别)模型，并可将模型用于新样本的验证或未知样本的预测等。
11	回归建模		回归建模是在同时提供数据(自变量)矩阵 **X** 和因变量属性 **y**(如化合物浓度或物理化学性质值等)后，通过算法构建该属性与数据矩阵间的定量关系模型，并可将模型用于新样本的验证或未知样本的预测等。
12	验证与预测		参照上述第 10 和第 11 章，用户所构建的分类或回归模型，可用于新样本的验证或未知样本的预测，这也是构建模型重要意义。
13	辅助功能		作为一个完整的产品，本软件在提供上述数据处理功能的同时，亦提供其他丰富的辅助功能，一方面用于更好地完成数据分析需求(比如节点及节点管理，批的构建及应用)，另一方面则提供用户向导、结果报表、程序运行信息、窗口管理和帮助等功能。

2.1.1. 数据载入

数据载入是数据分析处理的初始步骤。新建工程文件(或打开已有工程)后，接下来的第一步便是加载需要分析处理的数据到该工程中，以便使用本软件所提供的数据处理操作和功能，完成对数据的分析。

本软件提供如下三种将数据载入到工程中的方法：

- 从单个文件载入数据。

- 从文件夹载入数据。
- 从数据库载入数据。

从单个文件载入数据，在完全相同的实验条件下对多个不同样本分析所得到的数据，或理论计算数据，因数据间不存在漂移现象而可直接拼接，即加载单个文件，在工程中亦能得到含有多个样本信息的数据矩阵。

从文件夹载入数据，则在用户选择数据文件所存放的路径，并设置相应参数后，系统可智能地同时加载所选文件夹下的所有文件，并拼接成可被分析的数据矩阵。本软件在智慧加载文件夹下的数据时，可同时考虑如下影响数据载入的因素:

- 数据载入到工程中一个已经存在的数据，还是加载为新的数据。
- 数据是否需要进行转置操作，即行与列，或样本与变量的转置。
- 数据中不同列之间的分割方式。
- 数据中是否含有字符，以及字符的删除与保留方式。
- 数据中是否含有化学坐标，若无则考虑是否需要加入坐标，若有且数据载入到已经存在的数据中，则二个坐标是否一致。
- 数据是否需要根据实验条件自定义一个坐标，以及数据的载入。
- 数据的实际载入区间。
- 数据在单个文件中的存在形式，即数据坐标与响应的对应关系。
- 是否忽略对文件夹中某个数据文件的载入。
- 是否根据设定的数据载入参数，批量载入文件夹下的数据。

从上可以看出，本软件完整考虑并解决了文件夹数据批量载入的种种情形，功能强大，可实现数据的智慧载入。

从数据库载入数据，则与从文件夹载入数据类似，同样可实现数据的批量载入，并同样考

虑数据的上述各种复杂情形。数据库数据的来源，则是已经载入到工程中的基本数据，即通过导出工程中已经存在的数据，并附加每个样本的对应信息，包括实验条件和数据样本信息等，从而添加数据库中的记录。

基于数据库和数据库管理的功能，可实现当前数据与数据库中数据的任意比较和建模分析。由于数据库中的数据可能是不同来源或条件变化后所得到的，比如工业生产中原材料的变化，工艺条件和参数的变化，或是某个标准的数据等。因此通过载入并比较分析数据库中的数据，可实现过程分析、参数优化和质量控制等广泛功能。

可载入的数据类型

目前可被载入的数据类型包括 txt, CSV, excel, Mat, SPC, JCAMP-DX 等，可满足色谱、质谱和光谱数据的分析处理需求。

更详细的内容请参考第 6 章(“主页”)中数据载入的有关内容。

2.1.2. 数据操作

数据被载入后，将出现在当前工程的数据节点中。在数据被分析处理前，用户可对数据进行新的划分，产生用于分析的数据子集，或者对目标数据进行系列操作和管理，以便更好地理解被处理的数据等。概括地，本软件提供的数据操作包括如下功能内容：

- 用户任意划分数据，产生子数据：包括行划分、列划分和子矩阵，详见第三章(“技术术语与名词解释”)。
- 数据复制、剪切、粘帖、插入、删除。
- 插入说明性信息。
- 数据查找、替换。
- 数据添加、跳转。
- 数据查找、范围查找、数据合法性检查。
- 升、降序排列。

- 数据导出到数据库。
- 数据(自变量)矩阵 **X** 和因变量矩阵 **y** 的不同操作、管理和转换。

2.1.3. 数据库与数据管理

数据库同时实现对数据及其对应信息的保存和管理。本软件中的数据库可帮助用户实现如下功能:

- 保存和管理数据，以及每个数据的详细说明性信息，比如数据本身以及对应样本的信息，实验分析条件等。
- 保存和管理从工程中选择性导出的任意数据，方便其后的模型比较分析。
- 保存和管理因某一条件变化后获得的数据，比如原材料产地或工艺条件参数变化后的数据，用于与新数据的比较分析。
- 保存和管理建立某个特定模型的数据，方便其后的模型应用。
- 保存和管理标准数据，比如某一中药注射剂指纹图谱国家标准，用于实际样本的分析和评价，以及质量控制。

2.1.4. 图形

图形可简单直观地表达数据的变化。本软件提供多种二维和三维图形的绘制方式，并实现用户自定义绘图，详情请参见 9.2.8.。

二维图形:

- 曲线图。
- 散点图。
- 条形堆积图。
- 填充图。
- 棒状图。

三维图形:

- 三维散点图。
- 三维表面图。

用户自定义图形:

- 实现用户自定义个数的数据矩阵，并以可选择的绘图方式和绘图优先性绘出图形，详情请参见 9.2.8.。

2.1.5. 图形属性

本软件提供丰富的图形属性修改功能，可对涉及图形的所有属性进行修改，既包括图形的通用属性，如画布、标签、坐标轴、和标题等，亦包括不同类型图形的个性化属性，如线条颜色和粗线、符号形状和大小等。

图形属性修改时，系统提供良好用户体验，可在图形和属性修改界面中交互显示需要修改属性的数据样本。

2.1.6. 图形操作

图形操作是指在用户绘制的图形中，提供各种标注、缩放和标记的功能，以实现对图形本身和图形对应的数据的使用和操作。

标注: 在图形中加入各种符号或文字，以说明或强调图形中的关注点。

- 线条。
- 箭头。
- 矩形框。
- 椭圆。
- 文字。
- 选择。

缩放: 放大或缩小图形以查看图形细节。

- 区域缩放。
- 缩放重置。
- 后向缩放。
- 前向缩放。
- 平移。

标记：可从样本或变量方向，标记图形中所体现的数据，并可选择被标记数据，产生新的子数据矩阵。

- 单个标记。
- 以矩形框标记。
- 以矩形框取消标记。
- 反向标记。
- 取消反向标记。

每项功能的介绍详情请参见 9.4.。

2.1.7. 数据预处理

本软件提供非常丰富的数据预处理功能。这里所说的数据预处理，是一个宽泛的概念，整体来说，是通过对数据的某个变换或变化，提高对数据的理解或数据质量，规范化数据，或者去除数据本身的无用或干扰性信息，从而提高数据处理结果的准确性、可用性或模型的预测能力。

一般预处理：原始数据的常用预处理方法，包括:

- 数据转置。
- 数据插值。
- 加入噪声。

数据规范化：原始数据从样本和变量方向标准化或标度化等处理。

- 样本归一化。

- 变量标度化。
- SNV 变换。
- Quantile 标准化。

数据运算: 在原始数据的基础上，通过从数据样本或变量方向的运算，替换或产生新的数据样本或变量。

- 以数学表达式形式产生。
- 创建计算器的形式产生。

提高数据质量(第 I 类方法): 从纯数学的角度提高数据质量，或改善数据样本间的差异。

- 数据平滑，包括 5 个重要的平滑方法。
- 数据求导，包括 4 个重要的求导方法。
- 扣除数据背景，包括自动、手动，以及自动结合手动的方法。
- 数据样本间的漂移校正，包括自动、手动，以及自动结合手动的方法。

提高数据质量(第 II 类方法): 从解决实际问题的角度提高数据质量，消除数据中的无用或干扰信息(比如近红外分析中颗粒散射影响)，从而帮助建立更可靠的模型，或模型预测结果。

- 多元散射校正。
- 正交信号校正。
- 去趋势化。

2.1.8. 变量选择

变量选择是复杂多变量数据分析处理的重要步骤之一，其目的在于寻找或优化组合与建模属性密切关联，可解释性好，彼此间互补性强的关键特征，以构建稳健可靠、泛化能力强的模型。允余的变量不仅降低建模的质量和预测能力，且模型样本数与变量数的比值需大于某一数值（通常为 3∽10)，以保证模型的稳健性和预测精度。

变量选择的标准则需考虑化学(或物理)和数学二个方面，前者是对实际问题本质的理解，后者则从纯数学分析的角度，筛选对模型构建和预测作用最大的一个或多个重要变量。当然针对具体的问题，则可以将二者结合起来使用。

本软件系统同时提供解决分类和回归问题的系列变量选择方法。特别地，系统提供强制性加入或排除某些变量的功能，提高其可用性和用户体验。本软件包括的变量选择算法有:

经典方法

- 不加权方法: 计算每类样本中各变量标准偏差之均值与所有样本中各变量标准偏差的比值，以此判断变量重要性并选择合适的变量。
- 加权方法: 计算每类样本中各变量加权标准偏差之均值与所有样本中各变量标准偏差的比值，以此判断变量重要性并选择合适的变量。
- Fisher 比法: 计算每个变量类内与类间方差的比值，以此判断变量重要性并选择合适的变量。
- 逐步回归方法: 逐步引入或删除进入模型中的变量，考察对模型的影响，基于回归分析与 F 检验评价其不同变量的重要性，选取合适的变量引入模型。

常用方法

- VIP 法: 即 Variable Importance in the Projection，包括针对分类和回归分析的方法；同时考虑回归系数和载荷所构造的特征重要性评价指标，通常以指标值达到 1 作为引入该特征的依据。
- SR 法: 即 Selectivity Ratio，包括针对分类和回归分析的方法；计算每个特征被解释方差与残差方差比值所构造的特征重要性评价指标。
- UVE 法: 即 Uninformative Variable Elimination，包括针对分类和回归分析的方法；通过加入噪声考察回归系数的稳健性来选取合适建模特征的方法。
- MC-UVE 法: 即 Monte-carlo Uninformative Variable Elimination，包括针对分类和回归分析的方法；同时基于蒙特卡罗随机采样分析与无信息变量剔除的变量选择方法。

- MWPLS 法：即 Moving Window Partial Least Squares，以移动窗口扫描的方式构造一系列的子模型，通过评价这些模型的有效性以引入或剔除被选窗口内的特征。
- S-plot 法：S 形图(适合 O-PLS 分析)，包括针对分类和回归分析的方法；同时考虑特征间的协方差和相关性所定义的特征选择指标，因图形通常近似 S 形而得名。

模型集群分析

- CARS 法：即 Competitive Adaptive Reweighted Sampling，包括针对分类和回归分析的方法；通过计算多个重采样子模型的预测误差分布，实现变量集的全面评价，以选取优化的特征组合。
- Random Frog 法：包括针对分类和回归分析的方法；统计分析每个特征在 N 个不同维数模型中被选择的概率，实现变量选择，适合于在高维空间中获得较优的特征组合。
- MIA 法：即 Margin Influence Analysis，专门针对支持向量机方法提出，包括支持向量分类和回归分析的方法；基于蒙特卡罗随机采样构建多个模型，计算相应的 SVM 模型间隔，并统计分析每个特征对分类模型间隔的影响能力实现变量选择，适合于 SVC 方法建模。

本软件涵盖目前受到广泛使用和关注的变量选择算法，明显优于 Unscrambler 和 SIMCA 等软件，真正做到人物我有，人有我优。

2.1.9. 探索性分析

探索性分析是在信息了解有限的情况下，最大限度获取有价值数据信息的有力工具，可以帮助用户理解不同数据样本或变量的差异性或相似性，以及二者间的关联关系。

- PCA 法：即 Principal Component Analysis (PCA)，使用最广泛的数据降维方法之一，实现以最少的主成分数，包涵尽可能多的原始数据被解释方差。
- HCA 法：即 Hierarchical Cluster Analysis，计算样本间的相似性指标(如距离)，将指

标最优的二样本归为一类，重复计算直到所有样本均聚类为止。

- K-means 法： 计算每个样本到类中心的距离，以距离大小判别其类别归属；每次计算完成后，重新计算样本到新类中心的距离；重复划分样本类别，直至每个样本的类别不再发生变化为止。

本部分包括聚类算法。

2.1.10. 分类建模

模式识别是数据处理的最重要的工具之一，分为无监督和有监督二类方法。本部分所述内容主要为有监督的系列方法，即先通过已经类别的数据样本建立分类模型，并在需要时基于模型验证已知类别的新样本，或预测未知类别的样本。

在各行业的具体应用中，分类建模得到频繁使用和广泛关注，比如产品好坏鉴别或质量等级判别，真假辨别与掺假识别，疾病与健康状态分辨与病理过程(阶段)判断等等。
本软件提供如下分类建模的方法，适合于二类及以上分类问题。

- SIMCA 法：即 Soft Independent Modeling of Class Analogy，以 PCA 法获得样本聚类轮廓，构造已知类别样本的 PCA 模型，以此验证并评价未知样本，实现未知样本的分类。
- KNN 法：即 K-Nearest Neighbors，计算未知样本到 K 个已知样本的距离，以距离未知样本最多的已知样本类别作为未知样本的类别。
- PCA-MD 法：即 Principal Component Analysis with Mahalanobis Distance，构造已知类别样本的 PCA 模型，计算未知样本到这些模型的马氏距离，实现未知样本的分类。
- PLS-DA 法：即 Partial Least Squares Discriminant Analysis，构造已知样本的 PLS 模型，以此构造线性判别分类面，用于未知样本的分类。
- PLS2-DA 法：即 Partial Least Squares-2 Discriminant Analysis，构造已知样本的 PLS2

模型，以此构造线性判别分类面，可用于多类(大于二类)未知样本的分类。

- O-PLS-DA 法：即 Orthogonal Partial Least Squares Discriminant Analysis，基于 OPLS 构造已知样本的分类模型，以此构造线性判别分类面，用于未知样本的分类。
- SVC 法：即 Support Vector Classification，基于支持向量机的非线性分类器。

除上述方法外，本软件亦适时提供系列分类建模中所需的交互检验方法，以及模型和结果评价方法等。

2.1.11. 回归建模

回归分析是另一种获得二个数据间定量变化(相关)关系的重要工具。本软件将这二个数据分别称为自变量和因变量，前者亦称数据矩阵(如中药粉末近红外光谱)，而后者亦称属性(响应)矩阵(如中药中的重要药效组份的含量)。

本软件提供的算法，适合于含有一个及一个以上属性矩阵的情形，具体方法包括:

- PCR 法：即 Principle Component Regression，基于主成分分析的回归方法，适合非满秩或强相关数据矩阵的分析。
- MLR 法：即 Multiple Linear Regression，经典的回归分析方法，也是平方误差下的最好线性无偏估计器。
- PLS 法：即 Partial Least Squares Regression，构建数据矩阵 X 与 y 间的定量模型关系；可同时分解自变量矩阵 X 和因变量(响应)矩阵 y 构造模型，特别适合强相关数据矩阵的分析。
- PLS2 法：即 Principle Component Regression -2，不同于常用的偏最小二乘方法(1)，适合含有多个因变量(响应)矩阵 y 的分析。
- O-PLS 法：即 Orthogonal Projection to Latent Structures，正交投影移除自变量矩阵 X 中与因变量(响应)矩阵 y 不相关的变量，建立更稳健模型。
- SVR 法：即 Support Vector Regression，非线性变换数据到高维特征空间，构造线性决策函数以实现线性回归，可达致结构风险最小化(同时考虑期望风险，经验风

险和置信范围)。

除上述方法外，本软件亦适时提供系列回归建模中所需的交互检验方法，以及模型和结果评价方法等。

2.1.12. 验证与预测

验证与预测是指基于已经建立的模型(来自 PCA 分析，分类或回归建模中)，验证新的样本或预测未知样本的结果。显然验证与预测是数据分析处理的本质和根本目的所在。模型构建仅是数据分析的起点，用户往往更关注模型验证或预测的结果表现。试想用户希望获知复杂体系中目标化合物的含量，采用色谱等分离手段，显然费时费钱费力；若通过建立光谱(若近红外)与该化合物含量间的定量模型，基于该模型则可非常快速地预测组份的含量结果。

本软件提供卓越的用户体验，使得用户在模型构建时，便能同时完成新数据样本的验证或预测。详细内容请参见 2.2.关于本软件的重要特色介绍。

2.1.13. 辅助功能

本软件不仅具有丰富的数据分析功能，实现智慧型多变量数据处理，亦包括完整的辅助功能模块，实现从数据到报表输出，以使用户无需借助任何第三方的软件或系统，便可完成数据处理的全流程功能。本软件的主要辅助功能包括:

- 工程式文件管理: 数据处理过程涉及中的所有数据、图形、操作、算法流(批方法)，模型，以及中间结果和最终结果等，均可被完整保存，并在下次打开时恢复。
- 节点与节点管理: 上述工程式的文件管理，在系统中则以节点的形式存在，用户亦可单独保存或导入某个节点，比如某个表现较好的模型。
- 插入数据: 用户可在使用过程中，再次插入任意的数据。
- 偏好设置: 用户可预先设定自己的使用偏好，比如语言、系统更新状态和推荐，报表表头等。

- 参数设置: 用户可预先设置所有数据处理算法的默认参数，避免频繁修改方法参数的麻烦。用户在调用算法时，程序自动调用默认参数，亦可根据即时修改。
- 运行信息: 用户可即时获知数据处理过程中的说明性信息。
- 图表保存: 用户可即时以不同格式输出所产生的图表结果。
- 产生报表: 用户可以任意流程添加输出到报表中的内容。
- 窗口管理: 用户可便捷管理数据处理过程中产生的数据、图形或结果等输出窗口。
- 更新帮助: 用户可快速获得本软件的更新和帮助等。
- 用户向导: 提供快速使用本软件的帮助流程，且用户可在向导文件中实际操作使用本软件。

本软件所提供的人性化辅助功能，极大提高数据处理过程的使用体验。

2.2. 重要特色

本软件最大特色在于努力实现智慧型的复杂多变量数据处理。快捷建立复杂数据处理算法流(批方法)，构建基于批概念的数据处理模式，忽略繁复的单步数据分析操作，达致一键获得处理结果，极大降低对使用者的要求，真正超越传统的数据处理软件产品。具体地说，本软件的重要特色功能可归纳为如下几个方面:

- 算法流(批方法)。
- 一键处理与多模型处理。
- 同步建模、验证与预测。
- 数据批载入与智慧数据处理。
- 数据抽提与重建模。
- 多线程与多核并行计算。
- 卓越用户体验。

2.2.1. 算法流(批方法)

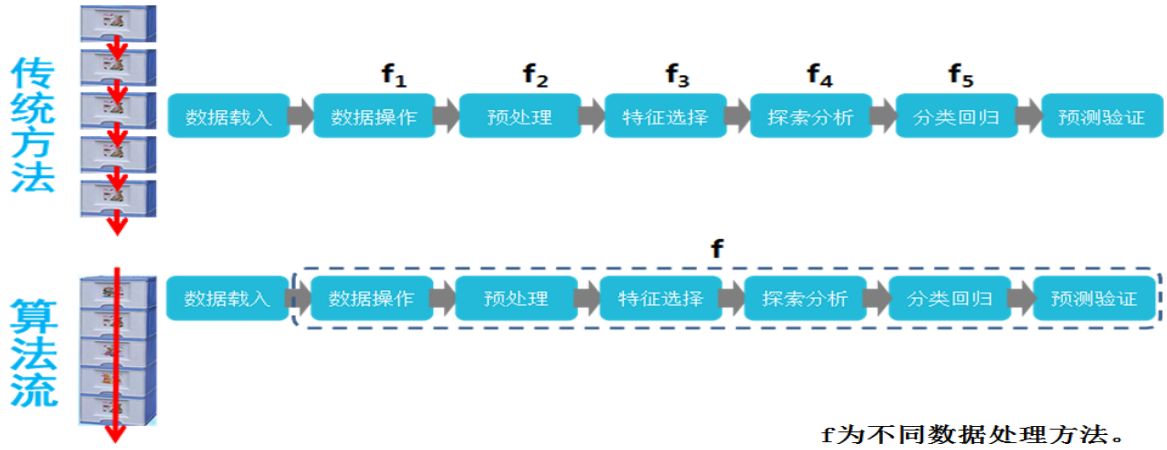
算法流(批方法)思想是本软件的主要亮点和特色之一，特别适合数据处理步骤多、过程繁复的“三高”数据分析与信息提取挖掘。

算法流(批方法)即构造包含不同数据处理方法的整合与优化流程，包括数据批载入、预处理、特征选择、模型构建与未知样本预测等，设置方法参数，即可将待分析数据“注入”算法流中(训练集、校正集、验证集和预测集等)，实现数据快速便捷，准确智能分析，达致智慧型数据分析与信息挖掘之目的。

特别地，算法流构造的变化，可实现复杂数据的一键处理和多模型处理，数据处理方法及参数对分析结果的影响，以及相同数据处理方法(算法流)对不同类型数据集处理的影响。

如下图比较了算法流与传统数据处理方法的差异。概括起来，算法流具有如下显著优势:

- 融合多变量数据处理的多个步骤，减少频繁的数据输入与结果输出。
- 数据样本或变量较多的“大”数据，其分析处理需要消耗较长时间，算法流(批方法)一次性构造需要加入的方法，随即添加被分析的数据，计算完成后即获得最终结果。传统上单个步骤完成后，均需人工干预，无法一次性获得最终的计算或模型结果。
- 优化整合不同行业领域的具体数据处理问题，构造个性化的算法流，可实现数据处理的快速批量化与标准化。
- 丰富任意的数据处理方法组合，便捷研究不同数据处理方法与参数设置对计算结果的影响。
- 特别强调: 算法流实现数据一键处理与多模型处理，以及同步建模、验证与预测。详细内容见 2.2.2 与 2.2.3.。



2.2.2. 一键处理与多模型处理

如上所述，算法流实现不同数据处理方法的逐级串联与优化整合。针对用户需要解决的实际问题，构造个性化的算法流，设置方法参数，并往算法流的入口添加目标数据即可实现全流程分析，且自动保存每步计算的中间结果和最终结果，这便是本软件产品提供的一键处理功能。

针对某个具体问题，用户便可在优选算法流的基础上，将新的待分析数据直接加载到算法流中，可快捷简便、智慧可靠地获得分析结果。

多模型处理则是指用户在构造算法流时，可同时添加多个不同的建模方法，程序便自动判断建模前的分析方法与建模方法，实现各建模方法的结果比较。程序先运行全部建模前的方法，并将结果作为输入分别加载到各建模方法中。

除此之外，用户亦可自由添加、删减或修改算法流中的方法，或者调节方法顺序，修改方法参数等，随心所欲实现快速数据处理。

2.2.3. 同步建模、验证与预测

用户先构造算法流，后添加待分析处理的数据。在添加被处理数据时，用户可一次性同时加载数据训练集、验证集和预测集等，极大减少用户频繁选择数据、数据处理方法，以及

模型的麻烦。传统上需先选择数据处理方法和数据以构建模型后，再将模型应用于新的验证样本或未知预测样本。

训练集、验证和预测的结果以节点文件夹的形式保存，层次清晰，查看方便。详细内容请参见对节点文件夹介绍，以及各章节对数据处理方法结果的操作和说明。

ⓘ 一个工程文件可管理多个被载入的数据，且单个数据可做任意的数据划分(行划分、列划分，以及子矩阵划分)，保证训练集、验证集和预测集数据的丰富来源。

2.2.4. 数据批载入与智慧型数据处理

数据载入到软件系统中，是数据处理的第一步。实现智慧型数据数据处理，则须先实现快速数据载入。传统数据分析处理，包括 Unscrambler 和 SIMCA 等软件，数据载入亦极其繁复费时，甚至需要用户将不同样本的数据事先拼接，给用户带来极大不便。

数据智慧载入的麻烦在于数据本身的复杂性和多样性，详情请参见 2.1.1.部分对数据载入的介绍。如前所述，本软件从技术上完整解决数据智慧导入的难题，使得用户甚至仅需告诉程序文件夹路径(文件保存位置)即可，系统自动考虑各种复杂的情形，实现多文件、多样本数据的载入。

ⓘ 更多关于数据载入的内容，请参见 2.1.1.及其详细介绍。

2.2.5. 数据抽提与重建模

算法流或传统单个数据处理方法运行后得到的可视化图形结果，直观表达数据样本、变量以及二者之间的关系，用户可以此做出某些决断，如奇异值的判别等。

本软件提供基于图形结果的数据抽取，并可将被抽提的部分(样本或变量)重新构建数据和模型，实现数据的二次分析，获得更加符合用户期待的结果。除上述基于图形的数据抽提，本软件亦提供基于数据表格的方式，详情请参见 6.3.1.。

数据重建模亦可在工程导航的右键菜单中完成，详情请参见 4.3.。

2.2.6. 多线程与多核并行计算

数据处理过程往往消耗较大的计算量，尤其当数据含有较多样本和变量时；同时 Monte-carlo 等方法本身运算次数多，亦加大运算的难度。

基于此，本软件加入多线程与多核并行计算的方法，极大地提高计算速度，从“大数据”的分析处理考察，显著优于传统软件系统。

2.2.7. 卓越用户体验

除了上述的丰富数据处理方法和强大的分析能力，本软件尤其追求最佳的数据处理用户体验。事实上，上述 2.1.13.关于系统辅助功能的介绍，绝大部分涉及提高用户体验。除此之外，本软件用户体验亦包括如下内容:

- Ribbon 界面：本软件界面采用 Ribbon 样式，美观简洁，使用方便。
- 我的收藏：工程导航栏中汇聚和管理所有数据、图形、算法流、模型及其结果等等，用户可快速选择有用的部分加入到收藏中，方便快速查看、管理和产生报表等。
- 节点搜索：如上所述，由于工程栏的复杂性，本软件提供快速搜索节点的策略，且可选择性搜索所有节点或当前被选择的节点，方便用户快速查看结果。
- 节点信息查看：针对导航栏中丰富的节点，本软件提供查看节点信息的功能，即用户可即时获得关于节点的全部信息，比如数据大小、行列范围，序号，或图形类型，图形对应的数据信息等。
- 添加注释：用户需对上述节点信息作出补充时，可在注释栏中添加用户的任意内容，方便标记和查看。
- 保存运行信息：程序的运行提供保存等功能，用户可随时打开浏览。
- 右键菜单：本软件邮件菜单丰富，功能强大，使用方便，为用户提供

2.3. 快速入门

快速了解并初步掌握本软件的使用，可从如下三个方面入手:

- 用户向导: 用户点击“主页”或“帮助”菜单下的“用户向导”,可快速了解本产品的功能及使用。
- 走近 ChemDataSolution: 本手册第二章所介绍的内容，整体描述本软件的情况，包括产品功能和重要特色等，可让用户快速入门。
- 快速浏览与使用: 快速浏览本手册，并实际使用本软件的主要功能。

2.4. 典型应用领域与实例

本软件的典型应用领域，已在 1.1.中有所介绍，请参阅，关于本软件所能解决的具体问题示例。

本手册所介绍数据处理实际问题案例，请参见第十六章。

第三章　技术术语与名词解释

在更详细地介绍本软件的使用操作，以及数据处理方法和分析结果前，先介绍重要的名词或术语，以使用户更快更好地掌握软件的使用，同时方便用户理解使用过程中遇到的技术或方法性问题。

- **数据结构。**
- **数据类型。**
- **化学坐标。**
- **数学坐标。**
- **数据等长处理。**
- **自变量。**
- **因变量。**
- **绘图优先性。**
- **外部绘图。**
- **内部绘图。**
- **工程。**
- **工程导航栏。**
- **节点文件夹。**
- **节点。**
- **算法流(批方法)。**
- **批处理。**

3.1. 数据结构

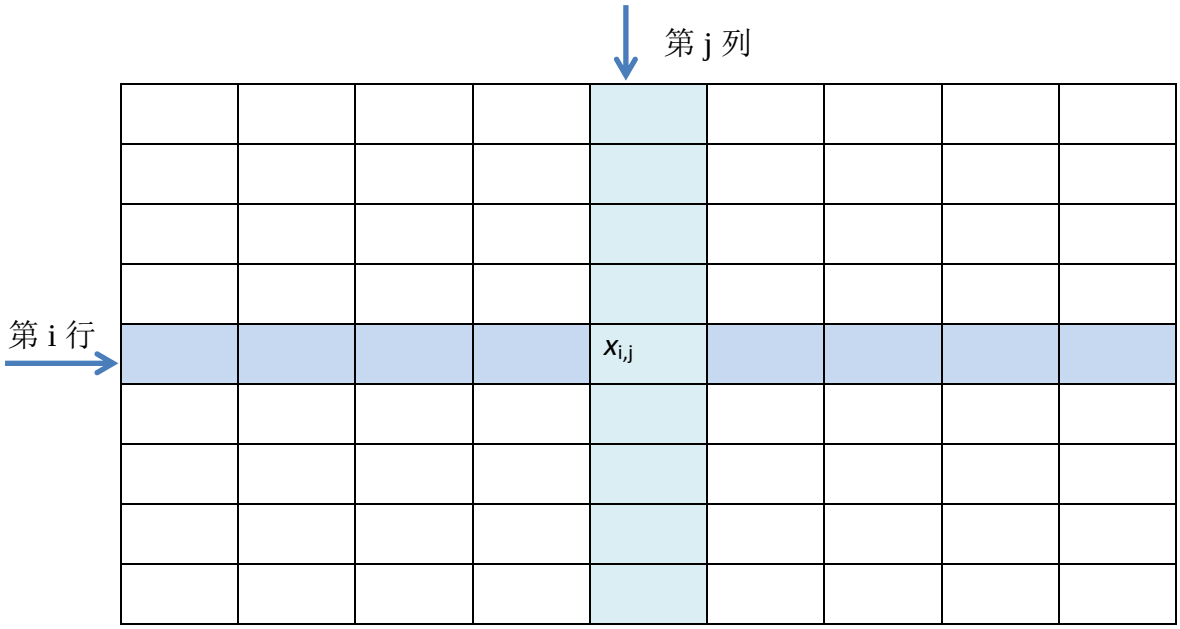
本节主要介绍本软件所处理的数据矩阵的结构，即软件所处理的数据究竟是什么样子的，主要包括如下四个方面的内容:

- 矩阵结构。

- 样本与变量。
- 基本数据表。
- 行划分、列划分与子数据。

3.1.1. 矩阵结构

矩阵是指纵横排列的二维数据表格，通常将含 m 行和 n 列的矩阵 **X** 定义为 $\mathbf{X}_{m\times n}$ 或 **X**(m,n)，其中的任一元素以 $x_{i,j}$ 表示，即指矩阵 **X** 中第 i 行和第 j 列所对应的元素，矩阵的任一行或列则为一向量。以一个 9×9 的数据矩阵为例，其展开形式可由下图表示，其中每个数据空格中有且只有一个标量数据。



上述数据矩阵 X，便是本软件所处理数据形式。当然，用户的实际数据通常情况都较复杂，如数据中同时含有字符等，本软件提供强大的数据载入功能，用户可在数据载入的同时处理这些问题，详情请参见 2.1.1.。

严格地，矩阵用大写、粗体字母表示；而矩阵的某一行或列，即行或列向量，则用小写、粗体字母表示，如 $\mathbf{x}_i$ 或 $\mathbf{x}_j$；而矩阵的某一元素为一个标量，则用小写、斜体字母表示，如 $x_{i,j}$。

本软件亦提供优异的矩阵计算功能，即可对已经存在于工程文件中的矩阵，从行或列的方向进行丰富的数据运算，产生新的行或列，或替换原有的行或列数据。详情请参见 10.10.。

3.1.2. 样本与变量

上述图中的矩阵数据，是从纯数学的角度来介绍的。事实上，本软件所解决的数据处理问题，通常是针对不同行业领域的实际应用，如表 1.1 中列举的问题。

本软件约定

对应图中所述的数据行和列，在实际应用中分别为样本和变量。即，在实际的数据处理中，系统默认数据矩阵的每一行为一个样本，该行向量为样本量测所得到的不同变量；数据矩阵的每一列为一个变量，该列向量为变量在不同样本中的量测值。

举例来说，某中药生产企业按照国家标准，采用色谱技术检测成品质量，每批次被分析的中药产品即为一个样本，而色谱分析得到的中药产品中各活性组份或非活性组份的含量即为描述该样本是否合格的变量(当然从指纹图谱的角度，则整个色谱图谱上的每个点均为一个变量)。如前所述，本软件的主要功能就是分析提取和挖掘不同样本或变量，以及二者间的相关关系。以上述中药生产企业的产品分析为例，则是提供原材料、工艺参数或控制条件等变化后产品质量是否仍然合格，与没有改变条件时相比，差异变化的大小，产品中哪些活性或非活性成份发生了变化，以及究竟哪些因素导致产品质量发生显著改变等，从而“对症下药”找到决策良方。

本软件分析处理多样本、多变量间的关系。当实际数据并非行为样本、列为变量时，本软件提供便捷数据转置功能，可在数据载入时或数据载入后进行相关操作。

3.1.3. 基本数据表

数据载入完成后，数据即导入到当前工程中，被称为基本数据表。如前所述，由于本软件

以直观便捷的工程导航栏形式管理数据、图形、算法流(批方法)、模型及其结果等，因而基本数据表便成为工程导航栏中的一个节点，关于节点的介绍可参考 3.13.和 3.14.。

被导入的数据可添加到已经存在的基本数据表中，亦可新建数据表，并将数据加载其中。因而一个工程文件可同时存在多个基本数据表。不同的基本数据表即可成为模型训练集、验证集或预测集的来源。

被导入的数据之所以被称为基本数据表，实因本软件可对该数据进行任意的划分或“切割”，得到新的数据，详见 3.1.4. 行划分、列划分与子数据。

3.1.4. 行划分、列划分与子数据

如前所述，用户对任意基本数据表进行更进一步的划分或“切割”，即用户可抽提基本数据表中的任一部分数据，重新构造成新的数据，新的数据亦成为模型训练集、验证集或预测集的来源，满足用户个性化的数据选择与建模需要。

- 行划分：指抽提基本数据中一行或多行中的全部数据(列)，构造成新的数据，即抽提一个或多个样本中的所有变量产生新的数据。
- 列划分：指抽提基本数据中一列或多列中的全部数据(行)，构造成新的数据，即抽提一个或多个变量中的所有样本产生新的数据。
- 子数据：排除上述行划分和列划分外的其他数据划分或“分割”形式所产生的数据，即无论针对行或列(样本或变量)，均只取部分数据。

行划分、列划分与子数据分别保存在工程导航栏的不同节点文件夹中，且清晰地对应在其被划分的基本数据表节点下，极大地丰富了本软件的数据产生形式，用户可在解决实际问题中用于标记数据或生成新数据，亦可用于比较不同数据的模型结果。

3.2. 数据类型

如前所述，本软件处理矩阵数据。以数据格式来区分，可载入 txt, CSV, excel, Mat, SPC,

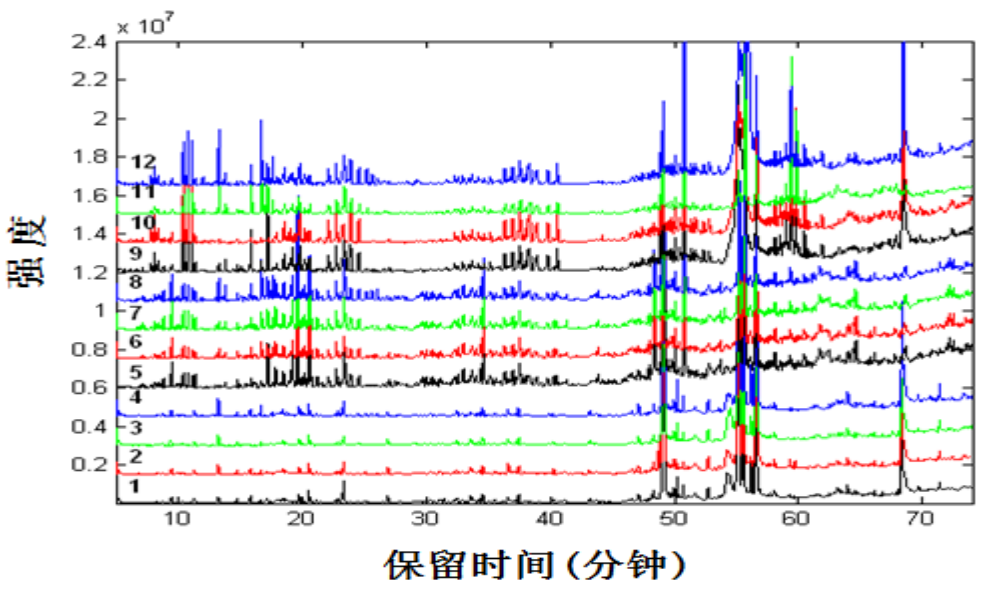
JCAMP-DX 等类型，并提供智慧型的数据载入机制，可智能加载多个样本，构造矩阵数据；以数据来源来区分，即可处理色谱、质谱和光谱等仪器类型数据，亦可处理理论计算所获得的数据，如 QSAR 与 QSPR 中的分子拓扑或量化计算描述符等。

3.3. 化学坐标

以化学量测数据为例，样本量测通常是响应(如光谱吸收值)随时间或某一具有化学意义的指标变化得到的系列数值。化学坐标即指每个响应值所对应的时间或具有化学意义的指标。如下图所示色谱量测中(12 个不同中药人参样本的色谱分析图)，横坐标显示从 7.0 到 73.0 分钟时间范围内的人参色谱分析结果，采样频率为 10 点/1 秒(即每采集一个点，耗时 0.1 秒)，则数据对应的化学坐标即为向量[7.0×60 : 0.1 : 73.0×60]秒。

化学量测所得到的化学坐标具有明确意义。其中 x 轴表示时间或波长等，而 y 轴表示吸光度或响应值，每个刻度代表一个单位，坐标原点代表 0 时刻或 0 波长。因此数据矩阵中的数据携带的信息包括坐标信息。

很显然，每个被量测的实际样本，必然包括一个对应的化学坐标。然而该坐标不一定包含在被载入的数据中。本软件在导入数据时，可根据用户设定的实验条件，自动产生一个对应的化学坐标。



3.4. 数学坐标

与化学坐标对应，数学坐标是指化学坐标中每个数据点的数学序号所组成的坐标。以上述图中的色谱分析为例，其化学坐标为化学坐标为[7.0×60： 0.1： 73.0×60]秒，一共 39601 个数据点，则数学坐标为向量[1： 1： 39601]。因而，在实际的数据载入中，化学坐标可能不存在，但数学坐标必定存在，通常对应从数学序号 1 开始，间距亦为 1，结束点为数据总长度的向量。

本软件以化学坐标优先，若数据存在化学坐标，则数据载入完成后的基本数据表，将对应一个化学坐标，若没有化学坐标，则程序自动为该数据产生一个数学坐标。

3.5. 数据等长处理

实因化学量测的实验条件差异，如色谱分析中的保留时间，质谱分析中的质荷比(m/z)，以及光谱量测中的波长等测量范围的不同，加上采样点间隔的差异，必然导致化学坐标以及对应的化学坐标点数不同，亦即数据长度不同。

如上所述，本软件分析矩阵数据，若不同样本数据长度存在差异，则无法组成矩阵形式进行处理。因而，本软件提供便捷的数据插值功能，以用户的设置为目标，在不改变数据本身特性的情况下，自动将不同长度的样本数据插值成相同长度便于分析。

本软件所涉及的所有差值运算，均为线性插值。

3.6. 自变量

自变量实为纯数学术语，通常以 **X** 表示，指可主动操纵，引起因变量发生某种变化的条件或因素，是因变量发生的原因。

本软件沿用此术语，但意义上有所不同，指构建 **y** = *f*(**X**)模型中的数据矩阵 **X**。举例来说，通过检测农产品(如蔬菜)的近红外光谱，在已构建模型的基础上，便可预测未知样本中的

农残含量，原因在于目标农药的结构特征已经被所量测的近红外光谱表征。此例中的近红外光谱则为自变量，亦称数据矩阵，而农残含量则为因变量。

本软件中所述的自变量和因变量，是从模型构建的角度来说的，不可完全套用纯数学中的概念。

3.7. 因变量

因变量是相对自变量而言的，是自变量产生的结果，即因变量 y 随自变量 X 的变化而变化，通常以 y 表示。在 3.6.中所述的实例中，已经说明因变量及其化学意义。

本软件中，因变量有时亦称为属性值等。

3.8. 绘图优先性

本软件提供二维、三维，以及自定义的绘图方式，详情请参见第九章。

对于一个数据矩阵，若绘制其二维图形，则存在二种可能的绘图操作。以绘制曲线图为例，其一是将数据的每一行绘成曲线图，表达的意义则是该样本中所有变量的变化情况(如前所述，本软件约定每一行表征一个样本。)，而另一种方式则是将数据的每一列绘成曲线图，表达的意义是该变量在所有样本中的变化情况。很显然，即使针对同一数据，上述二种绘图方式所得到的图形，其表达的意义完全不同。

绘图优先性是本软件提供的绘图功能之一，用户在绘制图形时，可自定义上述不同的绘图方式，以行(样本)或列(变量)为优先绘制图形。

3.9. 外部绘图

用户载入的数据，包括数据矩阵 **X** 与属性值 **y** 后，通常以上述化学坐标或数学坐标作为横坐标，以 **X** 或 **y** 的数值作为纵坐标绘图，得到 **X** 或 **y** 值随化学坐标或数学坐标的变化关系。此时绘图时用到外部坐标的信息，称为外部绘图。

3.10. 内部绘图

与外部绘图不同，内部绘图并不以化学坐标或数据坐标作为绘图时的横坐标，而是均以数据矩阵 **X** 中的不同行(样本)或列(变量)，或者样本的不同属性值 **y**(如实际量测值与模型预测值，或主成分分析所得的不同主成分等)作为横坐标与纵坐标。

ⓘ 内部绘图仅针对二维或三维散点图。

3.11. 工程

本软件以工程(文件)的方式组织整个数据分析过程，即数据分析所载入的数据、可视化图形、构造的算法流(批方法)、模型和模型对应的图形或表格结果，以及中间结果等，其表现形式、保存和打开以工程的形式统一管理。

此外，基本数据表划分或“分割”后得到的行划分、列划分或子数据，以及对图形增加图形标注或标记后得到的新形式，均以在保存或打开时存在于工程中，从而保证用户可完整恢复和管理数据、图形及其分析结果。

工程文件的打开与保存，亦可以单个节点的形式进行，即工程中所包含的上述内容，分别以节点形式存在于工程导航栏中，用户可选择性保存其中的某一项或全部内容，打开时则按照保存的内容完整恢复。

ⓘ 本软件工程文件的后缀名为“.CDS”。

3.12. 工程导航栏

工程导航栏是指本软件主界面左侧列表栏，多级管理各个节点，是本软件的中枢部分，连接数据、图形、算法流(批方法)，模型及其结果等。

- 数据导入到工程中后，将在工程导航栏中产生基本数据节点。

- 基本数据表划分或“分割”后得到的行划分、列划分或子数据，在对应数据下，自动在工程导航栏中产生数据节点。
- 数据绘图得到可视化图形，将在工程导航栏中产生对应图形，并统一管理。
- 用户构造的算法流(批方法)，自动在工程导航栏中产生对应节点。
- 用户建构模型后，该模型及对应的图形和表格结果，以及中间结果等，均在工程导航栏中产生节点，统一管理。

工程导航栏是本软件的连接中枢，方便用户统一管理产生的各种结果。系统同时提供工程导航栏的快捷右键菜单功能，针对不同的节点，功能亦动态变化。

3.13. 节点文件夹

如前所述，本软件以工程的形式管理数据处理及其结果，并在工程导航栏中集中体现。节点文件夹则是工程导航栏中管理多个节点的文件夹，并以多级形式存在，比如:

- 基本数据表是工程导航栏中的一个节点文件夹。
- 用户划分或“分割”后，得到行划分、列划分或子数据后，分别以次一级的节点文件夹列于基本数据表节点文件夹下。
- 用户通过多次数据划分或“分割”后，可能产生多个行划分、列划分或子数据，则它们分别按类列于对应的节点文件夹下。

节点文件夹是存在于工程导航栏中，是工程组成部分，以多级形式存在和管理。

3.14. 节点

节点是工程导航栏中的最小单元，亦是构成节点文件夹的最小单元。如前所述，用户在单次数据划分或“分割”中所得到的行划分、列划分或子数据，均在工程导航栏中构成一个节点。每个节点具有丰富的右键菜单，同一类型的节点，其右键功能相同。

算法流(批方法)外，用户直接单击节点，均将在界面主窗口中产生对应的图形或表格结果。

3.15. 算法流(批方法)

算法流(批方法)是本软件的核心之一，其详细介绍请参见 2.2.1.，2.2.2.，2.2.3.，以及 8.5.。

3.16. 批处理

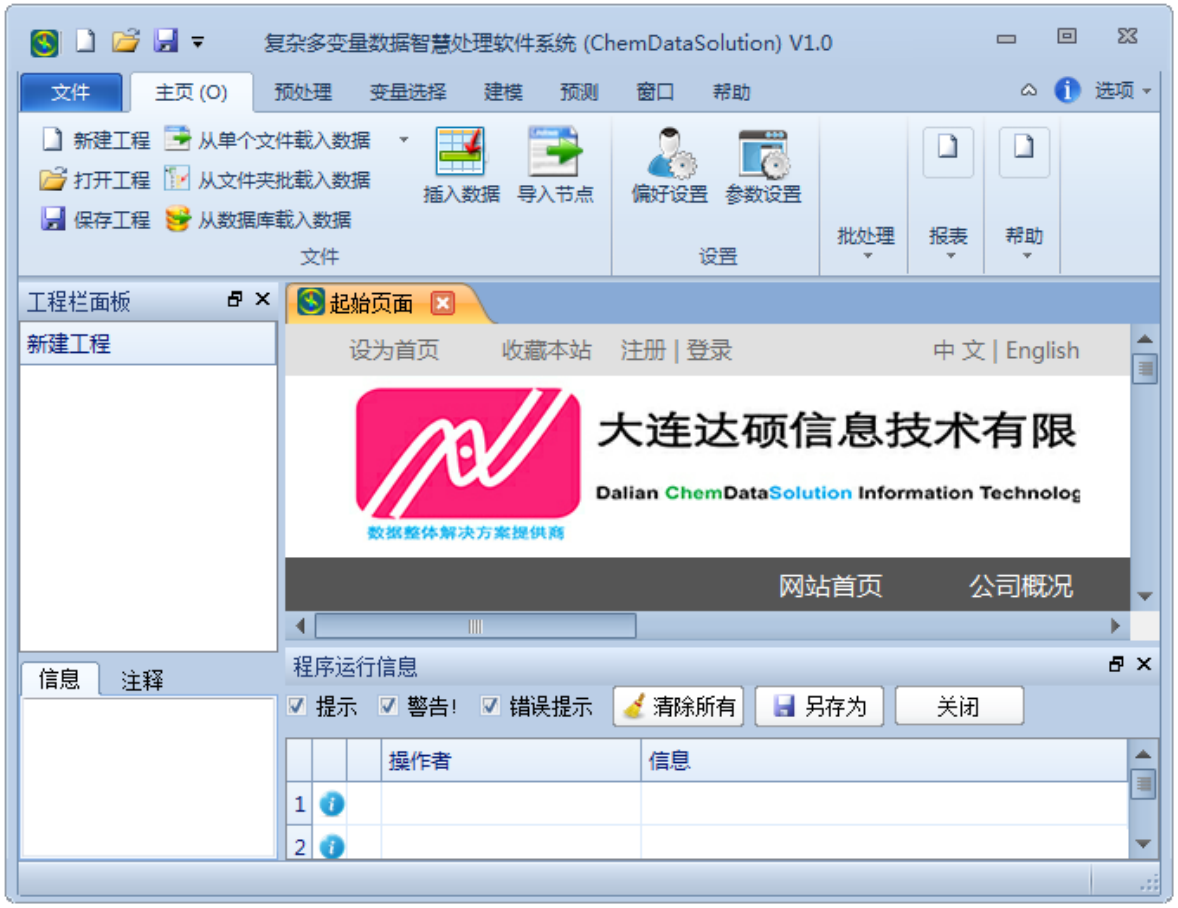
批处理是一个宽泛的概念，在本软件是指批量进行多个处理或操作，主要包括:

- 批载入：指一次载入多个相同或不同类型的数据文件。
- 算法流(批方法)：针对数据处理方法和数据的不同，具有多各方面的功能，详见 2.2.1.，2.2.2.，2.2.3.，以及 8.5.。
- 产生报表：用户从工程导航栏中任意添加并排列多个项目，输出报表。

第四章　用户界面

本章介绍系统的主界面及功能，以便用户快速了解 ChemDataSolution 整体框架与使用。

用户安装完成系统后(具体按照方法请参阅)，双击名为“ChemDataSolution.exe”的文件图标  即可打开软件，出现如下图所示的用户主界面。



在上图中：

1) 中间位置为起始页面，内容为我公司网站，用户点击相关链接，即可直接在浏览器中打开对应的内容；

2) 顶端为系统主要菜单，以 Ribbon 样式呈现，集合软件数据处理的完整功能；

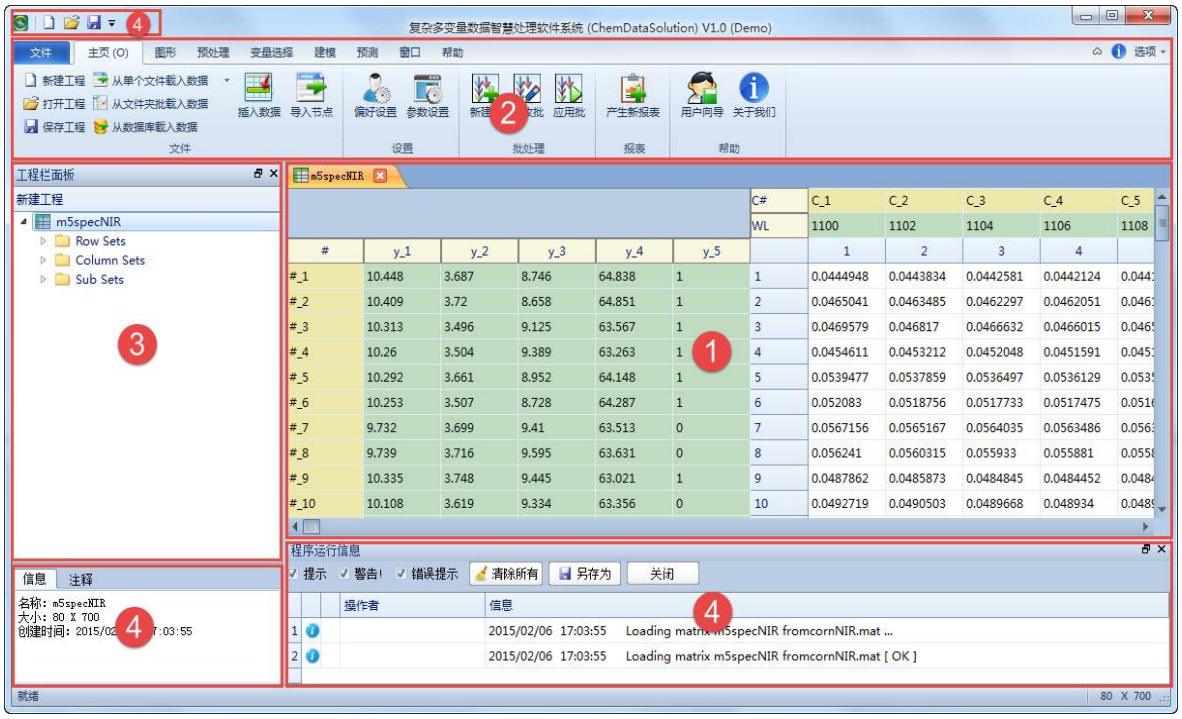
3) 左侧部分包括到导航栏和节点信息的显示及标注，导航栏汇集用户进行数据处理

操作后的各种结果；

4) 其他部分则为辅助功能区。

为方便介绍，系统主界面被划分为如下图所示的四个部分(①-④)，即：

- 主窗口
- 功能菜单区
- 工程导航栏区
- 其他辅助功能区



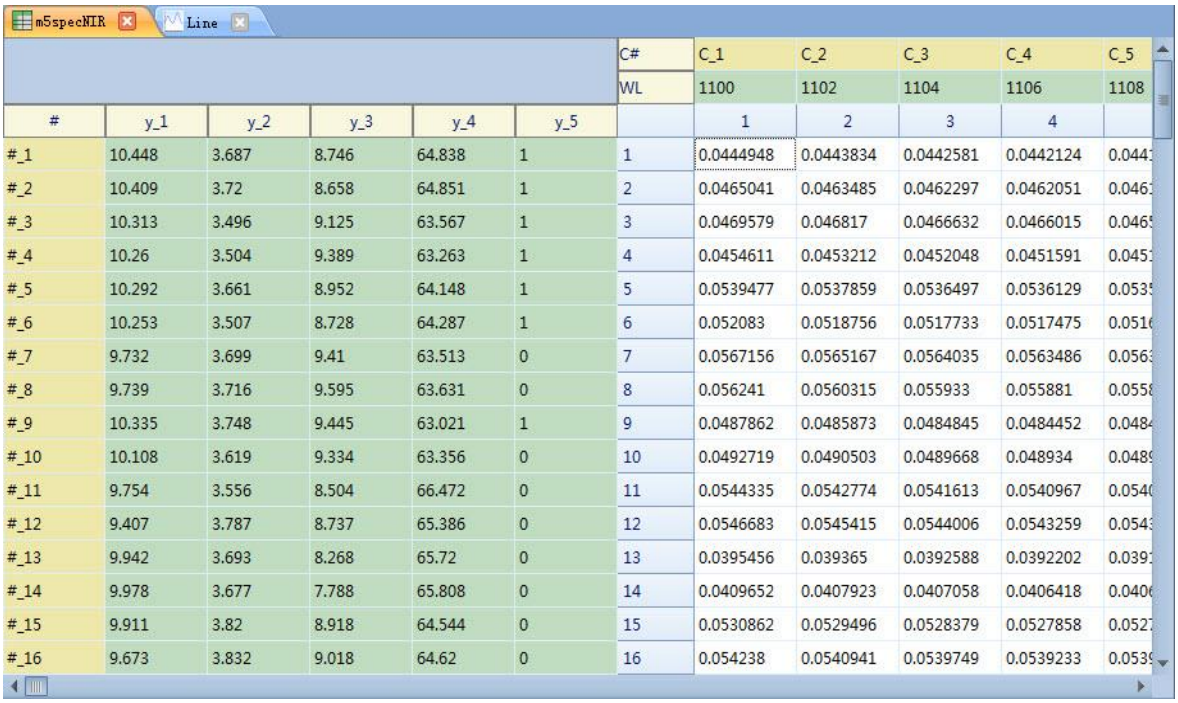
本章不包括实际操作的介绍，具体操作步骤和内容详见其后的各相关章节。

4.1. 主窗口

软件的主窗口包括二种模式，程序可根据实际情形自动转换:

1) 数据编辑模式：呈现数据集的视图表，如下图所示。

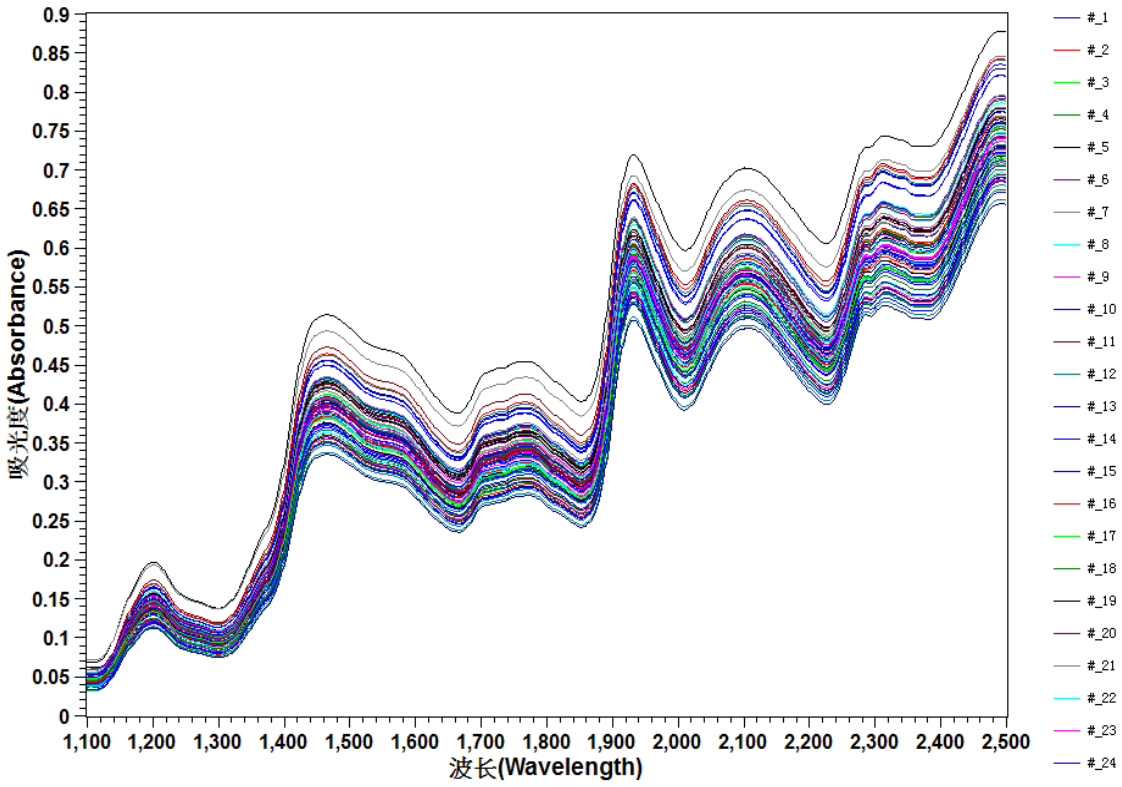
- 新建工程并载入数据后，将以此模式呈现被载入的数据；用户对数据的操作，以及分析处理后的数据结果与模型图表结果等均将以此模式呈现。
- 数据编辑模式下的主窗口，亦可被划分为多个不同部分，包括：数据矩阵(自变量 X)，响应向量或矩阵(因变量 y)，以及二者对应的行/列属性描述，具体内容请参见第六章。
- 建模后得到的模型图表结果，提供右键菜单功能，用户可操作实现数据/图形转换，具体内容请参见第十二章。
- 数据编辑模式下，可对数据进行 20 种不同操作，例如查找，排序，创建子数据等，具体内容请参见第六章。



						C#	C_1	C_2	C_3	C_4	C_5
						WL	1100	1102	1104	1106	1108
#	y_1	y_2	y_3	y_4	y_5		1	2	3	4	
#_1	10.448	3.687	8.746	64.838	1	1	0.0444948	0.0443834	0.0442581	0.0442124	0.044
#_2	10.409	3.72	8.658	64.851	1	2	0.0465041	0.0463485	0.0462297	0.0462051	0.046
#_3	10.313	3.496	9.125	63.567	1	3	0.0469579	0.046817	0.0466632	0.0466015	0.046
#_4	10.26	3.504	9.389	63.263	1	4	0.0454611	0.0453212	0.0452048	0.0451591	0.045
#_5	10.292	3.661	8.952	64.148	1	5	0.0539477	0.0537859	0.0536497	0.0536129	0.053
#_6	10.253	3.507	8.728	64.287	1	6	0.052083	0.0518756	0.0517733	0.0517475	0.051
#_7	9.732	3.699	9.41	63.513	0	7	0.0567156	0.0565167	0.0564035	0.0563486	0.056
#_8	9.739	3.716	9.595	63.631	0	8	0.056241	0.0560315	0.055933	0.055881	0.055
#_9	10.335	3.748	9.445	63.021	1	9	0.0487862	0.0485873	0.0484845	0.0484452	0.048
#_10	10.108	3.619	9.334	63.356	0	10	0.0492719	0.0490503	0.0489668	0.048934	0.048
#_11	9.754	3.556	8.504	66.472	0	11	0.0544335	0.0542774	0.0541613	0.0540967	0.054
#_12	9.407	3.787	8.737	65.386	0	12	0.0546683	0.0545415	0.0544006	0.0543259	0.054
#_13	9.942	3.693	8.268	65.72	0	13	0.0395456	0.039365	0.0392588	0.0392202	0.039
#_14	9.978	3.677	7.788	65.808	0	14	0.0409652	0.0407923	0.0407058	0.0406418	0.040
#_15	9.911	3.82	8.918	64.544	0	15	0.0530862	0.0529496	0.0528379	0.0527858	0.052
#_16	9.673	3.832	9.018	64.62	0	16	0.054238	0.0540941	0.0539749	0.0539233	0.053

2) 图形查看模式：呈现数据的可视化图形。

用户任意选择某一数据表格的一部分或全部，再选择图形菜单中的一种绘图方式，即可绘制用户需要的图形，如下图所示。



图形查看模式下可实现图形的缩放和标注，以及图形中样本/变量的选择和标记等，详情请参见 9.4.。

4.2. 功能菜单区

功能菜单区实现系统的主要功能，一级菜单/标签名称与功能概括见下表:

序号	一级菜单	主要功能简介
1	文件	新建、打开、保存、另存工程，以及最近的工程；打印，打印预览；退出等菜单项。
2	主页	载入数据、系统设置、批处理、报表、用户向导等，涵盖完成数据处理的完整流程。
3	图形	1) 数据编辑模式下，显示可绘制的图形类别。

		2) 图形查看模式下，显示可对图形进行的操作。
4	预处理	参见 2.1.7.所介绍的内容与功能。
5	变量选择	参见 2.1.8.所介绍的内容与功能。
6	建模	参见 2.1.9.，2.1.10.，以及 2.1.11.所介绍的内容与功能。
7	预测	参见 2.1.12.所介绍的内容与功能。
8	窗口	视图窗口操作。
9	帮助	参见 2.1.13.所介绍的内容与功能。
10	选项	软件整体样式风格、文字的设置。

以下对功能菜单区中一级菜单分别做简要介绍，详细内容请参见各相关章节。

4.2.1. 文件

点击文件菜单后，显示如下图所示的内容:

文件菜单中各部分的详细功能，见下表:

序号	功能名称	图标	快捷键	功能描述
1	新建工程		Ctrl + N	新建一个空白工程。
2	打开工程		Ctrl + O	打开已有工程。
3	保存工程		Ctrl + S	保存当前工程。
4	工程另存为			将当前工程另存到一个新的文件路径位置，或另存为一个新的文件名。
5	关闭工程			关闭当前工程。若未保存对当前工程文件的修改，则关闭前程序会自动提示。
6	打印		Ctrl + P	打印当前活动窗口中的内容。
7	打印预览			预览当前活动窗口中内容的打印效果。
8	退出			关闭系统。
9	最近的工程			显示最近打开过的工程文件名称，可直接点击文件，以打开对应的工程。

更多内容详情请参见第七章。

4.2.2. 主页

点击主页菜单后，显示如下图所示的内容:

主页中包含系统的关键功能，可实现基于算法流(批方法)的数据处理全流程。主页中主要提供针对工程，而非具体数据处理的功能。基于不同的功能类型，将主页分为如下 5 个功能组。

- 文件
- 设置
- 批处理
- 报表
- 帮助

上述 5 个功能组的功能名称及描述见下表:

序号	功能组别	功能名称	功能描述
1	文件	新建工程	新建一个空白工程。
		打开工程	打开已有工程。
		保存工程	保存当前工程。
		从单个文件载入数据	导入单个文件数据到工程中。
		从文件夹批载入数据	批量导入文件夹中的数据到工程中。

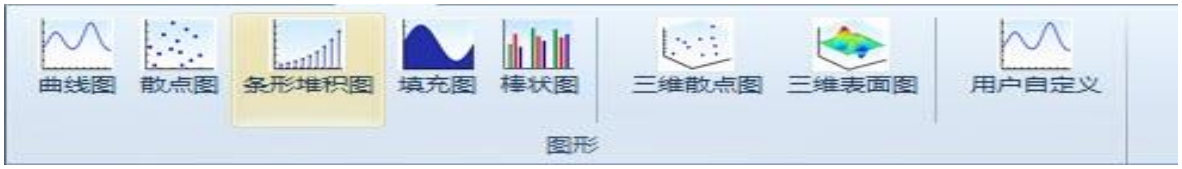
		从数据库载入数据	导入数据库中的数据到工程中。
		插入数据	插入一个新的数据到工程中。
		导入节点	从已有工程中导入某个节点到当前工程。
2	设置	偏好设置	设置用户使用软件的偏好。
		参数设置	统一设置数据处理方法的参数。
3	批处理	新建批	创建一个新的算法流(批方法)流程。
		修改批	修改一个已经创建好的算法流(批方法)流程。
		应用批	往算法流(批方法)中添加数据，执行数据处理流程。
4	报表	产生新报表	创建一个新的报表。
5	帮助	用户向导	帮助用户更快使用本软件。
		关于我们	访问关于本软件和本公司的信息。

4.2.3. 图形

图形标签是关联标签，当选中相应的项时图形标签才会显示。图形标签中的命令选项会根据所选中内容的不同而适时变化。如前所述，当选中工程导航栏中的数据项时，图形标签中的命令选项是选择绘图方式，如曲线图，散点图等；当选中的是图形项时，图形标签中的命令选项是各种图形工具，如加标注，放大/缩小等。

4.2.3.1. 绘图方式

在数据编辑模式下点击图形菜单后，显示如下图所示的内容:



选中工程导航栏中的数据节点，主窗口即进入数据编辑模式，此时便可访问绘图功能，选择其中的全部或部分数据，可绘制图形。本软件提供 8 种数据绘图方式如下表所示:

序号	绘图类型	图标	说明
1	曲线图		以曲线形式显示一个向量或矩阵，并可定义绘图优先性(行或列优先)，以及图形坐标。
2	散点图		类似于曲线图，但以不连续点的形式表达数据。 绘图时可选择内部作图或外部作图，并可定义绘图优先性(行或列优先)，以及图形坐标。 详细内容请参见 3.8.，3.9.和 3.10.。
3	条形堆积图		以条形堆积的形式显示一个向量或矩阵，可定义绘图优先性，当数据维数大于 1 时，则以不同颜色条形堆积的形式绘出。
4	填充图		颜色填充曲线图下面积后得到的图形。
5	棒状图		以棍棒状形式显示一个向量或矩阵。 类似于条形堆积图，差异在于线形更粗，当数据维数大于 1 时，不同线条并不重叠堆积在一起，而是相邻并排，以不同颜色标示。
6	三维散点图		类似于二维散点图。在二维的基础上，增加一个数据维度，在立体空间中显示数据点。

7	三维表面图		以颜色深浅表示数据值大小，相邻数据间以插值形式平滑过渡，构造覆盖整个数据区域表面的立体图形。
8	用户自定义		用户可同时选择多个矩阵数据，并根据所定义的绘图方式一并绘制出来。

4.2.3.2. 图形工具

在图形查看模式下，程序会自动跳转到图形工具，显示如下图所示的内容，此时可使用图形工具可对当前被打开的图形进行各种操作。

有关图形工具的更多功能详情请参见 9.4.。

4.2.4. 预处理

点击预处理菜单后，显示如下图所示的内容:

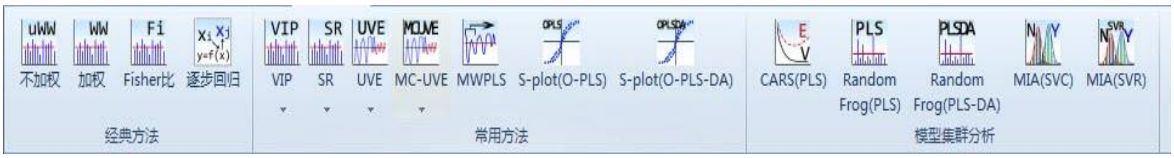
本软件完整包括主要数据预处理的功能，提高数据的可用性与数据质量，从而在后续的变量选择，建模与模型应用中得到更稳健可靠的结果。亦包括数据的运算功能，即原始数据样本或变量的变换，以及基于原始数据的运算产生新的样本或变量。

上述预处理方法特别适合色谱、质谱和光谱等科学仪器数据的分析。

预处理方法的具体内容介绍，请参考预处理章节。

4.2.5. 变量选择

点击变量选择菜单后，显示如下图所示的内容:



本软件同时涵盖可用于分类和回归建模的变量选择方法；每类方法包括迄今为止的经典重要方法，化学数据分析中较常用的方法，以及最新的方法等，可满足用户的广泛需求，并可基于本软件提供的不同方法处理用户数据，得到综合比较结果。

各变量选择方法的具体内容，请参考变量选择章节。

4.2.6. 建模

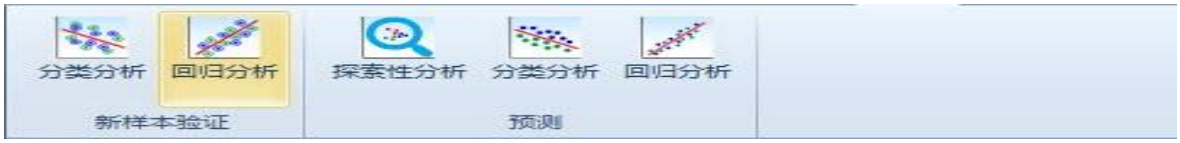
点击建模菜单后，显示如下图所示的内容:

本软件可同时进行探索性分析，以及分类与回归建模，并包含针对二类以上问题的分类方法和二个以上属性的回归方法，同时亦包括非线性建模方法。基于此，本软件可解决绝大部分化学与生物分析中的数据处理问题。

各建模方法的具体内容，请参考建模章节。

4.2.7. 预测

点击预测菜单后，显示如下图所示的内容:



基于 4.2.6.中方法所建立的模型，可对新的数据样本进行验证，以及对未知的数据样本进行预测；基于所用模型的不同，可分别处理分类与回归问题的预测。

各验证、预测方法与流程的具体内容，请参考预测章节。

4.2.8. 窗口

点击窗口菜单后，显示如下图所示的内容:

窗口菜单功能分为两组，一组是对工作区中的各标签页的窗口管理(如平铺或重叠)，另一组是对程序主窗口的视图管理，主要包括工程栏和程序运行信息的显示与隐藏。窗口管理提供以下功能:

序号	功能名称	图标	功能描述
1	平铺窗口		将工作区内已打开的窗口平铺显示。
2	层叠窗口		将工作区内已打开的窗口层叠显示。
3	上一个活动窗口		显示当前窗口的上一个活动窗口。

4	下一个活动窗口		显示当前窗口的下一个活动窗口。
5	关闭所有窗口		关闭工作区内所有已打开窗口。
6	关闭当前窗口		关闭工作区内当前显示窗口。
7	关闭其它窗口		关闭工作区内除当前显示窗口外的所有窗口。
8	关闭左侧窗口		关闭工作区内当前显示窗口的左侧窗口。
9	关闭右侧窗口		关闭工作区内当前显示窗口的右侧窗口。
10	切换窗口		工作区内已打开的窗口间快速切换。

视图管理提供以下功能:

序号	功能名称	图标	功能描述
1	工程栏		显示/隐藏用户界面左侧的工程导航栏。
2	程序运行信息		显示/隐藏程序运行信息窗口。

视图管理中的功能设为开/关双控键，点击图标可在显示与隐藏之间切换。

4.2.9. 帮助

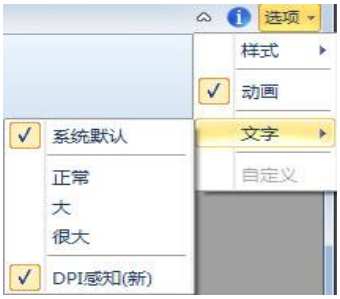
点击帮助菜单后，显示如下图所示的内容:

帮助菜单主要包括本软件的授权和帮助信息，以及登录访问本公司网站等。

序号	功能名称	图标	功能描述
1	修改授权		修改用户使用本软件的授权。
2	更新		在线更新本软件的功能或方法等。
3	版权		显示本软件的版权相关信息。
4	使用帮助		以 CHM 文件的形式，显示本软件的完整使用帮助。
5	用户向导		以简洁导向的形式，帮助用户快速入门，使用本软件。
6	关于我们		访问/获取本软件和本公司信息。

4.2.10. 选项

选项菜单提供对软件工作环境的设置功能，用户可对软件的窗体样式、动画、字体进行设置，同时用户亦可自定义功能区样式，如下图所示。



本软件采用 Ribbon 菜单模式，选项主要用于设置该类菜单样式，用户可根据个人喜好修改。

4.3. 工程导航栏区

如前所述，工程导航栏是本软件的中枢系统，以节点文件夹和节点的形式，集中、清晰、有序管理软件使用过程中所涉及的数据，可视化图形，算法流，数据处理中间结果，模型结果(表格和图形)，类似于文件夹与文件的管理形式。本小节介绍工程导航栏的基本功能与应用，主要内容包括:

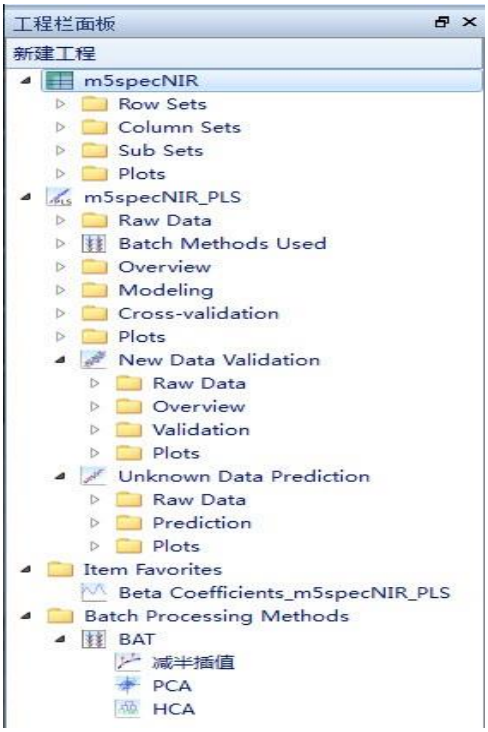
- 什么是工程导航栏
- 工程导航栏中的节点文件夹和节点
- 对节点的管理

4.3.1. 什么是工程导航栏

工程导航栏将工程中的内容(数据，图形，结果等)以树状结构的形式，分不同级别显示；最低层级即为具体的内容，称为节点，单击某具体借点后，相关内容将在主窗口中显示；节点的上一级或多级则为节点文件夹，单击文件左侧三角形  ，或者双击节点文件夹，即可打开。

算法流节点直接显示，不再出现在主窗口中，但算法流中各方法的详细信息和参数则显示在左侧下方的信息栏中，并可通过注释功能添加用户标注。

下图为一个典型的工程导航栏内容。从图中可以看出，工程导航栏以树状图的形式，分级别清晰管理不同的节点文件夹，以及每个节点文件夹下对应的具体节点。



4.3.2. 工程导航栏中的节点文件夹和节点

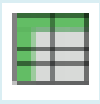
节点是工程导航栏的最小单元，分为数据，图形和算法流等；节点文件夹则分为不同级别，最低一级文件夹直接涵括节点，而高一级的文件夹则包含低一级的文件夹，如前所述，完全类似于文件和文件夹的关系。

一般来说，工程导航栏中可能出现的节点文件夹及节点包括如下几种类型:

序号	节点文件夹/节点类型	内容描述
1	基本数据表	包含样本和变量的数据矩阵。
2	行划分数据	包含所有变量，部分样本的数据矩阵。
3	列划分数据	包含所有样本，部分变量的数据矩阵。

4	子数据	包含部分样本，部分变量的数据矩阵。
5	图形	包含一种或多种图形。
6	模型	包含建模结果。
7	验证	包含新样本的验证结果。
8	预测	包含未知样本样本的预测结果。
9	算法流	成功新建一个包含多个数据处理方法的算法流后，程序所产生的节点。

节点文件夹和节点的产生及命名规则，如下表所述。用户可通过右键功能修改。

序号	节点文件夹或节点	名称	图标	命名规则
1	导入工程的原始数据(节点)	分二种情形		若导入到已有数据中，则保持名称不变；若导入为新的基本数据表，则采用被导入数据文件名。
2	行划分子数据(节点文件夹)	Row Sets		汇聚各相同类型节点，命名唯一。
3	行划分子数据(节点)	依次编号		第一个节点名称为 Rsets，其后的节点则依次添加序号，比如”Rsets_1”。
4	列划分子数据(节点文件夹)	Column Sets		汇聚各相同类型节点，命名唯一。
5	列划分子数据(节点)	依次编号		第一个节点名称为 Csets，其后的节点则依次添加序号，比如”Csets_1”。

6	其他类型子数据(节点文件夹)	Sub Sets		汇聚各相同类型节点，命名唯一。
7	其他类型子数据(节点)	依次编号		第一个节点名称为 Ssets，其后的节点则依次添加序号，比如”Ssets_1”。
8	可视化图形(节点文件夹)	Plots		汇聚各相同类型节点，命名唯一。
9	可视化图形(节点)	依图形类型不同	依类型而变化	依曲线图，散点图等类型不同而变化；以曲线图为例：第一个节点名称为 Line，其后的节点则依次添加序号，比如”Line_1”。
10	算法流(节点)	Batch Processing Methods		第一个算法流节点名称为 BAT，其后的节点则依次添加序号，比如”BAT_1”。算法流中的不同方法则按序显示在工程栏中。
11	模型(节点文件夹)	综合建模数据与方法		由二部分构成，前一部分为建模时所采用的数据，后一部分则为建模所用的算法流或方法。 最高级的模型结果节点文件夹中，通常再包含多个节点文件夹，如 Raw Data(原始数据)，Batch Methods Used(算法流)，Overview(关键结果汇总)，Modeling(模型表格结果)，Cross-validation(建模中交互检验结果)，Plots(模型图形结果)和 Intermediate Results(算法流中不同方

				法的中间结果)等。
12	模型(节点)	依表格与图形结果而不同	依表格与图形结果而不同	不同建模方法所得节点有差异，但相同建模方法所得结果固定。
13	新样本验证(节点文件夹)	New Data Validation		新样本验证结果作为一个节点文件夹，置于模型文件夹中，包括 Raw Data(原始数据)，Overview(关键结果汇总)，Validation(验证结果)和 Plots(图形)四个文件夹。
14	新样本验证(节点)	依表格与图形结果而不同	依表格与图形结果而不同	根据上述不同节点文件夹而变化。
15	未知样本预测(节点文件夹)	Unknown Data Prediction		未知样本验证结果作为一个节点文件夹，置于模型文件夹中，包括 Raw Data(原始数据)，Prediction(预测结果)和 Plots(图形结果)三个文件夹。
16	未知样本预测(节点)	依表格与图形结果而不同	依表格与图形结果而不同	根据上述不同节点文件夹而变化。
17	其他(节点)	Item Favorites		汇聚用户所有选择性添加的节点，命名唯一。

新建工程后，通过不同方式导入的数据，将数据节点的形式显示在工程导航栏中；

其后对数据的可视化绘图，则以节点文件夹的形式显示为下一级。

数据节点文件夹：如前所述，及其后部分亦将介绍到，本软件同时提供基本数据(从原始数据文件、文件夹或数据库中导入得到)，以及对基本数据进行某种行或列的划分，截取其中的部分或全部数据，以产生新的子数据。这些新产生的子数据将分成三类，分别存放在三个不同的数据节点文件夹中，即行划分，列划分和子数据(非行划分或列划分节点文件夹)。

数据节点：用户对基本数据表操作得到的具体行划分，列划分和子数据，则显示在对应的数据节点文件夹下。

用户构造的算法流，亦以节点的形式显示在工程导航栏中。用户构造模型时，首先产生一个节点文件夹，模型中的具体内容，拷贝执行不同方法得到的中间结果，表格，图形等则以具体节点的形式，显示在该节点文件夹内。新样本验证与未知样本预测中所产生的节点文件夹及节点，插入在所应用的模型节点文件夹下。

若使用算法流分析数据，且该算法流中包含多个建模方法，则模型结果以不同节点文件夹的形式分别显示，即建模所分析的数据和所使用的算法流相同，模型结果亦单独显示在不同模型节点文件夹下。

Item Favorites，即我的收藏节点，汇集用户随时加入的重要内容，可以是数据，图形或模型结果，方便快速产看关键内容，尤其是用户可以在数据处理的过程中即时、逐步添加，尤其对于需要添加到报表中的内容，用户可添加到该文件夹，使用时可快速便捷选择。

除算法流之外，节点文件夹和节点名称以英文为主，用户可自行修改。

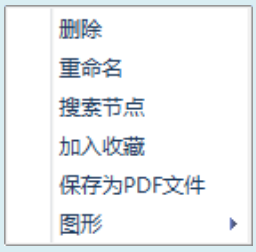
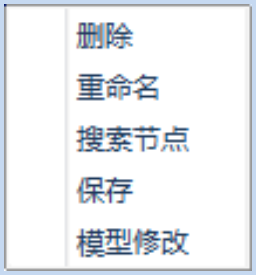
4.3.3. 节点文件夹与节点的操作

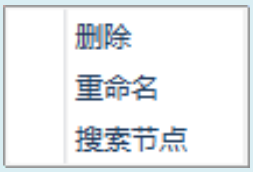
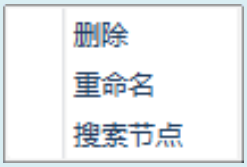
工程导航栏中节点文件夹和节点均具有完整的右键菜单功能，通过这些功能，可非常方便地实现对文件夹和节点的操作及管理。

除模型以外的节点文件夹，均仅具有重命名，删除和搜索三个功能。详情请参见第五章。

模型节点文件夹以及具体节点，其右键菜单功能与类型具有关联性，即不同节点文件夹及节点类型，右键菜单功能亦不尽相同，详细描述见下表。

序号	节点文件夹或节点	右键菜单功能	说明
1	基本数据节点	删除 重命名 搜索节点 保存 加入收藏 预处理 变量选择 建模 图形 添加数据 检查数据合法性 复制数据 保存为Txt文件 添加到数据库	数据是数据处理的基础和根本。因此针对基本数据节点的右键菜单功能非常丰富、全面。主要包括如下几个大类： 1) 节点的操作和管理。 2) 数据的直接分析处理和建模。 3) 数据的基本处理，如数据的检查与保存。
2	行划分数据节点	删除 重命名 搜索节点 加入收藏 预处理 变量选择 建模 图形 检查数据合法性 复制数据 保存为Txt文件 添加到数据库	与基本数据节点雷同。

3	列划分数据节点		与基本数据节点雷同。
4	子数据节点		与基本数据节点雷同。
5	图形节点		针对图形节点的功能，则包括对节点的管理，以及图形的类型转换和图形保存等。
6	模型节点文件夹		针对模型节点的功能，最重要的在于模型的修改，这也是本软件的重要体验之一。

7	验证节点		包括对节点的基本操作功能。
8	预测节点		包括对节点的基本操作功能。
9	算法流(批方法)节点		包括对节点的基本操作功能。
10	节点文件夹(模型节点文件夹除外)		包括对节点的基本操作功能。 系统固定此节点名称，不支持重命名，以免给用户造成困扰。

实因系统在使用过程中将产生大量数据，图形和模型结果等，因而本软件提供便捷的搜索功能，且可选择性对某一具体节点文件夹或者整个工程导航栏进行搜索，方便用户获得快速查看结果，详情请参见 5.3.。

更多详细内容请查看节点管理章节。

4.4. 其他辅助功能区

4.4.1. 快速访问工具栏

快速访问工具栏默认位于功能菜单区上方，如下图所示，用户可根据使用个人习惯设置常用命令的快捷方式。

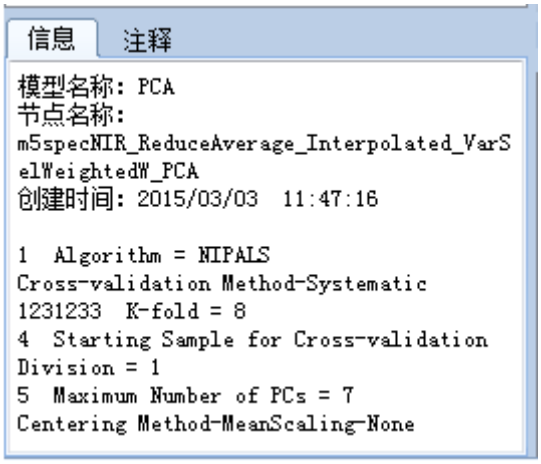
4.4.2. 自定义快速访问工具栏

点击快速访问工具栏右侧的下三角按钮，用户即可对快速访问工具栏进行如下图所示的自定义设置。

- 添加/删除快速访问工具栏上的功能按钮：添加/删除方法是选中/不选中对应功能的多选框。
- 移动快速访问工具栏到功能区上方/下方。
- 最小化功能菜单区。

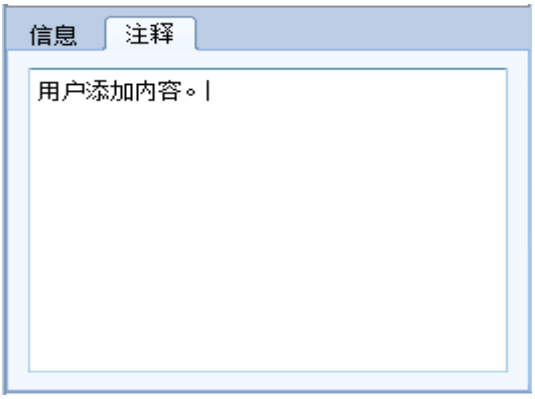
4.4.3. 信息显示区

信息显示区的内容，跟随用户选择的节点文件夹或节点的不同动态变化。以一个基于 PCA 的建模为例，信息显示区如下图所示，包括针对 PCA 分析方法的各参数内容。



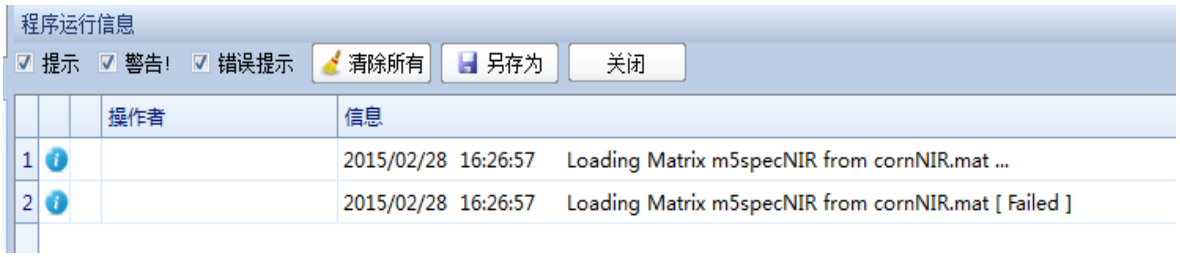
4.4.4. 添加注释区

添加注释区是信息显示区的有效补充，方便用户添加信息显示区内不足的内容，以及用户需要标注的内容，方便日后查看和记忆等，如下图所示。



4.4.5. 程序运行信息显示区

该区域显示程序运行过程中的信息，或者对系统进行某种操作后的说明性信息，如下图所示。



不同程序运行信息以分类的形式，分别标记，且可清除或者保存。根据实际情形，信息以中文或英文的形式显示。

第五章　节点文件夹与节点的管理

事实上，尽管此章节前已部分介绍节点文件夹及节点(如 4.3.3.)，但并未提及任何产生节点文件夹与节点的方法和过程。实因保证本实用手册可读性的要求，在第四章的基础上，随即介绍对节点文件夹和节点管理的内容。

节点文件夹和节点的管理通过其右键菜单功能来实现。不同类型的节点具有不同的操作命令(右键菜单的对应关系请查看 4.3.3.)。如前所述，节点的右键菜单功能包括如下三种情形:

总结上图中节点文件夹和节点的功能，可概括如下表。

序号	操作名称	功能描述	说明
1	删除	删除节点文件夹或节点，操作具有不可逆性，须谨慎使用。	无。
2	重命名	为节点文件夹或节点赋予一个新的名字，新名字不得与该节点	无。

		文件夹或节点中同级目录下的其他名字相同。	
3	搜索节点	在整个工程中搜索含有用户自定义字符的节点文件夹或节点，搜索成功后跳转并选中结果。	在搜索界面，用户可选择仅存在于某子节点文件夹或节点中的信息。
4	保存	仅保存此节点文件夹或节点到工程文件中。	支持基本数据和模型节点文件夹或节点。
5	加入收藏	将用户感兴趣的节点或重要节点加入到收藏节点文件夹中。	仅支持数据节点和图形节点。
6	预处理方法选择	以用户被选节点作为数据来源，任意选择预处理方法处理数据。	仅支持数据节点。
7	变量选择	以用户被选节点作为数据来源，任意选择变量选择方法处理数据。	仅支持数据节点。
8	建模	以用户被选节点作为数据来源，任意选择建模方法处理数据。	仅支持数据节点。
9	图形	1) 对数据节点，此操作的意义是：以当前节点作为数据绘出不同图形，如曲线图，散点图，条形堆积图等； 2) 对图形节点，此操作的意义是：以当前节点所使用的数据，重新绘出不同的图形，如曲线图，散点图，条形堆	仅支持数据和图形节点，但对两类节点的意义不同。

		积图等。	
10	添加数据	添加数据到当前节点。	仅支持基本数据节点，功能类同于从文件夹批量载入数据时，选择载入数据到已有的数据中。
11	检查数据合法性	检查数据是否包含非法数据(如 nan 等不可被处理的内容)，方便用户快速修改或替换数据。	仅支持数据节点。
12	复制数据	复制当前数据节点的内容为一个新的数据节点，即产生一个相同的节点，方便用户保存或重新处理数据。	仅支持数据节点。
13	保存为 txt 文件	将当前节点内容另存为 txt 文件。	仅支持数据节点。
14	添加到数据库	将当前节点内容添加到数据库中。	仅支持数据节点。
15	保存为 PDF 文件	将当前图形另存为 PDF 文件。	仅支持图形节点。
16	模型修改	通过修改数据或建模方法，重新产生新的模型。	仅支持模型节点。

数据节点包括基本数据表，以及在此基础上所得到的行划分、列划分、子数据节点和数据经过某些数据处理过程后得到的新数据。

接下来依次介绍表中所述的节点文件夹和节点操作功能。

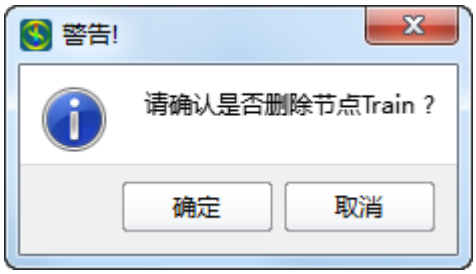
5.1. 删除

删除功能与普通的文件操作无异。

操作步骤:

步骤 1: 选中节点文件夹或节点，点击右键。

步骤 2: 右键菜单中单击**删除**菜单项，此时弹出下图所示的对话框:



步骤 3: 选择确定，删除该节点；选择取消，则取消删除操作。

删除节点文件夹或节点后，该操作具有不可逆性，请谨慎使用。

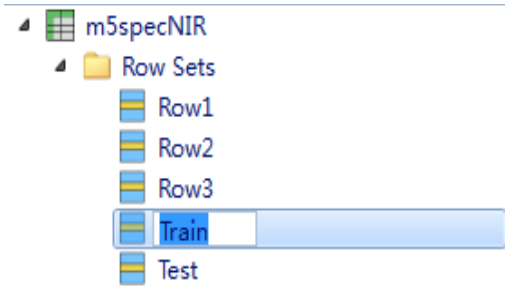
5.2. 重命名

为节点文件夹或节点赋予一个新的名字，新名字不能与该节点同级目录下的其他名字相同。

操作步骤:

步骤 1: 选中节点，点击右键。

步骤 2: 右键菜单中单击**重命名**菜单项，当前节点会变为编辑状态，如下图所示:



步骤 3: 输入用户属意的新名字，若输入的名字与同级目录下的其他名字相同，则会以图的形式提示用户；否则则重命名成功。

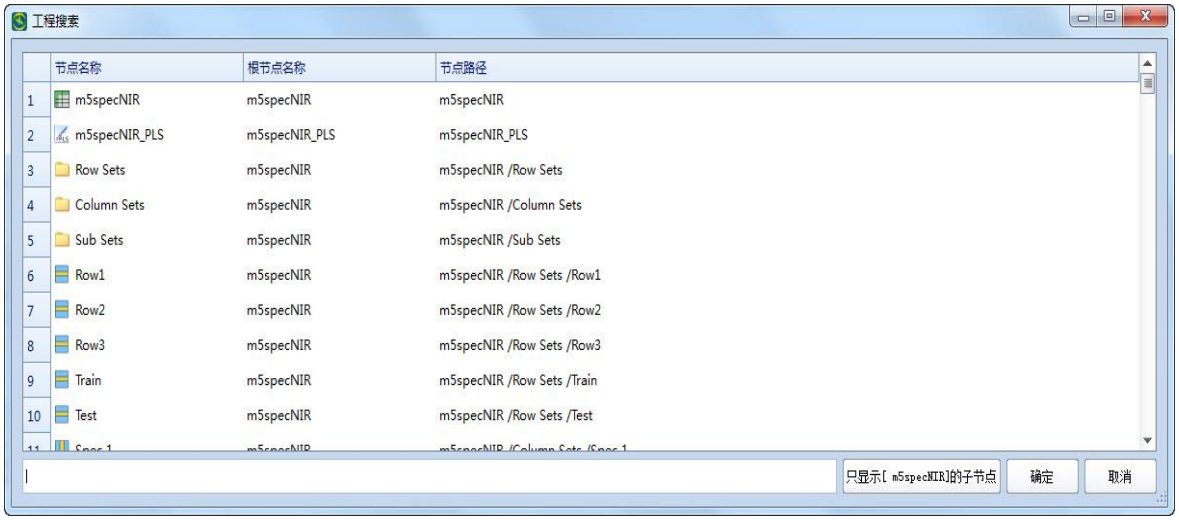
5.3. 节点搜索

节点搜索是本软件提供的个性化功能之一。用户在使用系统的过程中，往往在工程导航栏中产生大量的信息列表，包括数据和图形，特别是模型等。通过节点搜索功能，用户可快速获得含有某一字符的名称，找到需要的结果。

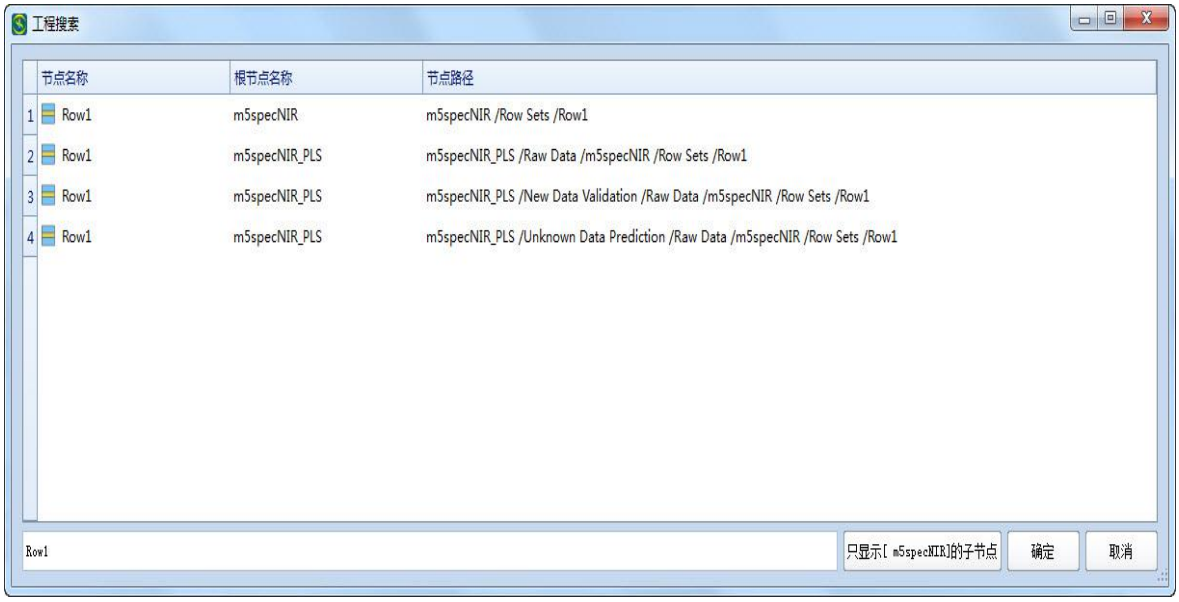
操作步骤:

步骤 1: 选中节点，点击右键。

步骤 2: 右键菜单中单击**搜索节点**菜单项，弹出如下对话框:



步骤 3: 在编辑框中输入要搜索的节点文件夹或节点名称，搜索结果则会随输入内容的改变而动态变化，如下图所示:



用户亦可仅显示某子节点文件夹下所含有的内容，如点击按钮**只显示[m5specNIR]的子节点**，即可过滤掉其他满足搜索条件的节点，以快速提炼结果。

步骤 4: 在搜索结果列表中选中某一项内容，点击**确定**按钮，则可跳转至工程中的对应节点上，方便用户即时查看。点击**取消**按钮，则取消操作，并关闭对话框。

5.4. 保存

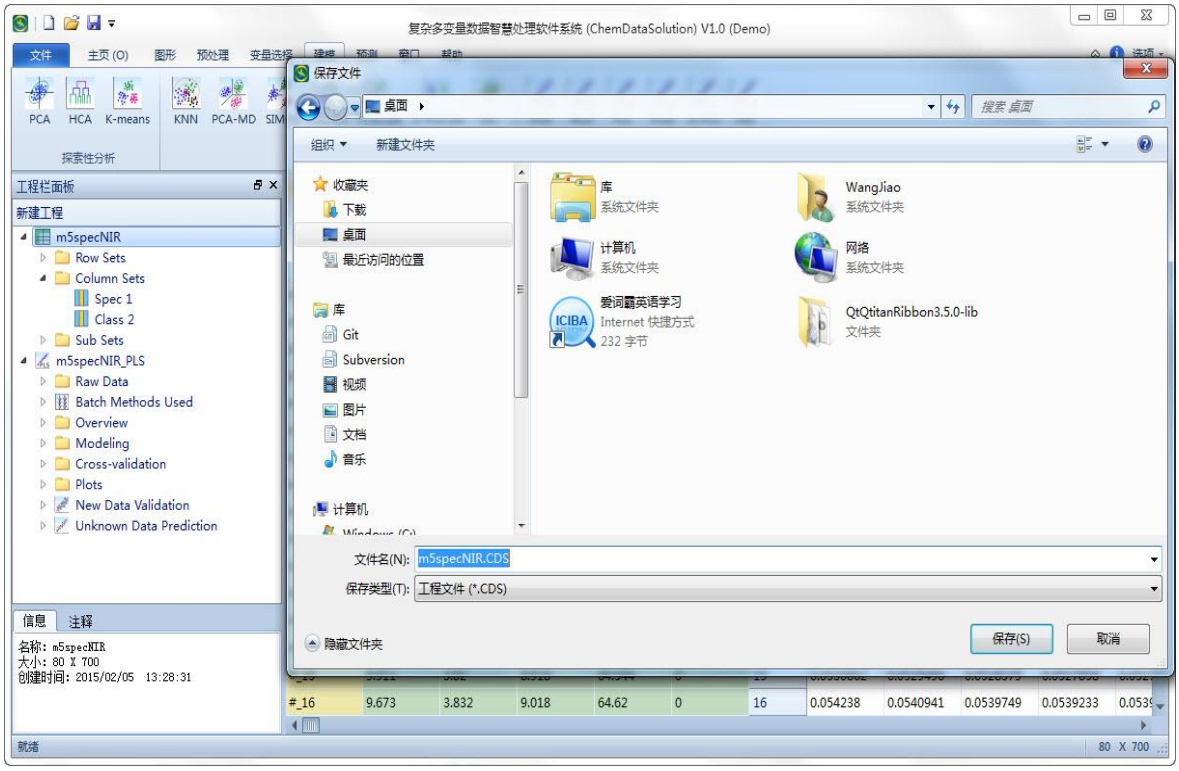
如前所述，本软件以工程形式管理工程导航栏中的数据、图形和模型结果及信息。因此，这些结果和信息的保存必须以工程文件的形式存在。以工程形式保存的结果和信息，用户可以工程文件的形式打开还原。

操作步骤:

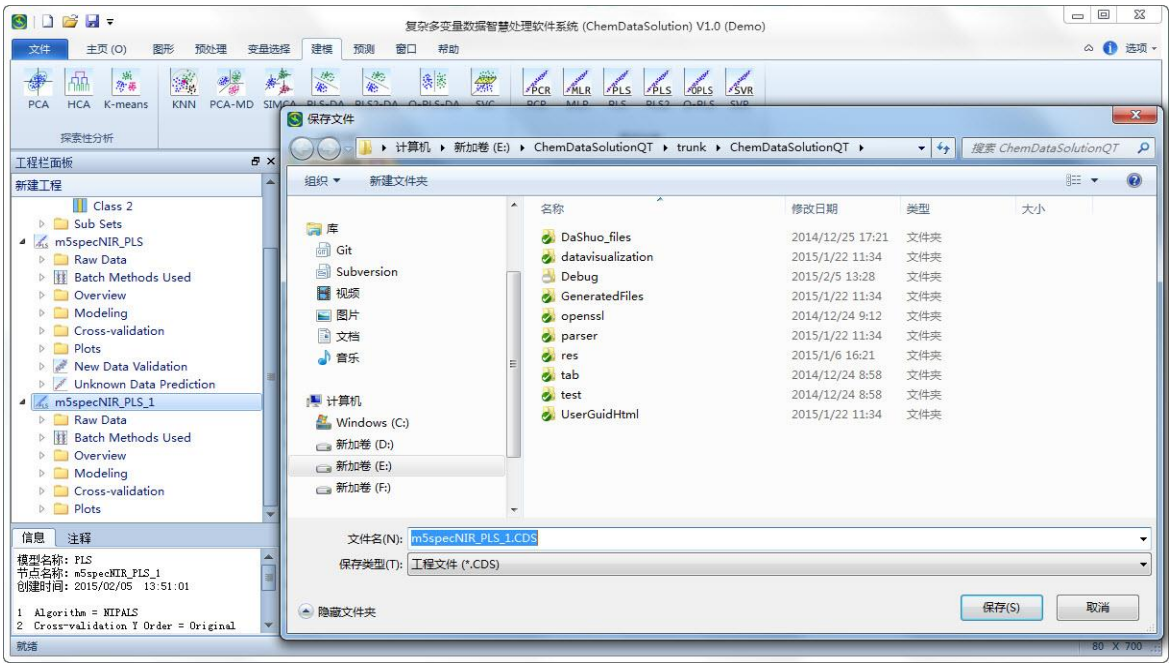
步骤 1: 选中节点文件夹或节点，点击右键。

步骤 2: 右键菜单中单击**保存**菜单项。

若所选节点为基本数据，则弹出如下对话框:

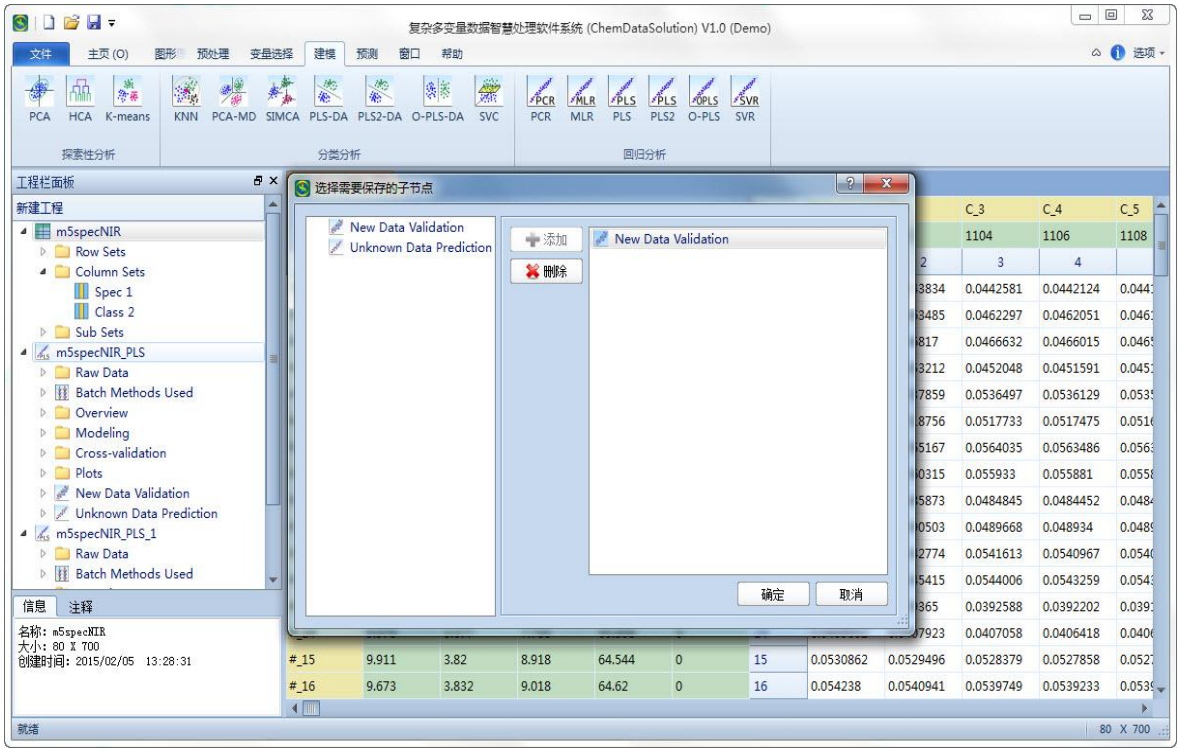


默认文件名为节点的名字，文件后缀为.CDS。若所选节点为模型，且子节点中无验证节点和预测节点，则弹出如下对话框:

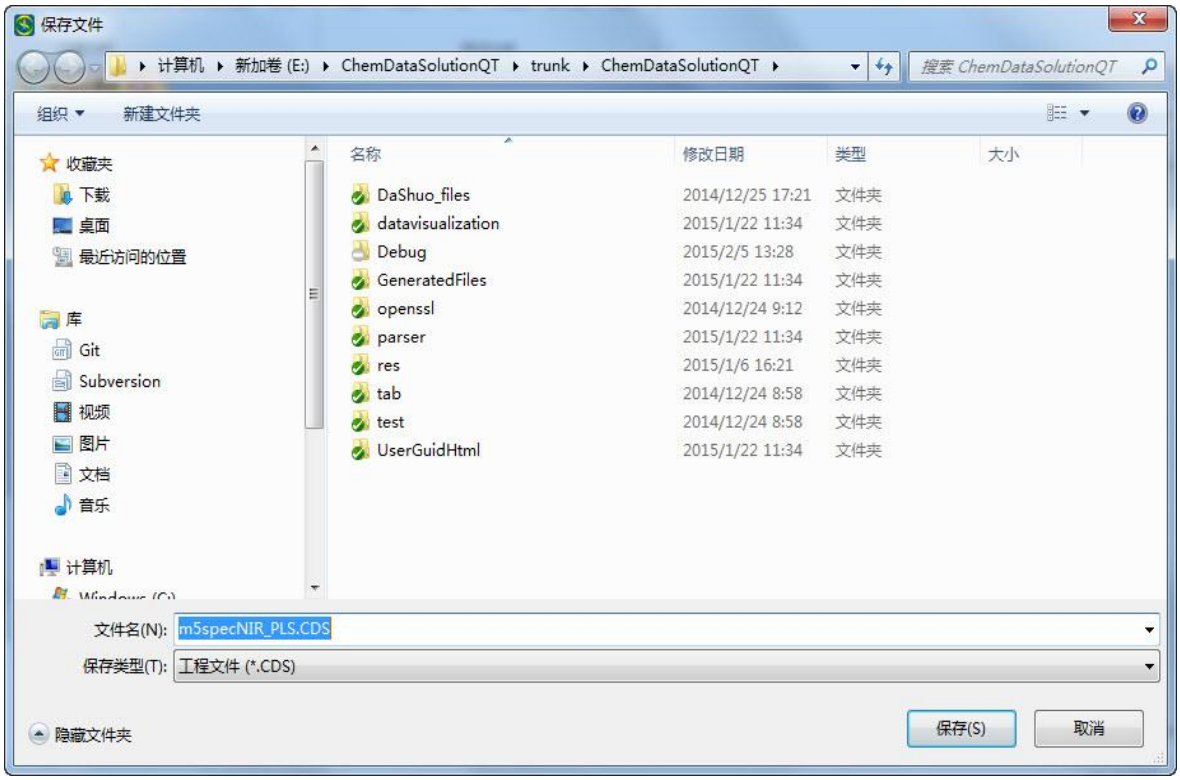


默认文件名为节点的名字，文件后缀为.CDS。若所选模型节点中包含验证节点或预测节点，

则先弹出如下对话框:



在这个对话框中可以选择是否选择，以及选择保存哪些验证或预测子节点，选择完成后点确定按钮，再弹出如下对话框:。默认文件名为节点的名字，后缀亦为.CDS。



步骤 3: 选择保存路径，点击保存按钮，即可保存节点文件夹内容到工程文件中。

此功能仅支持基本数据节点和模型节点。

5.5. 加入收藏

用户将节点加入收藏夹，方便管理和使用。

操作步骤:

步骤 1: 选中节点，点击右键。

步骤 2: 右键菜单中单击**加入收藏**菜单项，便将该节点加入到 Item Favorites 节点中，如下图所示:

收藏夹下的节点不仅包含原节点的名字，亦包含原节点的根节点的名字(原节点是根节点时除外)。任何对收藏夹下的节点的操作(删除功能除外)，亦是对原节点的操作。

收藏节点的二个主要功能是：方便用查看和比较重要表格及图形结果；方便用户产生报表，即将集中于收藏文件夹中的项目，同时快速加入到报表中，而无需从大量的节点文件夹和节点中逐个选择。

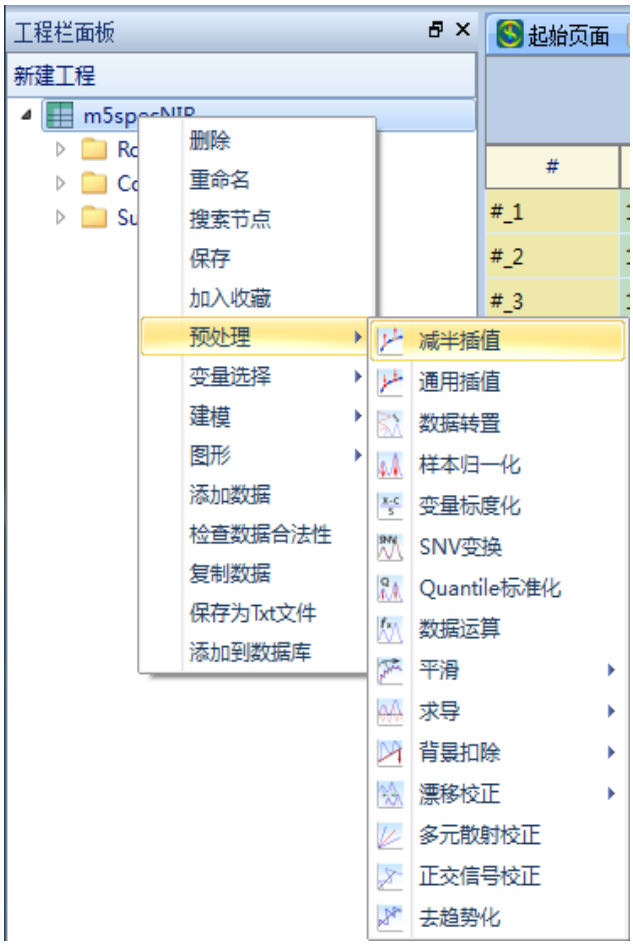
5.6. 预处理

以用户自选节点作为数据来源，采用本软件所提供的各种预处理方法分析数据，仅支持数据节点。

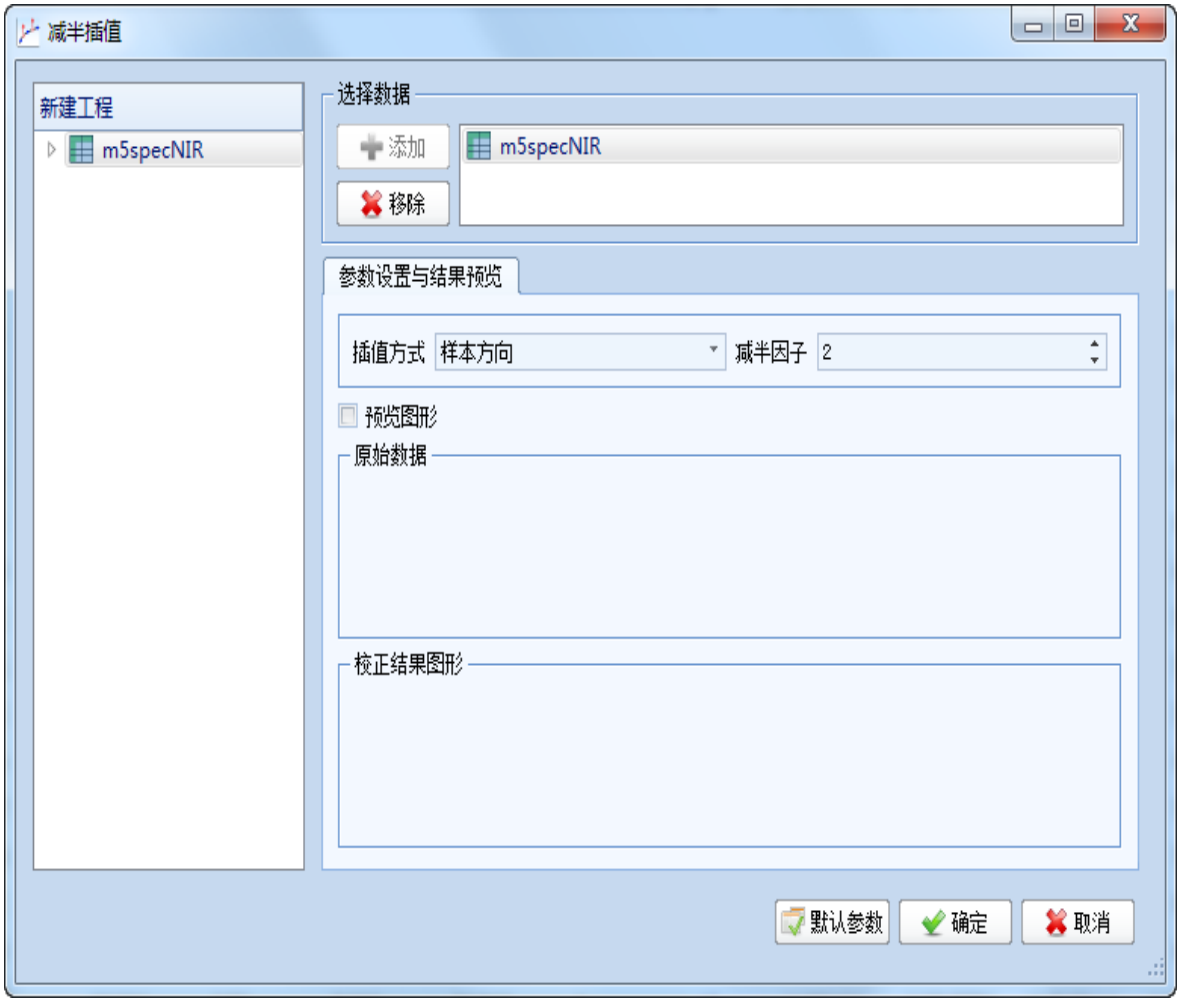
操作步骤:

步骤 1: 选中节点，点击右键。

步骤 2: 右键菜单中单击**预处理方法选择**菜单项，选择一个具体的预处理方法，如下图所示:



单击菜单项，弹出被选的预处理方法对话框，所选节点数据自动加载到数据预处理窗口中，如下图所示(以减半插值为例):



步骤 3: 接下来的操作及更多预处理方法对话框请参考**预处理**章节。

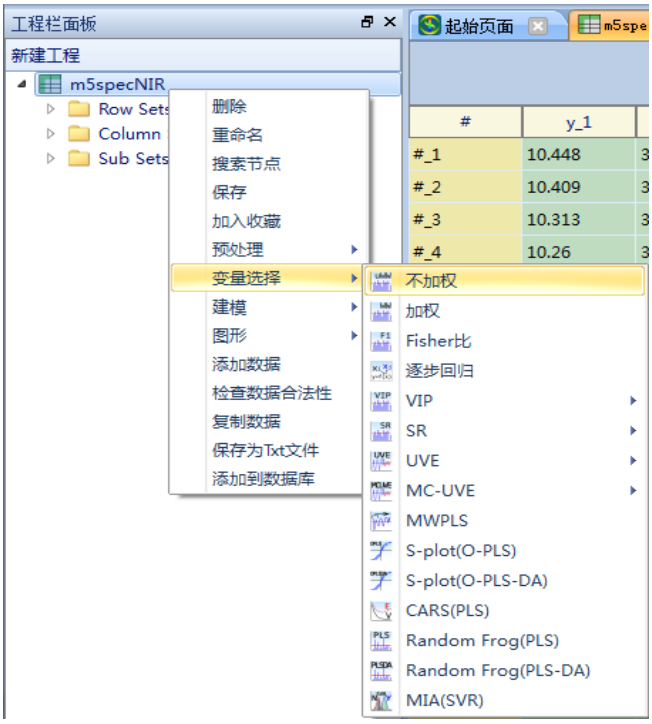
5.7. 变量选择

以用户自选节点作为数据来源，运用本软件提供的各种变量选择方法分析数据，仅支持数据节点。

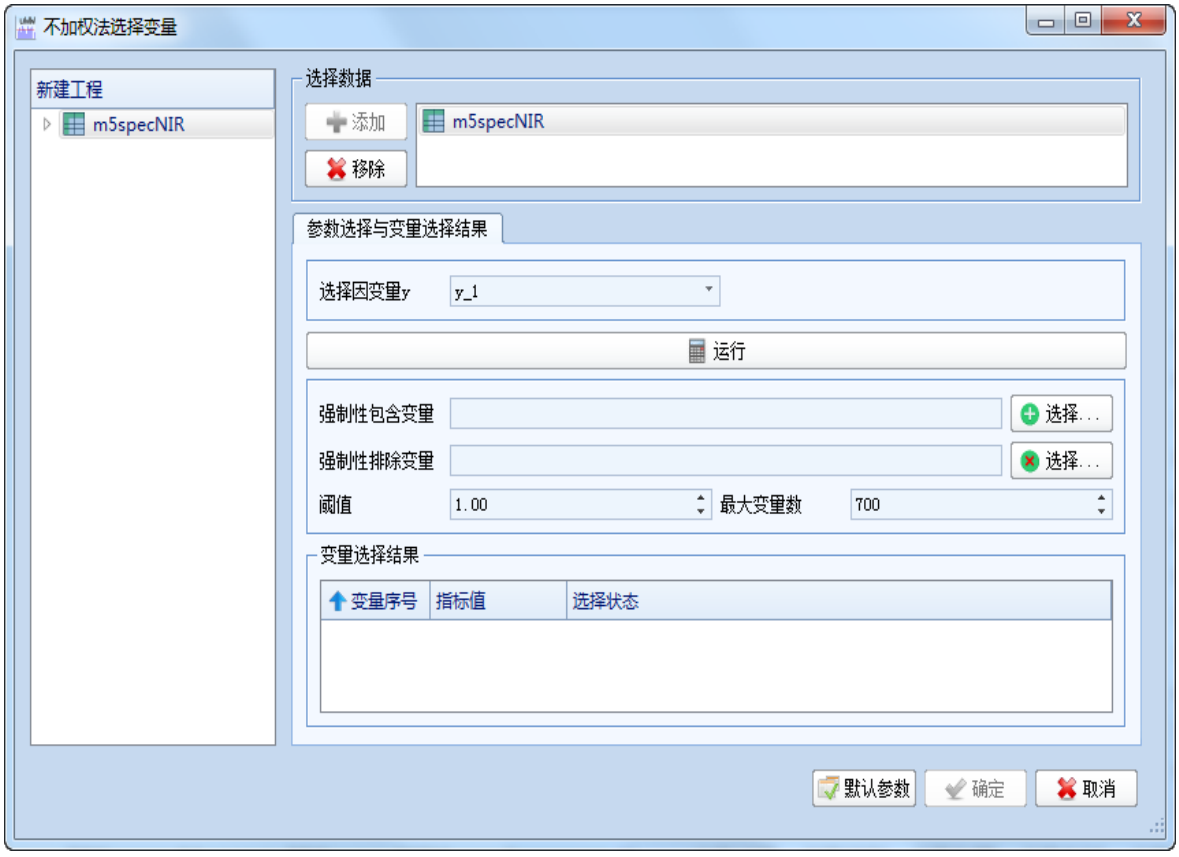
操作步骤:

步骤 1: 选中节点，点击右键。

步骤 2: 右键菜单中单击**变量选择**菜单项，再选择一个具体的变量选择方法，如下图所示:



单击菜单项，弹出被选变量选择方法对话框，所选节点数据自动加载到变量选择窗口中，如下图所示(以不加权方法为例)：



步骤 3: 接下来的操作及更多变量选择方法对话框请参考**变量选择**章节。

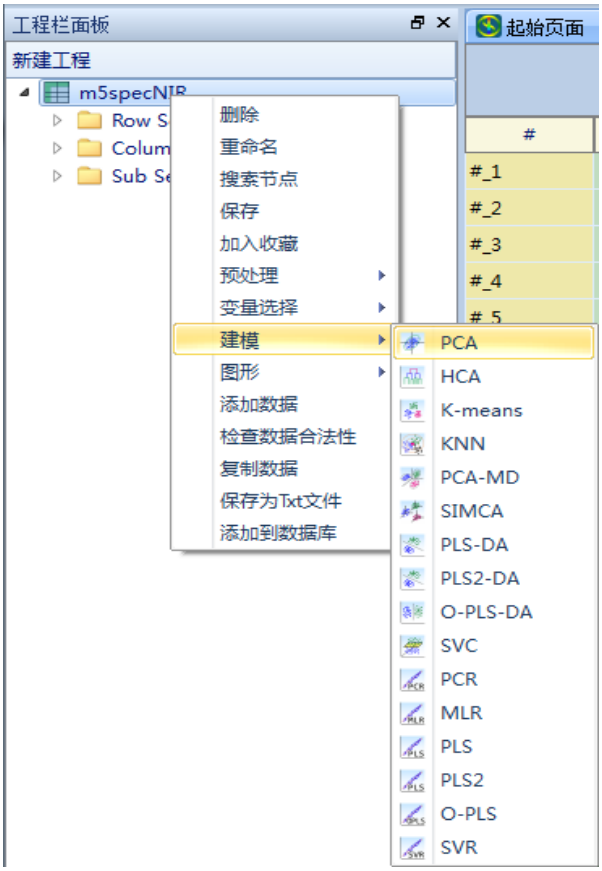
5.8. 建模

以用户自选节点作为数据来源，运用本软件提供的各种建模方法分析数据，仅支持数据节点。

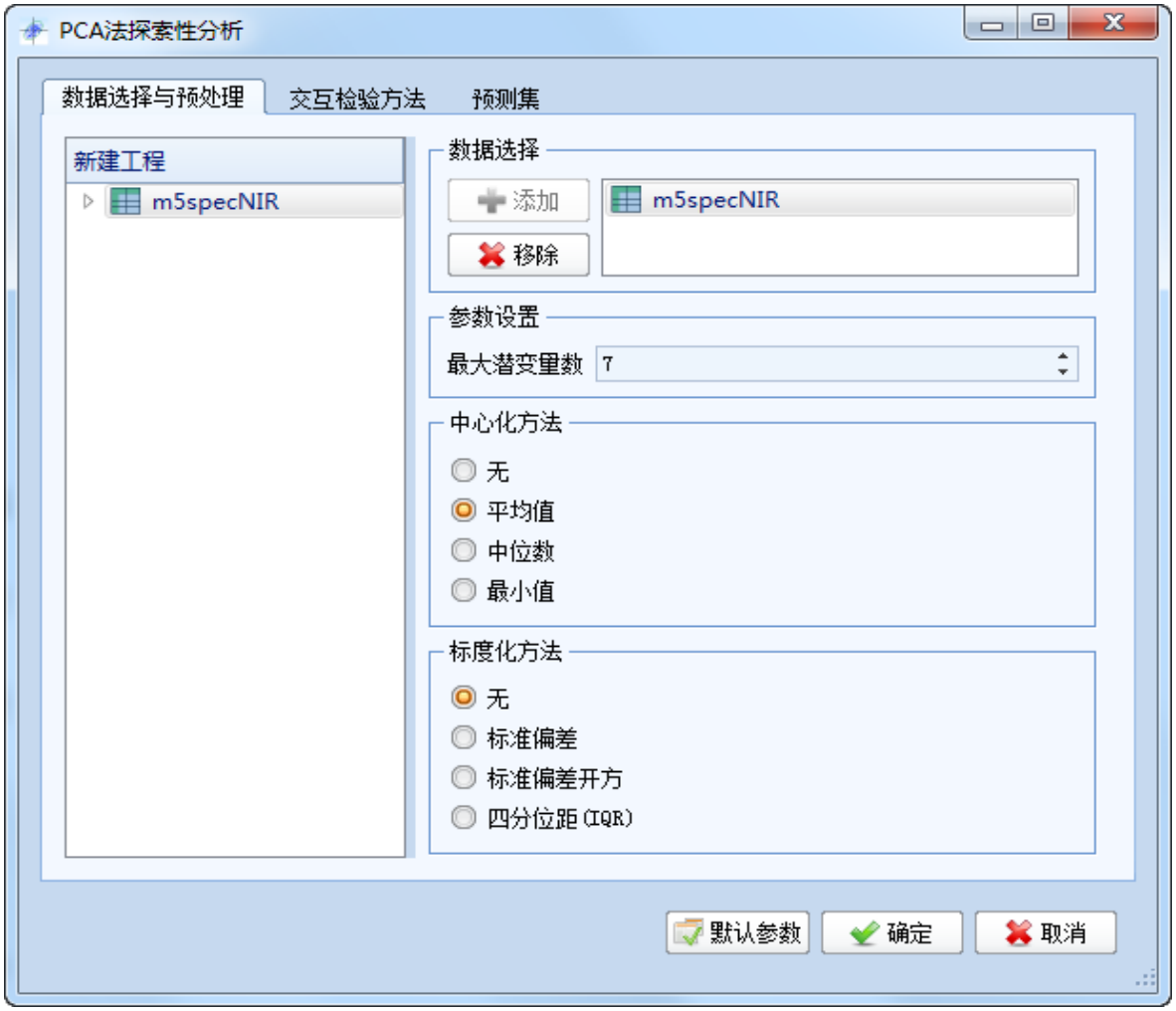
操作步骤:

步骤 1: 选中节点，点击右键。

步骤 2: 右键菜单中单击**建模**菜单项，再选择一个具体的建模方法:



单击菜单项，弹出被选变量选择方法对话框，所选节点数据自动加载到变量选择窗口中，如下图所示(以 PCA 法为例) :



步骤 3: 接下来的操作及更多建模方法对话框请参考**建模**章节。

5.9. 图形

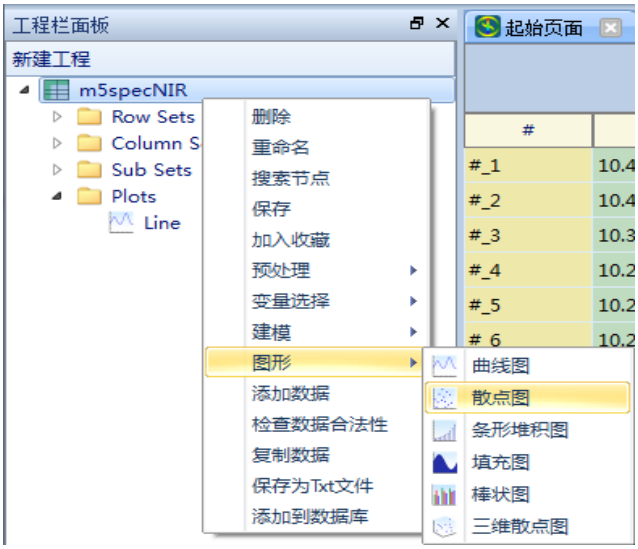
关于图形功能的描述及说明，详情请参见第九章。

操作步骤:

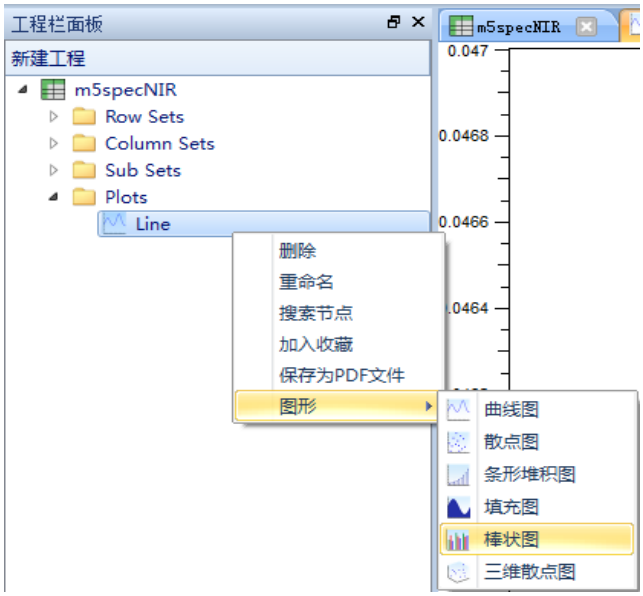
步骤 1: 选中节点，点击右键。

步骤 2: 右键菜单中单击**图形**菜单项，再从子菜单中任意选择一项。

- **第一种情形: 若选择数据节点，则界面如下图所示。**



- **第二种情形：若选择图形节点，则界面如下图所示。**



步骤 3：选择某种绘图方式后，弹出该绘图方式所对应的对话框，以棒状图为例，对话框如下图所示。绘图方式不同，该对话框可能亦随之改变。

步骤 4: 接下来的操作以及更详细的绘图方式请参考**图形**章节。

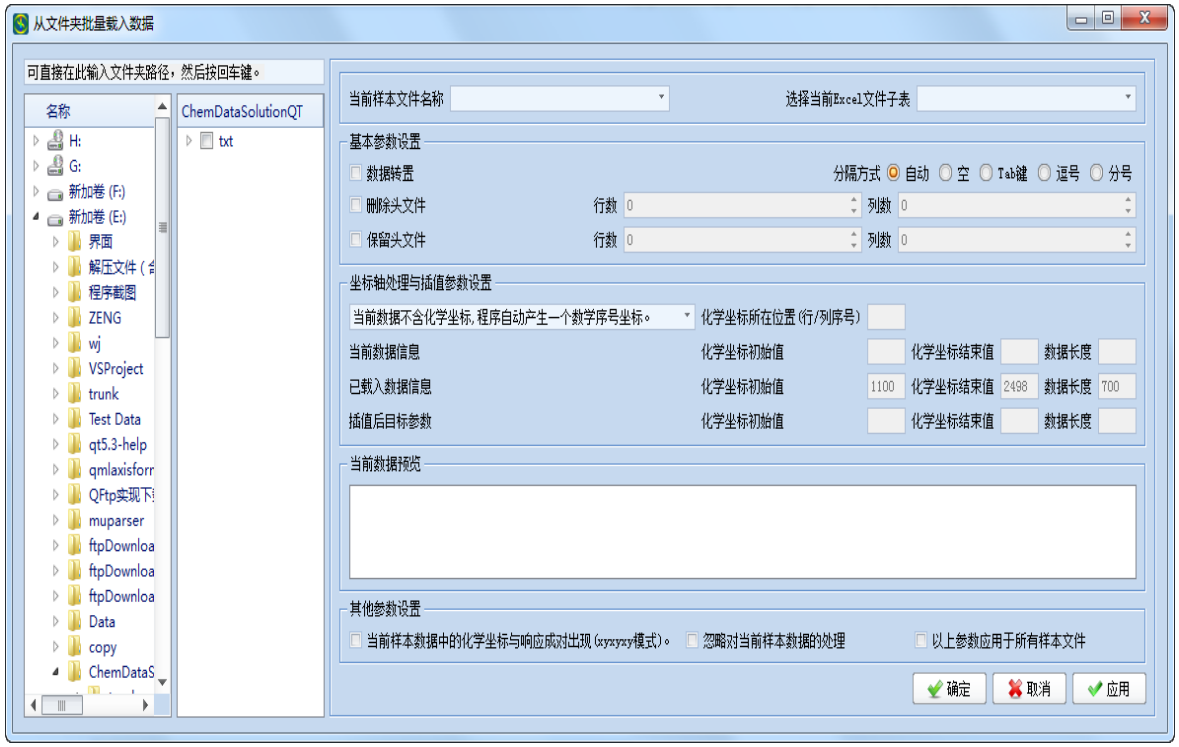
5.10. 添加数据

添加数据到当前节点，功能雷同于从文件夹批量载入数据，且选择载入新数据到已经存在的基本数据表中的情形，仅支持基本数据节点。

操作步骤:

步骤 1: 选中节点，点击右键。

步骤 2: 右键菜单中单击**添加数据**菜单项，弹出如下对话框(即从文件夹批量载入数据的界面，详见下图:



步骤 3: 接下来的操作以及更详细的内容请参考主页 -> 从文件夹载入数据章节。

本软件提供用户在处理数据的过程中，再次导入数据的功能，极大方便数据分析比较。

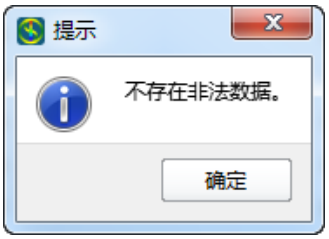
5.11. 检查数据合法性

检查数据合法性是指检查数据中是否包含程序无法处理的数据或非法字符，仅支持数据节点。所谓非法数据通常是指 nan 和 inf 二种情形。

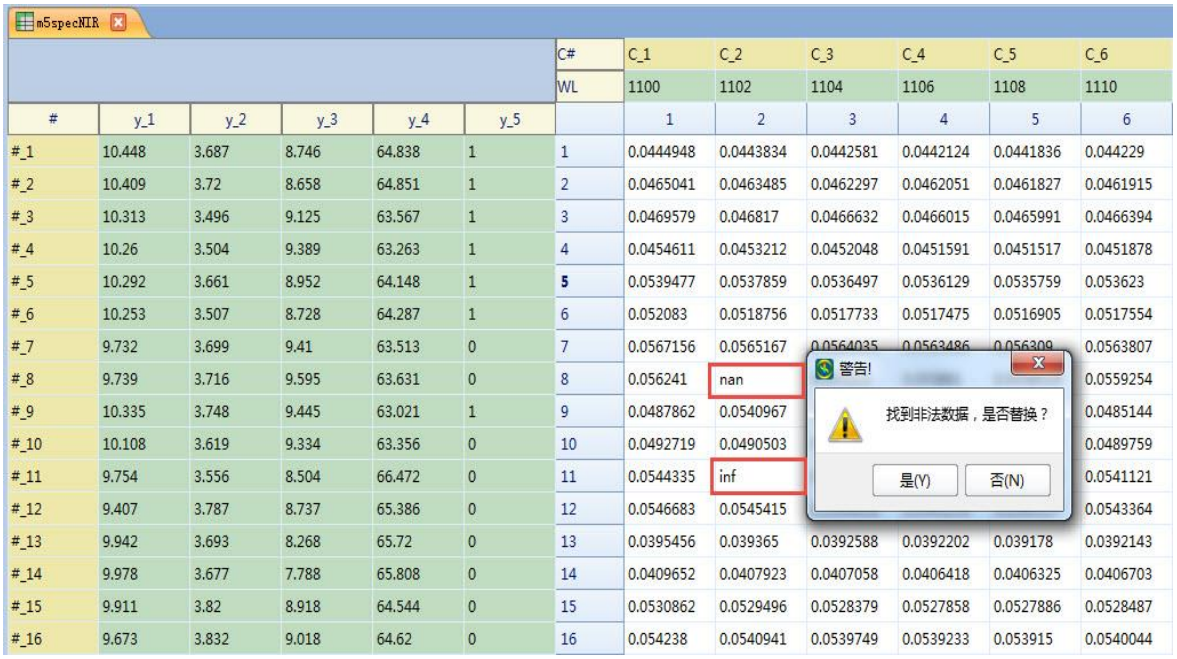
操作步骤:

步骤 1: 选中节点，点击右键。

步骤 2: 右键菜单中单击**检查数据合法性**菜单项,若所选数据节点中全部数据均合法，则出现如下图所示的提示信息。



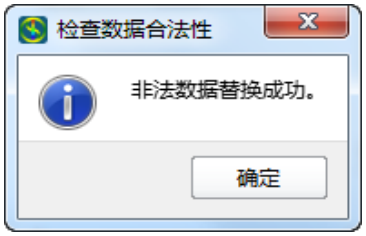
若数据表中含有非法数据，则出现如下图所示的提示信息。



选择**否**，退出提示对话框，数据表保持原样，不做任何修改；若选择**是**，则弹出如下图所

示的对话框，允许用户输入需要替换的数值:

点击确定便可将原数据表中的非法字符进行替换；若点击取消按钮，则取消替换，数据表仍然保持原样。数据替换成功后，出现如下信息提示用户。



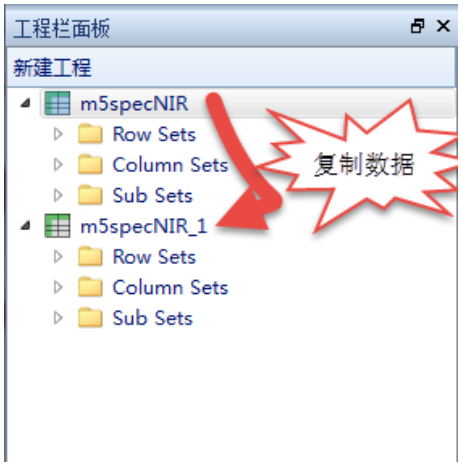
5.12. 复制数据

复制当前数据节点中的内容为一个新节点，仅支持数据节点。

操作步骤:

步骤 1: 选中节点，点击右键。

步骤 2: 右键菜单中单击**复制数据**菜单项，即可得到一个复制后的新节点，如下图所示:



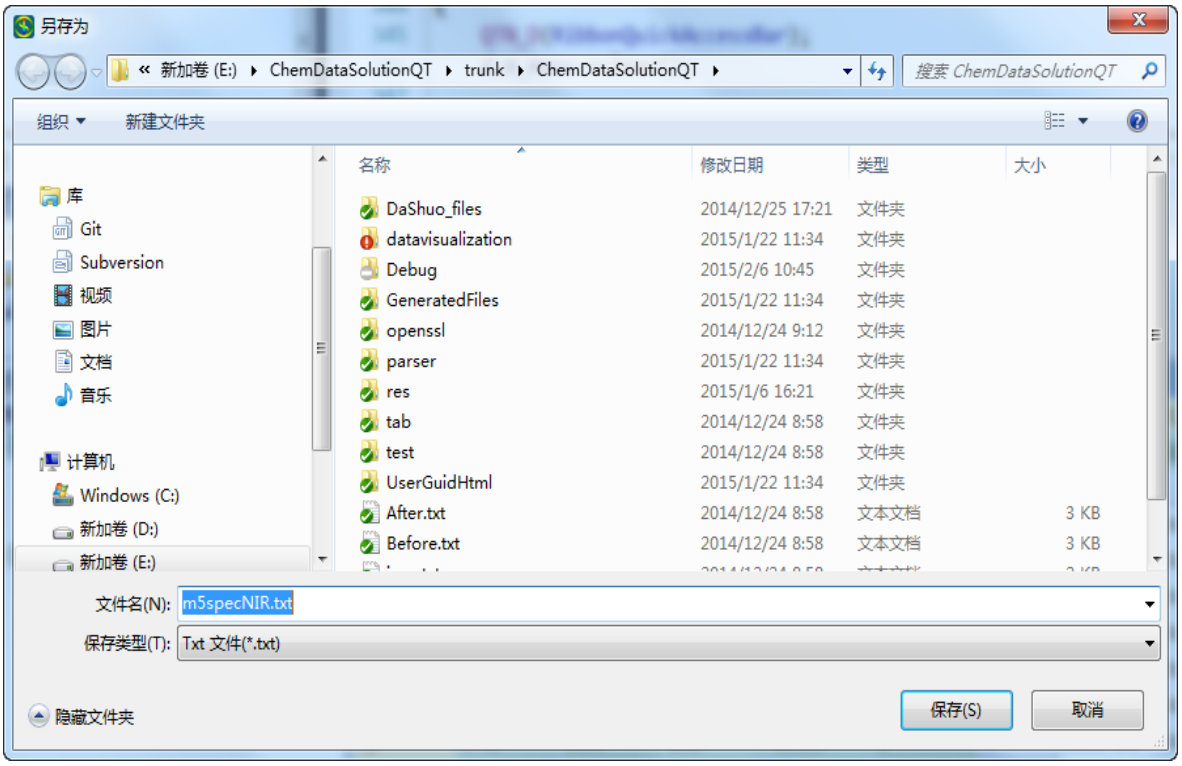
5.13. 保存为 txt 文件

将当前数据节点中的内容另存为一个 txt 文件。

操作步骤:

步骤 1: 选中节点，点击右键。

步骤 2: 右键菜单中单击**保存为 txt 文件**菜单项，弹出保存对话框，如下图所示:



文件名默认保存为节点名.txt。用户选择路径后，点击保存即可保存数据节点的内容到节点名.txt 文件中。

5.14. 添加到数据库

将当前数据节点中的内容添加到数据库中，仅支持数据节点。

操作步骤:

步骤 1: 选中节点，点击右键。

步骤 2: 右键菜单中单击**添加到数据库**菜单项。

步骤 3: 接下来的步骤请参考主页 -> 从数据库载入数据章节。

用户可在将数据添加到数据库时，对每个数据样本添加说明信息。

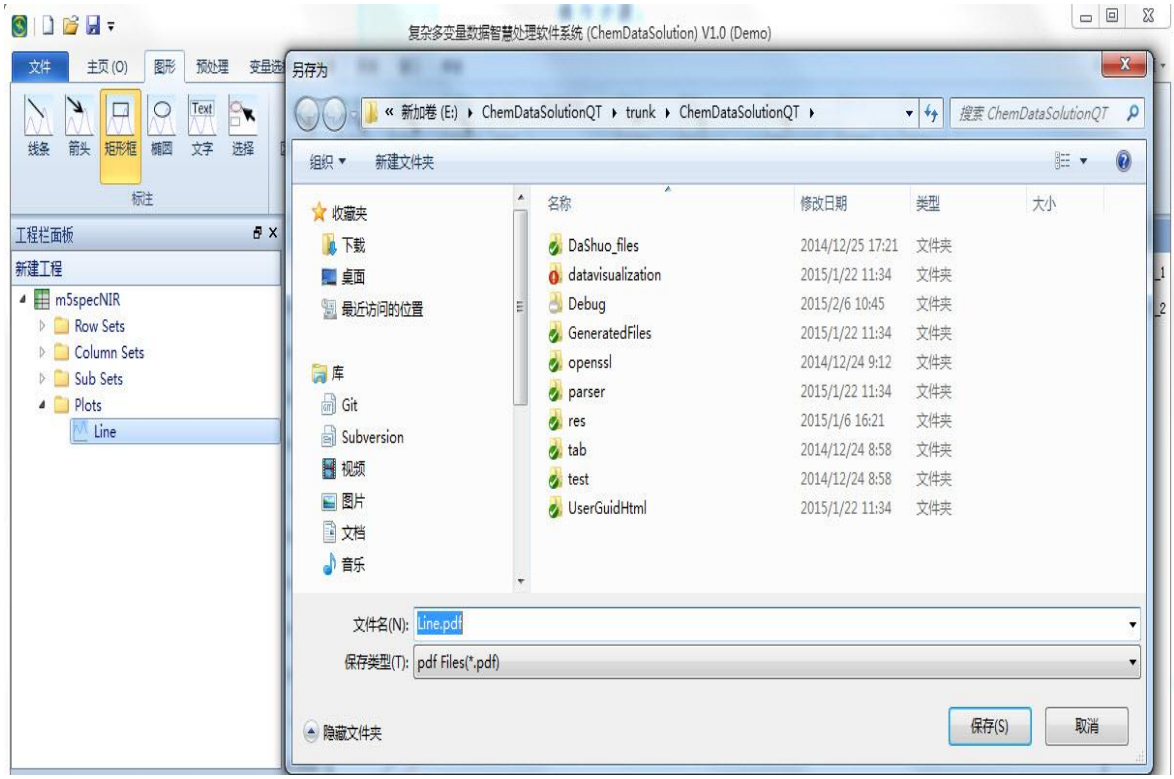
5.15. 保存为 PDF 文件

将当前图形另存为 PDF 文件，仅支持图形节点。

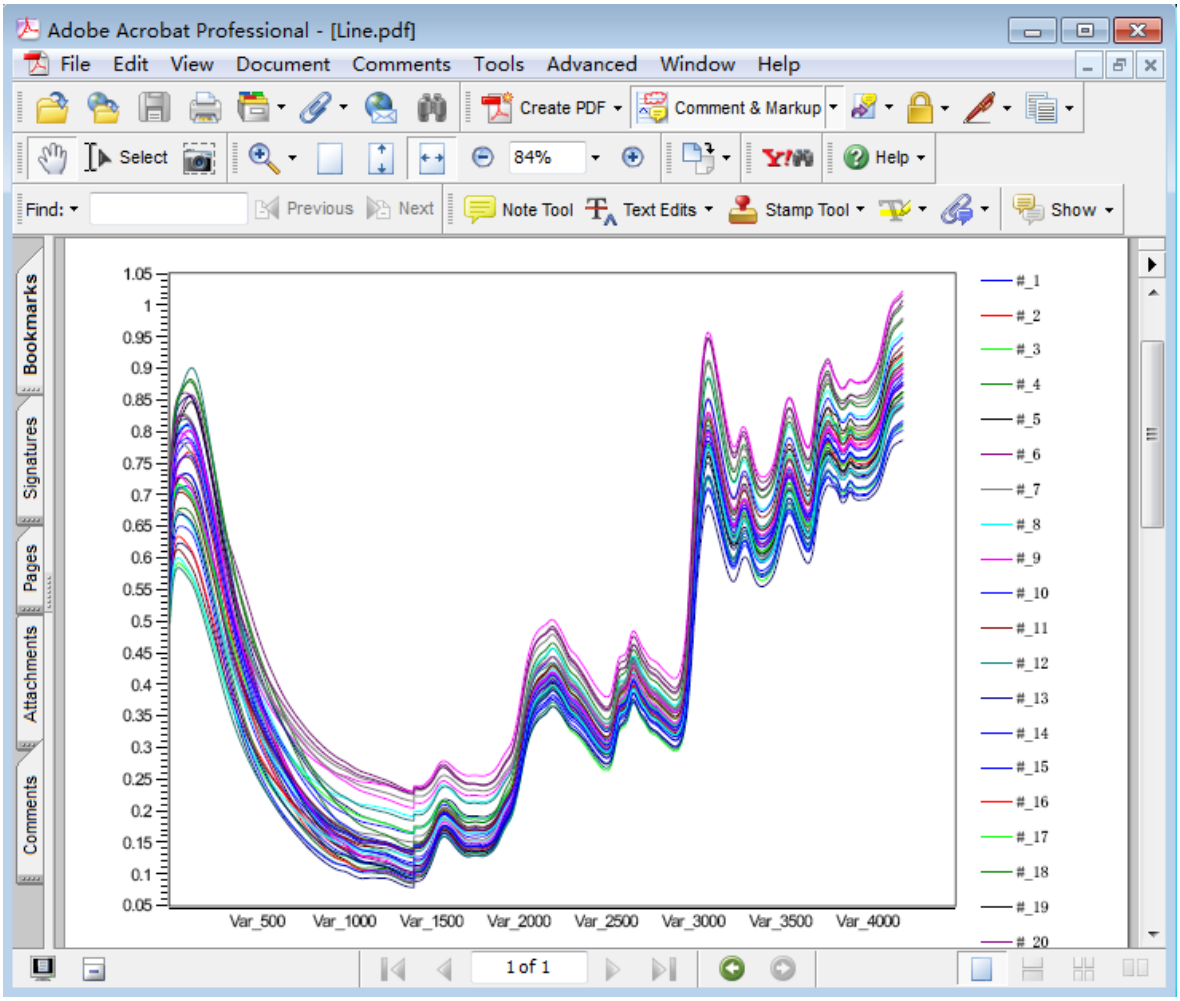
操作步骤:

步骤 1: 选中节点，点击右键。

步骤 2: 右键菜单中单击**保存为 PDF 文件**菜单项，弹出如下图对话框。文件名默认为节点名.pdf。



步骤 3: 选择保存路径，点击**保存**即可保存节点内容到 PDF 文件中，如下图是所保存 PDF 文件的截图。



从图中可以看出，保存为 PDF 文件的图形分辨率高，清晰度好，完全能满足各种苛刻的应用需要，如发表高质量的研究论文，或出版书籍著作等。

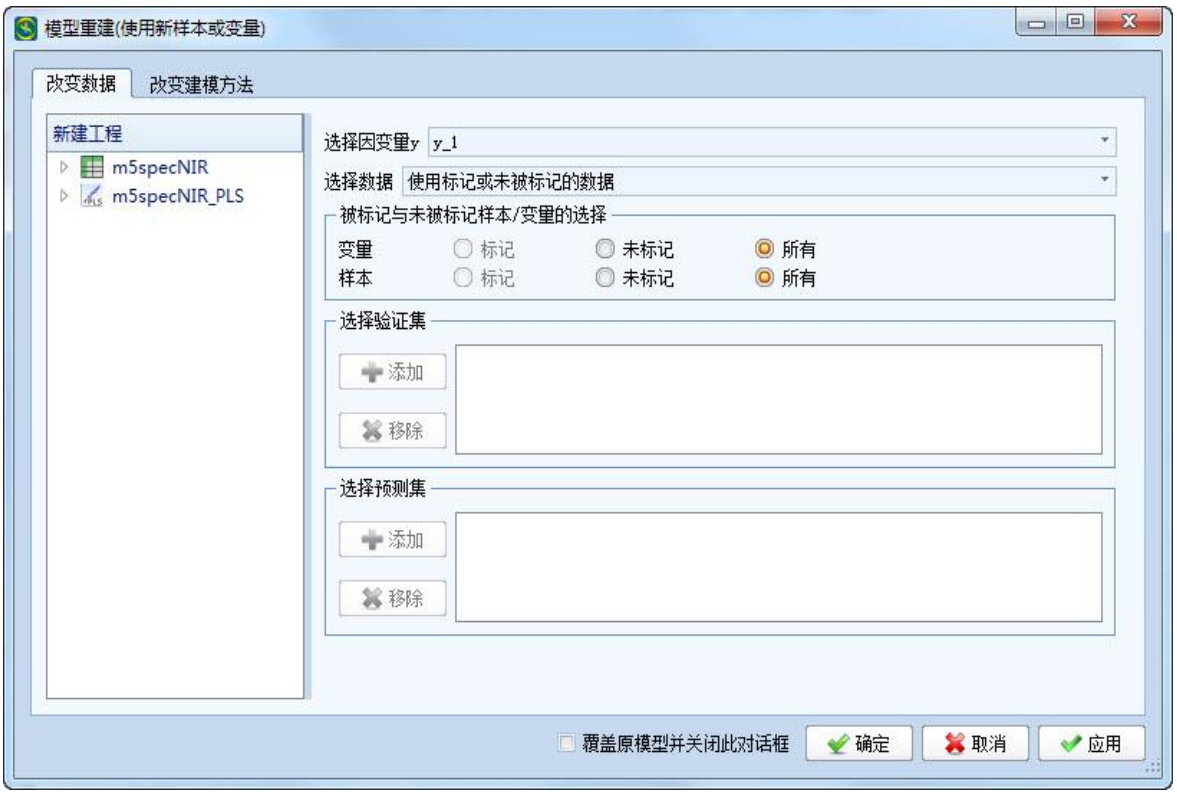
5.16. 模型修改

模型的构建往往需要进行较多的前期处理，包括数据准备和数据质量提高等。若用户构建模型后，其结果并不理想，或者希望修改数据或方法建立新的模型，则可通过模型修改功能快速获得新的模型结果，而无需重新开始所有的分析处理。毫无疑问，本功能仅支持模型节点。

操作步骤:

步骤 1: 选中节点，点击右键。

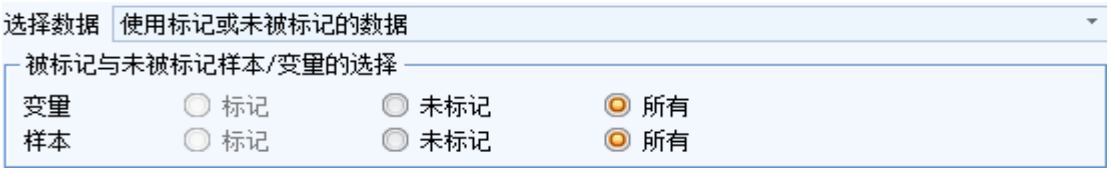
步骤 2：右键菜单中单击**模型修改**菜单项，弹出如下对话框：



修改模型可修改以下二部分的内容：

1) 改变数据：通过第一个标签页改变建模数据。数据来源可以是标记或未被标记的数据，即通过已建模型的图形结果，从图中直接标记数据样本或变量，亦可以是新数据。

若在选择数据框中选择使用标记或未被标记的数据，则通过图形中标记的样本或变量选择数据，如下图所示：



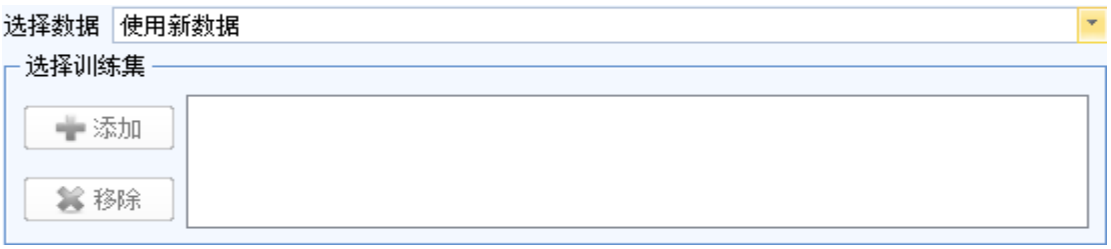
变量和样本的选择均包括三种情形，即

- 标记。
- 未标记。
- 所有。

其中标记单选框仅在原数据中有被标记的数据才有效，同样未标记单选框也只有在原数据中有未被标记的数据才有效，所有选项则总是有效的。此时，用户可选择如下表所示的方式，产生不同情形的新数据。

		变量		
		标记	未标记	所有
样本	标记	**情形 1**	**情形 2**	**情形 3**
	未标记	**情形 4**	**情形 5**	**情形 6**
	所有	**情形 7**	**情形 8**	**情形 9**

从上表可以清楚地看出，获取新数据的丰富形式，极大地提高了本软件的用户体验与可用性。若选择**新数据**时，则显示如下界面，此时用户可选择全新的数据进行建模。

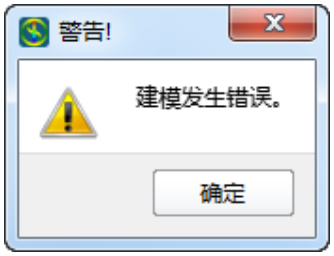


使用新数据指的是用户从工程中选择一个不同的数据节点作为训练集构建模型。有关训练集，验证集和预测集的更多内容，请参考主页 -> 批处理 -> 应用批章节。

2) 改变建模方法：通过第二个标签页实现改变建模方法，从而重构模型，如下图所示。建模方法选择，以及参数设置等更多内容，请参考**主页** -> **批处理** -> **修改批**章节。

步骤 3：点击应用或确定，即可重新建模。

在上述图中，若复选框 覆盖原模型并关闭此对话框 处于未被选中的状态，则重新建模的结果将作为一个新的模型节点加入到工程中；若该复选框处于选中状态，则重新建模的结果节点将替换以前的模型节点。若重新建模失败，则显示如下图所示的对话框，提示建模出错。



模型修改具有优异的用户体验，使用得当，可极大提高数据分析的效率。

第六章　基本数据表

基本数据表是数据处理的起始点，在正式详细介绍本软件的菜单功能前，先介绍这一部分，以使用户对数据表及其操作功能有更好的认识。本软件所述的基本数据表是指用户新建工程后，通过如下三种方式载入到工程中的数据，详情请参见 2.1.1.以及 8.1.。

- 从单个文件载入数据。
- 从文件夹载入数据。
- 从数据库载入数据。

实际的数据分析处理，可以是基本数据表，亦可以是在基本数据表的基础上，通过如下三种方式得到的新数据：

- 对基本数据表重新提取划分后，得到的行划分、列划分或子数据。
- 对基本数据表进行某种预处理或分析后得到的新数据。
- 综合以上二种方法得到的新数据。

本软件支持一个工程内含有多个基本数据表。在通过上述三种方式导入数据时，用户可选择将数据导入已有的基本数据表，或者导入到新的基本数据表。行划分、列划分、子数据表，以及经过数据分析处理后得到的数据表，使程序自动记录相关信息后，所得到基本数据表的关联表。其右键菜单操作和功能与基本数据表不完全相同。

用户可在完成对基本数据表的操作和管理后，再对其进行划分或分析，得到行划分、列划分、子数据表，或数据处理得到新的数据等。

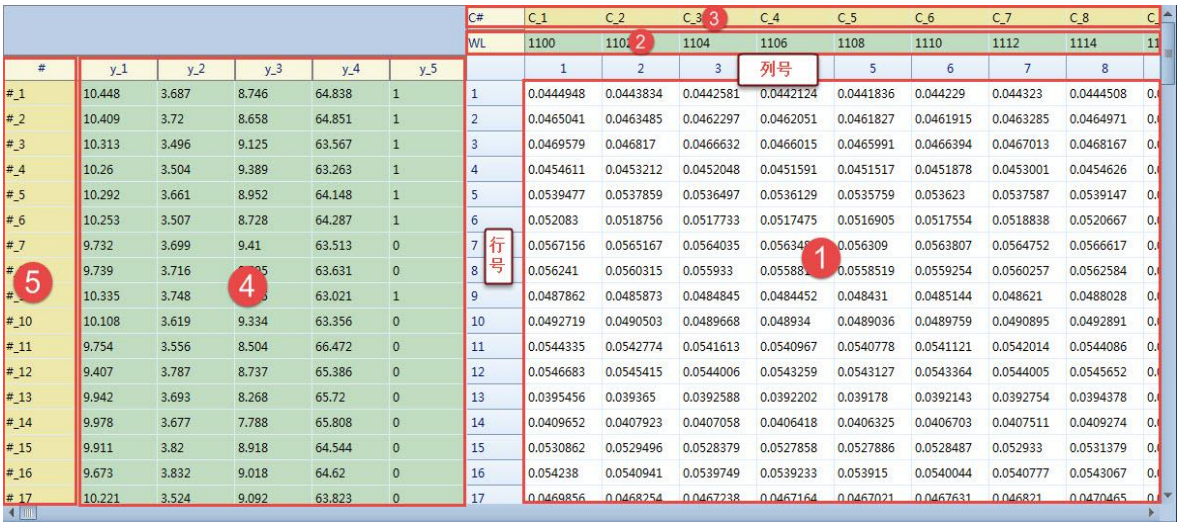
具体来说，用户创建一个新的工程，导入数据后即产生一个基本数据节点，其内容就是基本数据表。对行划分，列划分和子数据表的操作是对基本数据表操作中的一部分，实因各数据表的结构一致，故此处仅对基本数据表进行详细介绍。

6.1. 基本数据表的获取

基本数据表通过从文件或文件夹中载入数据，或者从数据库中导出数据获得，相关内容已在 2.1.1.中有基本介绍，用户可以参考阅读；详细的数据载入操作则可参考 8.1.。

6.2. 基本数据表的结构

基本数据表可划分为五个部分，如下图所示:



数据表中的行、列均有序号标记(图中所述行号和列号)。双击数据表中的任一单元格即进入编辑状态，用户可修改对应的数据或字符。图中数字所代表的区域依次表征如下内容:

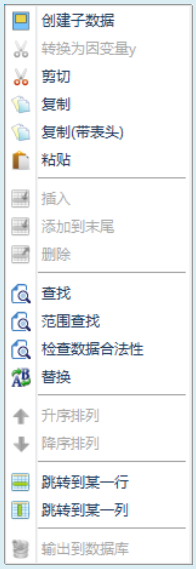
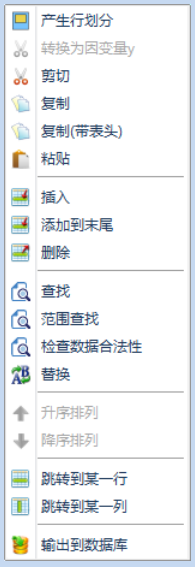
- 数据矩阵 X(自变量)。
- 变量的化学坐标(或数学坐标)。
- 变量属性(如变量名称等)。
- 属性矩阵 y(因变量)。
- 样本属性(如样本名称等)。

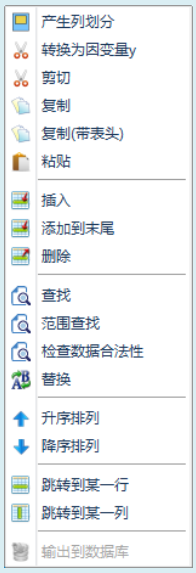
接下来依次介绍各区域的操作功能。

6.3. 对数据矩阵 X(自变量)的操作

数据矩阵 X(自变量)的操作，除编辑操作外都是通过右键菜单实现的。如上图所示，数据矩阵 X 的基本元素是单元格，由此衍生整行或整列。

用户点击单元格，整行或整列时的右键菜单功能亦是有所不同的，如下表所示。

序号	数据区域	右键菜单功能	说明
1	单元格		单元格是数据表格最小元素，其右键功能包括： 1) 创建子数据。 2) 剪切、复制、复制(带表头)、粘帖。 3) 查找、范围查找、检查数据合法性、替换。 4) 跳转到某一行、跳转到某一列。
2	整行		整行是指包含某一行中所有元素(变量)的一个向量。其右键功能包括： 1) 产生行划分。 2) 剪切、复制、复制(带表头)、粘帖。 3) 插入、添加到末尾、删除。 4) 查找、范围查找、检查数据合法性、替换。 5) 跳转到某一行、跳转到某一列。 6) 输出到数据库。

3	整列		整行是指包含某一行中所有元素(变量)的一个向量。其右键功能包括： 1) 产生列划分。 2) 转换为因变量。 3) 剪切、复制、复制(带表头)、粘帖。 4) 插入、添加到末尾、删除。 5) 查找、范围查找、检查数据合法性、替换。 6) 升序排列、降序排列。 7) 跳转到某一行、跳转到某一列。

右键菜单的样式基本相同(些许差异)，其差异通过其激活状态和功能的添加和删除来体现。接下来依次介绍上述所有功能。

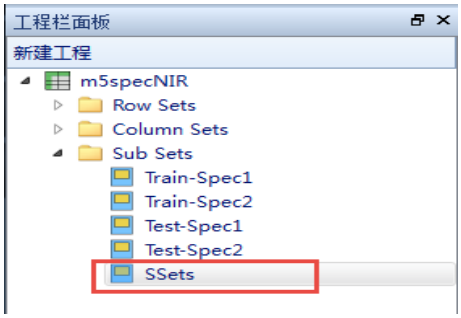
6.3.1. 创建子数据

用基本数据表中的部分样本，部分变量构建一个新的子数据(整行或整列除外)。

操作步骤:

步骤 1: 选择数据(部分行以及部分列)。

步骤 2: 单击右键，在右键菜单中选择**创建子数据**菜单项，即可创建子数据，如下图所示。在 Sub Sets 节点文件夹下，系统自动产生名为 SSets 的数据节点，关于节点的名称，可参见 4.3.2.。



事实上，相对于基本数据表而言，行划分和列划分数据均为子数据。本软件定义行和列划分数据，在于将他们与更一般的子数据区分(既非整行，亦非整列的数据)。

6.3.2. 转换为因变量 y

本功能即将数据矩阵 **X** 中的某一列，转换为为因变量 **y**，且将其移除矩阵数据 **X**，添加到因变量 y 中。

用户数据表中往往同时含有 X 和 y，此功能可简便实现 X 与 y 的同时导入。

操作步骤:

步骤 1: 选择数据(整列)，如下图所示。

步骤 2: 单击右键，在右键菜单中选择**转换为因变量 y** 菜单项，即可开始转换。

						C#	C_1	C_2	C_3	C_4	C_5
						WL	1100	1102	1104	1106	1108
#	y_1	y_2	y_3	y_4	y_5		1			4	5
#_1	10.448	3.687	8.746	64.838	1	1	0.0444948			0.0442124	0.0441836
#_2	10.409	3.72	8.658	64.851	1	2	0.0465041			0.0462051	0.0461827
#_3	10.313	3.496	9.125	63.567	1	3	0.0469579			0.0466015	0.0465991
#_4	10.26	3.504	9.389	63.263	1	4	0.0454611			0.0451591	0.0451517
#_5	10.292	3.661	8.952	64.148	1	5	0.0539477			0.0536129	0.0535759
#_6	10.253	3.507	8.728	64.287	1	6	0.052083			0.0517475	0.0516905
#_7	9.732	3.699	9.41	63.513	0	7	0.0567156			0.0563486	0.056309
#_8	9.739	3.716	9.595	63.631	0	8	0.056241			0.055881	0.0558519
#_9	10.335	3.748	9.445	63.021	1	9	0.0487862			0.0484452	0.048431
#_10	10.108	3.619	9.334	63.356	0	10	0.0492719			0.048934	0.0489036
#_11	9.754	3.556	8.504	66.472	0	11	0.0544335			0.0540967	0.0540778
#_12	9.407	3.787	8.737	65.386	0	12	0.0546683			0.0543259	0.0543127
#_13	9.942	3.693	8.268	65.72	0	13	0.0395456			0.0392202	0.039178
#_14	9.978	3.677	7.788	65.808	0	14	0.0409652			0.0406418	0.0406325
#_15	9.911	3.82	8.918	64.544	0	15	0.0530862			0.0527858	0.0527886
#_16	9.673	3.832	9.018	64.62	0	16	0.054238			0.0539233	0.053915
#_17	10.221	3.524	9.092	63.823	0	17	0.0469856			0.0467164	0.0467021
#_18	9.857	3.3	9.452	63.913	0	18	0.0455244			0.0452332	0.0452291
#_19	10.302	3.46	9.333	62.826	1	19	0.046162	0.0459942	0.0459172	0.0458935	0.0458906
#_20	9.818	3.446	9.073	64.292	0	20	0.0487734	0.0485991	0.0485026	0.0484662	0.0484444
#_21	10.169	3.541	9.711	63.099	0	21	0.0457413	0.0455802	0.0455001	0.0454223	0.0454158
#_22	10.034	3.417	9.694	63.246	0	22	0.0477657	0.0476088	0.0475163	0.0474519	0.0474392
#_23	9.691	3.645	8.685	65.474	0	23	0.0497307	0.0495645	0.0494652	0.0493908	0.0493728
#_24	9.78	3.71	8.729	65.427	0	24	0.0486839	0.0485081	0.0484192	0.0483356	0.048331

数据转换后的效果如下图所示。

							C#	C_2	C_3	C_4	C_5
							WL	1102	1104	1106	1108
#	y_1	y_2	y_3	y_4	y_5			1	2	3	4
#_1	10.448	3.687	8.746	64.838	1	0.0444948	1	0.0443834	0.0442581	0.0442124	0.0441836
#_2	10.409	3.72	8.658	64.851	1	0.0465041	2	0.0463485	0.0462297	0.0462051	0.0461827
#_3	10.313	3.496	9.125	63.567	1	0.0469579	3	0.046817	0.0466632	0.0466015	0.0465991
#_4	10.26	3.504	9.389	63.263	1	0.0454611	4	0.0453212	0.0452048	0.0451591	0.0451517
#_5	10.292	3.661	8.952	64.148	1	0.0539477	5	0.0537859	0.0536497	0.0536129	0.0535759
#_6	10.253	3.507	8.728	64.287	1	0.052083	6	0.0518756	0.0517733	0.0517475	0.0516905
#_7	9.732	3.699	9.41	63.513	0	0.0567156	7	0.0565167	0.0564035	0.0563486	0.056309
#_8	9.739	3.716	9.595	63.631	0	0.056241	8	0.0560315	0.055933	0.055881	0.0558519
#_9	10.335	3.748	9.445	63.021	1	0.0487862	9	0.0485873	0.0484845	0.0484452	0.048431
#_10	10.108	3.619	9.334	63.356	0	0.0492719	10	0.0490503	0.0489668	0.048934	0.0489036
#_11	9.754	3.556	8.504	66.472	0	0.0544335	11	0.0542774	0.0541613	0.0540967	0.0540778
#_12	9.407	3.787	8.737	65.386	0	0.0546683	12	0.0545415	0.0544006	0.0543259	0.0543127
#_13	9.942	3.693	8.268	65.72	0	0.0395456	13	0.039365	0.0392588	0.0392202	0.039178
#_14	9.978	3.677	7.788	65.808	0	0.0409652	14	0.0407923	0.0407058	0.0406418	0.0406325
#_15	9.911	3.82	8.918	64.544	0	0.0530862	15	0.0529496	0.0528379	0.0527858	0.0527886
#_16	9.673	3.832	9.018	64.62	0	0.054238	16	0.0540941	0.0539749	0.0539233	0.053915
#_17	10.221	3.524	9.092	63.823	0	0.0469856	17	0.0468254	0.0467238	0.0467164	0.0467021
#_18	9.857	3.3	9.452	63.913	0	0.0455244	18	0.0453193	0.0452344	0.0452332	0.0452291
#_19	10.302	3.46	9.333	62.826	1	0.046162	19	0.0459942	0.0459172	0.0458935	0.0458906
#_20	9.818	3.446	9.073	64.292	0	0.0487734	20	0.0485991	0.0485026	0.0484662	0.0484444
#_21	10.169	3.541	9.711	63.099	0	0.0457413	21	0.0455802	0.0455001	0.0454223	0.0454158
#_22	10.034	3.417	9.694	63.246	0	0.0477657	22	0.0476088	0.0475163	0.0474519	0.0474392
#_23	9.691	3.645	8.685	65.474	0	0.0497307	23	0.0495645	0.0494652	0.0493908	0.0493728
#_24	9.78	3.71	8.729	65.427	0	0.0486839	24	0.0485081	0.0484192	0.0483356	0.048331

转换后的数据进入因变量 **y** 区域，效果见上图红色框中的内容。

6.3.3. 剪切

剪切数据文本，以便粘帖到新的区域。

操作步骤:

步骤 1: 选择数据区域。

步骤 2: 单击右键，在右键菜单中选择**剪切**菜单项，即可剪切掉当前选择的数据。

上述步骤完成后，即可使用数据粘帖功能，将被剪切的数据粘帖到新的区域。

6.3.4. 复制

复制数据文本，以便粘帖到新的区域。

操作步骤:

步骤 1: 选择数据。

步骤 2: 单击右键，在右键菜单中选择**复制**菜单项，即可复制当前选择的数据。

复制与剪切功能对应，上述步骤完成后，便可将被复制的数据粘帖到新的区域。

6.3.5. 复制(带表头)

与复制功能相同，但复制数据文本的同时复制数据表头，即表中区域 3 的说明性信息。表头是指变量的属性，即说明性信息。

操作步骤:

步骤 1: 选择数据。

步骤 2: 单击右键，在右键菜单中选择**复制(带表头)**菜单项，即可复制当前选择的数据以及对应的表头内容。如选择如下图中被标记数据进行**复制(带表头)**操作，则实际被复制的内容如图所示，包括了其上部分的说明性信息。

							C#	C_2	C_3
							WL	1102	1104
#	y_1	y_2	y_3	y_4	y_5	y1		1	2
#_1	10.448	3.687	8.746	64.838	1	0.0444948	1	0.0443834	0.0442581
#_2	10.409	3.72	8.658	64.851	1	0.0465041	2	0.0463485	0.0462297
#_3	10.313	3.496	9.125	63.567	1	0.0469579	3	0.046817	0.0466632
#_4	10.26	3.504	9.389	63.263	1	0.0454611	4	0.0453212	0.0452048

							C_2	C_3
							1102	1104
#_2	10.409	3.72	8.658	64.851	1	0.0465041	0.0463485	0.0462297
#_3	10.313	3.496	9.125	63.567	1	0.0469579	0.046817	0.0466632
#_4	10.26	3.504	9.389	63.263	1	0.0454611	0.0453212	0.0452048

6.3.6. 粘贴

粘贴上述被剪切或复制的文本，可以包含数据或表头。

操作步骤:

步骤 1: 在数据表中的任一单元格内，单击右键。

步骤 2: 在右键菜单中选择**粘贴**菜单项，即可粘贴剪切或复制得到的数据文本，并置于当前单元格(一个或多个)区域。

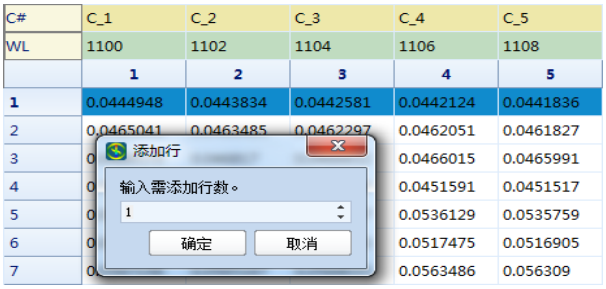
6.3.7. 插入

往基本数据表中插入用户自定义数量的行或列数，初始值均为 0(零)。

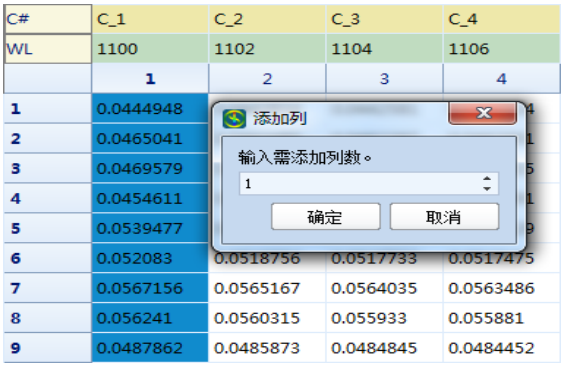
操作步骤:

步骤 1: 选择整行或整列数据。

步骤 2: 单击右键，在右键菜单中选择**插入**菜单项。若选择整行数据，则弹出如下图所示的对话框:



若选择整列数据，则弹出如下图所示的对话框:



步骤 3: 输入用户需要添加的行或列的数目，点击**确定**，即可插入对应的行或列数到被选择的行(或列)前，并以 0(零)作为初始值填充。若点击取消，则取消操作，并关闭对话框。

插入行或列后，用户则可将通过其他途径剪切或复制得到的数据粘帖其中，从而扩充基本数据表。

6.3.8. 添加到末尾

与上述 6.3.7.的插入功能雷同。不同点在于所添加的行或列自动置于行或列的末尾。

6.3.9. 删除

删除被选的行或列数据。

操作步骤:

步骤 1: 选择数据(行或列)。

步骤 2: 单击右键，在右键菜单中选择删除菜单项，即可删除被选的行或列。选择行或列数据时，Shift 和 Ctrl 功能键可用。

6.3.10. 查找

查找数据表中的数据。

操作步骤:

步骤 1: 在数据表中的任一单元格内，单击右键。

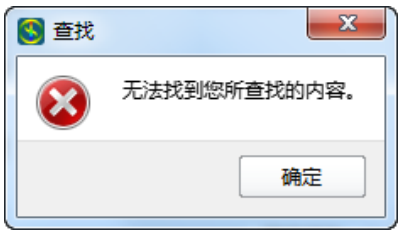
步骤 2: 在右键菜单中选择**查找**菜单项，弹出如下图中所示的对话框:

步骤 3：输入需要查找内容(数据)，点击**查找所有**，即找到所有符合条件的数据位置，并弹出如下图所示的新对话框，将查找得到的所有结果以表格的形式均列举出来，该表包括如下内容，以便用户获取查找内容的详情。

- 数据表名称
- 行号
- 列号
- 数值

选中查找结果中的某一行，即被找到的某一内容，系统将自动选中数据表中对应的单元格。若使用图中查找另一处的功能，程序自动跳到符合条件的下一处单元格并选中；点击关闭，

则关闭对话框。另外，若没查找到输入的数据，则显示如下对话框提示用户。



图中第二个标签页的功能为替换，用户可在此两个标签页间切换，该标签页功能请参考本章节之替换。

6.3.11. 范围查找

查找数据表中位于某一范围内的所有数据。

操作步骤:

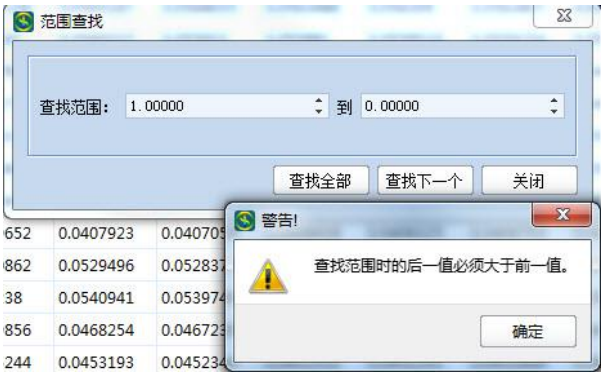
步骤 1: 在数据表中的任一单元格内，单击右键。

步骤 2: 在右键菜单中选择**范围查找**菜单项，弹出如下对话框:

步骤 3: 输入查找范围，点击**查找全部**，即查找所有符合条件的数据位置，并将查找得到的结果以如下图所示的形式列出来，该表亦包括与图所示的相同内容，以便用户获取查找内容的详情。接下来的操作和使用与 6.3.10.相同。

	数据表名称	行号	列号	数值
1	m5specNIR	4	3	0.0452048
2	m5specNIR	4	4	0.0451591
3	m5specNIR	4	5	0.0451517
4	m5specNIR	4	6	0.0451878
5	m5specNIR	4	7	0.0453001
6	m5specNIR	4	8	0.0454626
7	m5specNIR	4	9	0.0456637
8	m5specNIR	4	10	0.046025

需要注意的是，上图中数据的输入范围中，前一个数值须小于后一个数值，否则，程序以如下图提示用户。



6.3.12. 检查数据合法性

检查数据合法性是指检查数据中是否包含程序无法处理的数据或非法字符。

操作步骤:

步骤 1: 在数据表中的任一单元格内，单击右键。

步骤 2: 在右键菜单中选择**检查数据合法性**菜单项即可。

更多内容请参考 5.11.**节点文件夹与节点管理 -> 检查数据合法性。**

6.3.13. 替换

使用新内容替换数据表格中被查找的内容。

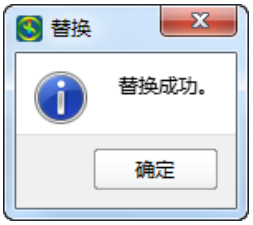
操作步骤:

步骤 1: 在数据表中的任一单元格内，单击右键。

步骤 2: 在右键菜单中选择**替换**菜单项，弹出如下对话框:

步骤 3: 输入查找内容和替换内容，点击**替换所有**，则数据表中的全部被查找内容，将改变为替换内容。点击替换，则每次仅替换一个。

在替换数据表中内容的同时，程序亦将选中被替换的单元格。替换成功后，出现图所示界面提示用户。



若点击**关闭**，则关闭对话框。点击**查找所有**，**查找下一处**操作，请参考本章节之**查找**。第一个标签页的内容为查找，两个标签页间可切换。

6.3.14. 升序排列

将数据表中的某列数据按从小到大的顺序排列。

操作步骤:

步骤 1: 选择某一整列数据。

步骤 2: 单击右键，在右键菜单中选择**升序排列**菜单项，即可将此列数据按升序排列(从小到大的顺序)。

6.3.15. 降序排列

与 6.3.14.雷同，数据顺序将按降序，即从大到小的顺序排列。

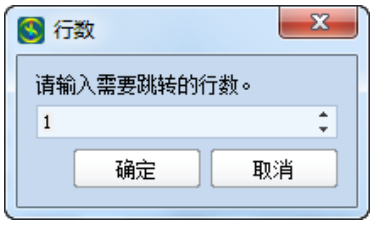
6.3.16. 跳转到某一行

跳转到用户指定的某一行，并选择标记。

操作步骤:

步骤 1: 在数据表中的任一单元格内，单击右键。

步骤 2: 在右键菜单中选择**跳转到某一行**菜单项，弹出如下对话框:



步骤 3: 输入需要跳转的行数(范围从 1 到数据表的最大行数)，点击**确定**，则系统将跳转并选中指定行，如下图所示。点击**取消**，则取消操作并关闭对话框。

		1	2	3	4	5	6	7	8	9
	2	0.2168950...	0	0	0.0299360...	0.3179138...	0.7697766...	0.0124207...	0.3062428...	0.3047681...
	3	0.7770003...	0	0	0.1767008...	0.4056156...	0.5022483...	0.7983619...	0.1275769...	0.4922994...
	4	0.1545440...	0	0	0.2558354...	0.1509502...	0.9736946...	0.9728593...	0.4160632...	0.1158763...
	5	0.9445655...	0	0	0.3353518...	0.2977693...	0.3613424...	0.9300068...	0.3775662...	0.1296551...
	6	0.4127845...	0	0	0.3668876...	0.2713971...	0.8204255...	0.9633683...	0.1222135...	0.0173481...
	7	0.7598351...	0	0	0.3865882...	0.5516526...	0.4559561...	0.1562075...	0.1201666...	0.3523045...
	8	0.9676799...	0	0	0.5289803...	0.0694683...	0.3076518...	0.4283391...	0.0235122...	0.3591004...
	9	0.2348464...	0	0	0.5616123...	0.0152376...	0.1424519...	0.7352804...	0.8737419...	0.8230415...
	10	0.2467601...	0	0	0.6126629...	0.4587966...	0.6687435...	0.2361767...	0.2238506...	0.3521672...
	11	0.9516471...	0	0	0.6337946...	0.0669544...	0.3563037...	0.9729138...	0.2351630...	0.6754733...
	12	0.4166275...	0	0	0.6621313...	0.8887370...	0.9445078...	0.7833127...	0.3407043...	0.6893942...
	13	0.5093996...	0	0	0.7884541...	0.5350666...	0.8260380...	0.8315242...	0.1995168...	0.6005059...

6.3.17. 跳转到某一列

与 6.3.16.雷同，唯一区别在于从行跳转与标记变成列跳转与标记。

6.3.18. 输出到数据库

将用户选定的某一行或多行样本数据导出到数据库中。

操作步骤:

步骤 1: 选中数据表中的一行或多行(Shift 和 Ctrl 键可用)，单击右键。

步骤 2: 在右键菜单中选择**输出到数据库**菜单项，弹出如下图所示对话框:

实因添加到数据库中的数据样本，必定对应说明性信息，因而在将数据添加到数据库前，用户可通过此界面预览数据，并输入样本信息。

本软件中所处理的数据以行作为样本，列作为变量。输入到数据表中的每一个数据

均表示一个样本，因而当且仅当用户选择整行数据时，输出到数据库功能才可用。

上图所示的界面，与用户之间打开数据库，直接添加数据记录时相同，其操作可参考 8.1.3. 从数据库载入数据章节。

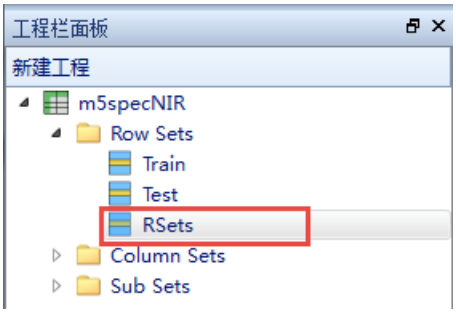
6.3.19. 产生行划分

选择基本数据表中的部分或全部样本对应的所有变量，重新构建一个新的数据，称为行划分。

操作步骤:

步骤 1: 选择数据(部分或全部行，所有列)。

步骤 2: 单击右键，在右键菜单中选择**产生行划分**菜单项，即可产生行划分，如下图所示:



关于行划分的更多内容，可参考 4.3.2.，以及第五章。

6.3.20. 产生列划分

与 6.3.19.雷同，唯一区别在于由对行的操作变为列操作，用户亦可参考 4.3.2.，以及第五章中的内容。

6.3.20. 产生子数据

如前所述，行划分和列划分所得到的数据均为子数据。本软件将子数据与行划分及列划分

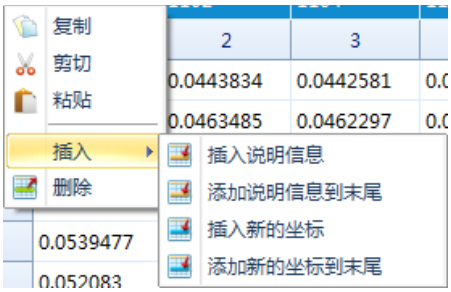
区分开来，在于突出前者的重要性。

本处所指的子数据是指基本数据表的一部分，但不包括行划分和列划分的情形，即基本数据表中非全部行，亦非全部列数据。

具体操作与 6.3.19.雷同，用户亦可参考 4.3.2.，以及第五章中的内容。

6.4. 对变量化学坐标(或数学坐标)的操作

如前所述，数据表中每列固定为样本的属性信息。以近红外数据的分析为例，若数据表中每行均为一个样本，则每列所对应变量的化学坐标则为某一具体波长，而数学坐标则为数学序号。在图中，首先全选任一组变量属性，然后点击右键，则出现如下图所示的功能。



其中，复制，剪切，粘贴，删除操作请参考本章节之对数据矩阵 X(自变量)的操作。

6.4.1. 插入新坐标

在图中的区域 3 内，插入一整行，为数据表产生一个新的化学或数学坐标。

操作步骤:

步骤 1: 在数据表的变量化学坐标或属性(亦称说明信息)区域内，选择整行数据(或信息)。

步骤 2: 单击右键，在**插入**菜单项子菜单中选择**插入新的坐标**菜单项，即可在指定位置插入一行新的坐标，结果如下图所示:

C#	C_1	C_2	C_3	C_4	C_5	C_6	C_7
V	V_1	V_2	V_3	V_4	V_5	V_6	V_7
x	1	2	3	4	5	6	7
WL	1100	1102	1104	1106	1108	1110	1112
	1	2	3	4	5	6	7
1	0.0320147	0.031834	0.0317745	0.0317151	0.0317356	0.0317282	0.0318142
2	0.0321434	0.032039	0.0319197	0.0318557	0.0318506	0.0318514	0.0319355
3	0.0332697	0.0331372	0.0330245	0.0329294	0.0328864	0.0328749	0.0329346

用户亦可编辑该坐标，或者使用剪切(复制)功能修改坐标。在绘制图形时，用户可选择已有的或新插入的坐标。

6.4.2. 添加新坐标到末尾

在数据表变量已有化学坐标的末尾，再插入一行新的坐标，即一个数据表可同时有多个化学或数学坐标。

具体操作步骤与 6.4.1.雷同，差异在于新插入坐标的放置位置。

6.5. 对变量属性(如变量名称等)的操作

对变量属性的操作与对变量坐标的操作雷同。事实上，这二个功能在同一界面内，此处分成二个小的章节，以突出其完全不同的意义和差异性。

如前所述，数据表中每列固定为样本的属性信息，比如：若数据表中每行均为一个人的血液抽检后的检测样本，则其对应的列则为血液中所检测到的小分子代谢物。基于此，每列变量的属性则为该小分子代谢物的信息，比如中/英文名称，分子量，分子式，CAS 号等等。

所有操作均可参考本章节之对变量化学坐标(或数学坐标)的操作。

6.5.1. 插入说明信息

在指定位置插入一行，并添加变量新的属性(说明性)信息。

与 6.4.1.雷同，差异在于插入的内容由坐标变为说明性信息。

6.5.2. 添加说明信息到末尾

在说明性信息的末尾，添加变量新的信息。

与 6.4.2.雷同，差异在于插入的内容由坐标变为说明性信息。

6.6. 对属性矩阵 y(因变量)的操作

本软件所说的数据处理，实际上包括如下二种可能：

1) 仅对数据矩阵 **X** 进行分析，如 PCA 和 HCA 等。

2) 同时对数据矩阵 X 和因变量 y 进行分析，如 PLS 和 PLS-DA 等。

基于此，基本数据表中显然需要包括对因变量 **y** 的操作。与此同时，本软件中因变量 **y** 亦包括如下二种类型：

1) 样本类别向量或矩阵(非连续)。在二类问题中，如产品合格与不合格，通常分别以 0 和 1 或者-1 与表示，用于模型构建；在多类问题中，如产品质量等级等，则以 1，2，3，。。。等序号表示。当然，针对多类问题，在实际的建模过程中，则基于类别信息转换为一个仅含 0 和 1 的数据矩阵进行处理。

2) 连续变化向量或矩阵。解决多变量的回归问题，亦是本软件的核心之一。如在复杂体系的分析中，用户需要得知一个或多个关键组份的含量(如中药活性成份或生物标志物等)。通过理论计算或实验量测的方式获取描述目标化合物含量的关联属性(自变量矩阵 X)，如近红外光谱，建立二者间的回归模型。对未知的体系中目标化合物的含量(预先获得描述矩阵 X，如量测得到近红外光谱)，便可能获得快速、准确的预测。该类向量或矩阵(如同时计算多个目标化合物的含量)中的元素，则可能是任意大小的数值，比如 0.01，1.0 或者 10.5 等。

在属性矩阵 y(因变量)区域点击右键菜单，则出现如下图所示的功能界面：

复制，剪切，粘贴，删除，查找，升序排列，降序排列，跳转到某一行操作请参考本章节之对数据矩阵 X(自变量)的操作。

对上面未介绍到的功能，逐个介绍如下。

6.6.1. 转换为自变量 X

将因变量 **y** 转换为自变量 **X**，实现数据矩阵与因变量矩阵的互换。

操作步骤:

步骤 1: 选择数据(整列)。

步骤 2: 单击右键，在右键菜单中选择**转换为自变量 X** 菜单项，即可开始转化:

数据转换后的结果如下图所示:

					C#		C_1	C_2	C_3	C_4	C_5
					WL	0	1100	1102	1104	1106	1108
#	y_1	y_2	y_3	y_4		1	2	3	4	5	6
#_1	10.448	3.687	8.746	64.838	1	1	0.0444948	0.0443834	0.0442581	0.0442124	0.0441836
#_2	10.409	3.72	8.658	64.851	2	1	0.0465041	0.0463485	0.0462297	0.0462051	0.0461827
#_3	10.313	3.496	9.125	63.567	3	1	0.0469579	0.046817	0.0466632	0.0466015	0.0465991
#_4	10.26	3.504	9.389	63.263	4	1	0.0454611	0.0453212	0.0452048	0.0451591	0.0451517
#_5	10.292	3.661	8.952	64.148	5	1	0.0539477	0.0537859	0.0536497	0.0536129	0.0535759
#_6	10.253	3.507	8.728	64.287	6	1	0.052083	0.0518756	0.0517733	0.0517475	0.0516905
#_7	9.732	3.699	9.41	63.513	7	0	0.0567156	0.0565167	0.0564035	0.0563486	0.056309
#_8	9.739	3.716	9.595	63.631	8	0	0.056241	0.0560315	0.055933	0.055881	0.0558519
#_9	10.335	3.748	9.445	63.021	9	1	0.0487862	0.0485873	0.0484845	0.0484452	0.048431
#_10	10.108	3.619	9.334	63.356	10	0	0.0492719	0.0490503	0.0489668	0.048934	0.0489036
#_11	9.754	3.556	8.504	66.472	11	0	0.0544335	0.0542774	0.0541613	0.0540967	0.0540778
#_12	9.407	3.787	8.737	65.386	12	0	0.0546683	0.0545415	0.0544006	0.0543259	0.0543127
#_13	9.942	3.693	8.268	65.72	13	0	0.0395456	0.039365	0.0392588	0.0392202	0.039178
#_14	9.978	3.677	7.788	65.808	14	0	0.0409652	0.0407923	0.0407058	0.0406418	0.0406325
#_15	9.911	3.82	8.918	64.544	15	0	0.0530862	0.0529496	0.0528379	0.0527858	0.0527886

6.6.2. 插入说明信息

在指定位置插入一整列新的说明信息。如前所述，本软件约定行为样本，如某药品、烟草或者组学研究中的分析对象。因而每行均可能有系列描述属性，如药品的生产厂商，批号，原材料，以及制造工艺等，以对样本进行完整说明。用户便可通过此功能增加新的样本描述信息。

操作步骤:

步骤 1: 选择数据(整列)。

步骤 2: 单击右键，在**插入**菜单项的子菜单中选择**插入说明信息**菜单项，即可在指定位置插入一行新的说明信息，结果如下图所示:

#1	#	y_1	y_2	y_3	y_4	y1
S_1	#_1	10.448	3.687	8.746	64.838	1
S_2	#_2	10.409	3.72	8.658	64.851	1
S_3	#_3	10.313	3.496	9.125	63.567	1
S_4	#_4	10.26	3.504	9.389	63.263	1
S_5	#_5	10.292	3.661	8.952	64.148	1
S_6	#_6	10.253	3.507	8.728	64.287	1
S_7	#_7	9.732	3.699	9.41	63.513	0
S_8	#_8	9.739	3.716	9.595	63.631	0
S_9	#_9	10.335	3.748	9.445	63.021	1
S_10	#_10	10.108	3.619	9.334	63.356	0
S_11	#_11	9.754	3.556	8.504	66.472	0
S_12	#_12	9.407	3.787	8.737	65.386	0

用户可进一步编辑单元格中的信息，修改成合适的内容。

6.6.3. 添加说明信息到末尾

在说明性信息的末尾，添加变量新的信息。

与 6.6.2.雷同，差异在于插入的内容由用户指定位置变为末尾。

6.6.4. 插入因变量 y

在指定位置插入一行新的因变量 **y**。

操作步骤:

步骤 1: 选择数据(整列)。

步骤 2: 单击右键，在**插入**菜单项的子菜单中选择**插入因变量 y** 菜单项，即可在指定位置插入一行新的因变量 **y**，结果如下图所示:

#1	#	y_1	y2	y_2	y_3	y_4	y1
S_1	#_1	10.448	0	3.687	8.746	64.838	1
S_2	#_2	10.409	0	3.72	8.658	64.851	1
S_3	#_3	10.313	0	3.496	9.125	63.567	1
S_4	#_4	10.26	0	3.504	9.389	63.263	1
S_5	#_5	10.292	0	3.661	8.952	64.148	1
S_6	#_6	10.253	0	3.507	8.728	64.287	1
S_7	#_7	9.732	0	3.699	9.41	63.513	0
S_8	#_8	9.739	0	3.716	9.595	63.631	0
S_9	#_9	10.335	0	3.748	9.445	63.021	1
S_10	#_10	10.108	0	3.619	9.334	63.356	0
S_11	#_11	9.754	0	3.556	8.504	66.472	0
S_12	#_12	9.407	0	3.787	8.737	65.386	0
S_13	#_13	9.942	0	3.693	8.268	65.72	0
S_14	#_14	9.978	0	3.677	7.788	65.808	0
S_15	#_15	9.911	0	3.82	8.918	64.544	0
S_16	#_16	9.673	0	3.832	9.018	64.62	0
S_17	#_17	10.221	0	3.524	9.092	63.823	0
S_18	#_18	9.857	0	3.3	9.452	63.913	0
S_19	#_19	10.302	0	3.46	9.333	62.826	1
S_20	#_20	9.818	0	3.446	9.073	64.292	0

6.6.5. 添加因变量 y

在因变量 **y** 的末尾，添加变量新的因变量。

与 6.6.4.雷同，差异在于插入的内容由用户指定位置变为末尾。

6.6.6. 等值刷

将选中区域单元格中的内容刷成相同的数值。

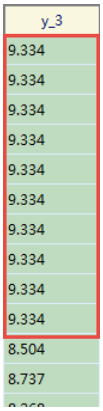
操作步骤:

步骤 1: 选择数据(整列)。

步骤 2: 单击右键，选择**等值刷**菜单项，即可把选中区域单元格中的内容刷成同样的值。

使用等值刷功能前:

使用等值刷功能后见下图，最后数值为用户右键点击等值刷时所选中的单元格内的值。



此功能主要用于用户插入或添加新的因变量 y 后，当其为分类变量时，可快速将系列样本的分类值修改成用户的目标类别值。

6.7. 对样本属性(如样本名称等)的操作

除转换为自变量 **X** 不可用外，其它功能与 6.6.对属性矩阵 **y**(因变量)的操作雷同。用户亦可参考 6.5.对变量属性(如变量名称等)的操作。

第七章　文件

如前所述，本软件采用工程文件的形式管理数据分析处理的结果，包括数据、图形，以及模型等。用户可以文件的形式新建、保存和打开工程，被打开的工程可完整呈现保存时的结果。

用户点击文件菜单，则出现如下图所示的界面，亦可参考 4.2.1.详述的功能列表。

接下来依次介绍上图中的各个功能。

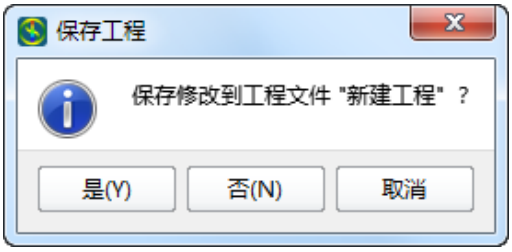
7.1. 新建工程

新建一个空白工程。

操作步骤:

步骤 1: 选择**文件**菜单。

步骤 2: 在菜单中单击**新建工程**菜单项(或直接使用快捷键 Ctrl + N),若当前已有被打开的工程，则弹出如下对话框提示用户是否保存该工程，如下图所示:



点击**是**，则保存当前工程，再新建一个名为“新建工程”的工程。保存工程的操作请查看本章内容之**保存工程**。点击**否**，则对当前工程不做任何处理而直接关闭，再新建一个名为“新建工程”的工程。点击**取消**，则取消新建操作，并关闭提示对话框。

本软件在同一时间，可且仅可使用一个工程。

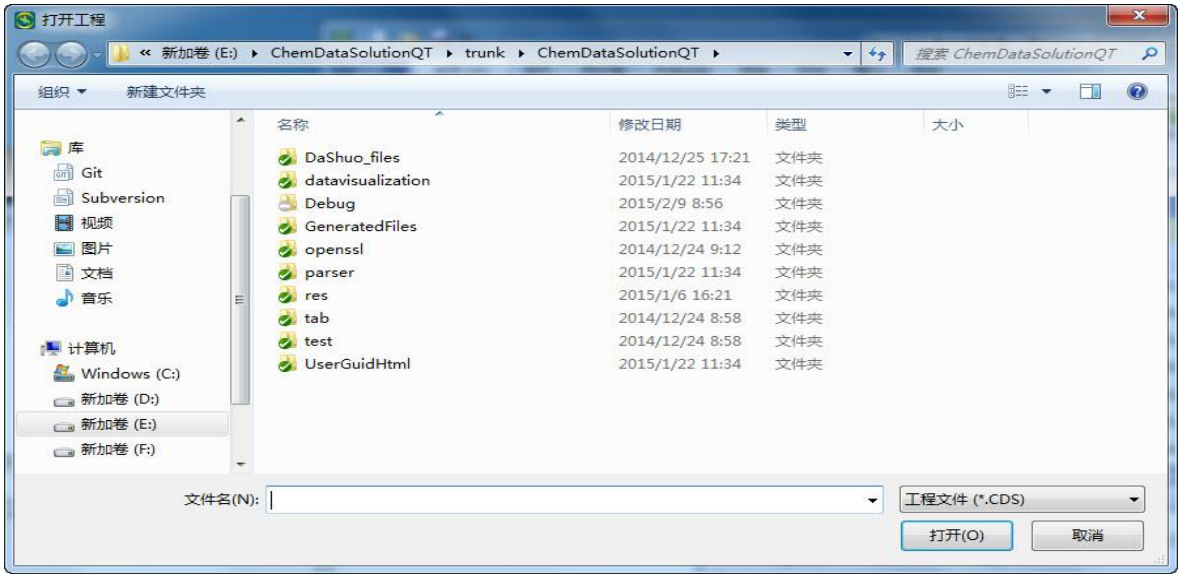
7.2. 打开工程

打开一个已有工程。

操作步骤:

步骤 1: 选择**文件**菜单。

步骤 2: 在菜单中单击**打开工程**菜单项(或直接使用快捷键 Ctrl + O)，如果当前已有被打开工程，则同样显示图所示的弹出如下对话框，其后的操作请参见 7.1.。进入打开工程页面，则显示如下图所示的界面。



步骤 3: 找到要打开的工程的路径，并点击打开即可。点击取消，则取消打开操作，并关闭提示对话框。

7.3. 保存工程

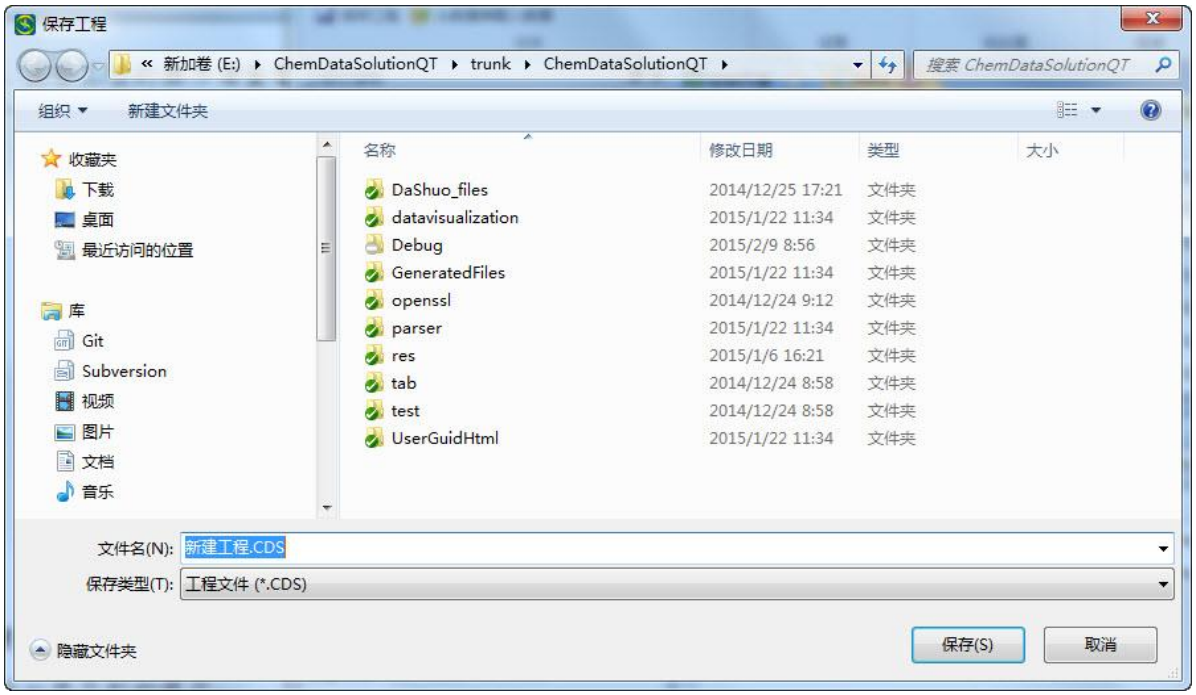
保存当前工程。

操作步骤:

步骤 1: 选择**文件**菜单。

步骤 2: 在菜单中单击**保存工程**菜单项(或直接使用快捷键 Ctrl + S)。若当前工程是已经存在的工程，则程序即时开始保存动作，若保存成功，则若干秒后在系统状态栏给出如下图所示的提示信息:

若为新建工程，则弹出如下图所示的保存工程对话框:



保存的默认文件名为新建工程.CDS，用户可自定义任意文件名。再选择保存工程的文件路径并点击保存按钮，即开始保存操作。若保存成功，则同样显示图所示的信息。

7.4. 工程另存为

将当前工程另存到一个新的文件夹路径，或另存为一个新的文件名。具体操作与 7.3.相同。

7.5. 关闭工程

关闭当前工程。若未保存对当前工程的修改，则关闭前提示用户是否保存修改。

操作步骤:

步骤 1: 选择**文件**菜单。

步骤 2: 在菜单中单击**关闭工程**菜单项，如果当前工程有修改，则关闭前将提示用户是否保存当前工程。

步骤 3: 接下来的操作请参考本章内容之保存工程。

7.6. 打印

打印当前活动窗口中的内容。

操作步骤:

步骤 1: 选择**文件**菜单。

步骤 2: 在菜单中单击**打印**菜单项(或直接使用快捷键 Ctrl + P)，弹出打印设置对话框:

步骤 3: 点击**打印**(确保打印机连接正确)，则开始打印当前活动窗口中的内容。当前活动窗口是指当前主窗口。

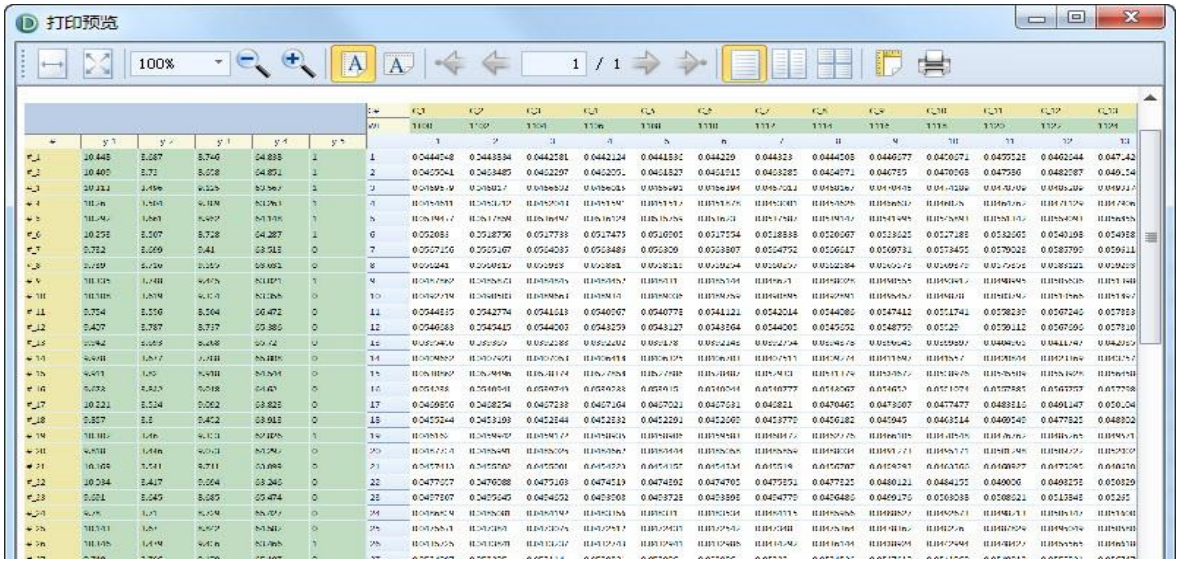
7.7. 打印预览

预览打印效果。

操作步骤:

步骤 1: 选择**文件**菜单。

步骤 2: 在菜单中单击**打印预览**菜单项，即可预览打印效果，如下图所示。对话框的上侧工具栏提供当前预览效果图的系列工具，如放大缩小，横向纵向显示，以及打印等。



7.8. 退出

退出，关闭软件。

操作步骤:

步骤 1: 选择**文件**菜单。

步骤 2: 在菜单中单击**退出**菜单项，若当前工程有修改，则退出前将提示是否保存当前工程。

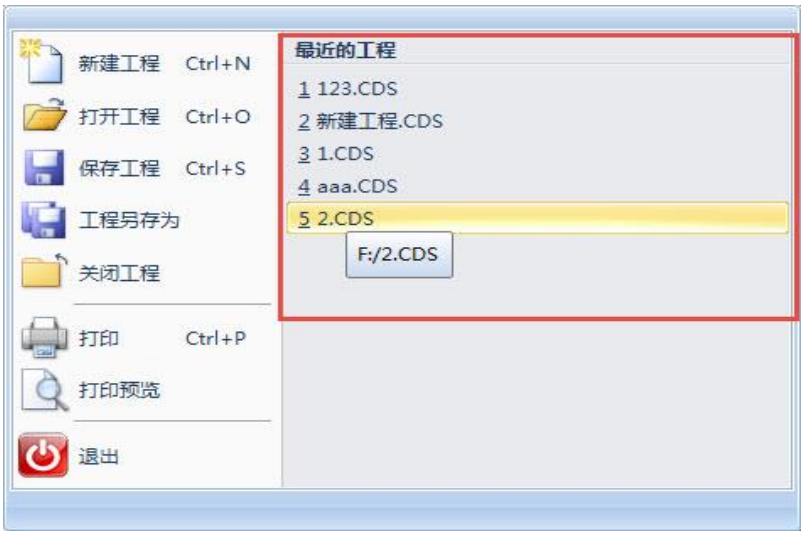
步骤 3: 接下来的操作请参考本章内容之**保存工程**。

7.9. 最近的工程

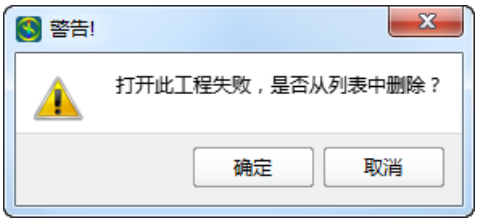
显示最近打开过的工程文件名称，可直接点击打开该工程。

操作步骤:

步骤 1: 选择**文件**菜单，即可看到最近工程列表(见下表)，将鼠标放在列表文件上可以看到该文件的完整文件路径。



步骤 2: 单击要打开的工程文件，若该工程文件已经不存在，则将出现如下对话框提示用户:



点击**确定**则将此文件从**最近的工程**列表中删除，点击**取消**则不操作。若该工程文件确实存在，则将被打开。

第八章　主页

主页是本软件关键入口，通过该部分可实现数据处理的整个流程，从新建工程到数据导入，从参数设置到新建算法流，从应用算法流分析处理数据到报表，从用户偏好设置到用户向导等。

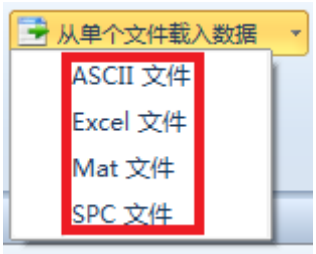
点击主页菜单，出现如下图所示的功能:

其中，新建工程，打开工程，保存工程功能及操作请参考上一章文件章节之新建工程，打开工程，保存工程；用户向导，关于我们功能及操作请参考帮助章节之用户向导(见 15.5.)和关于我们(见 15.6.)。本章节重点介绍主页中与数据处理相关的功能，接下来依次进行介绍。

8.1. 载入数据

8.1.1. 从单个文件载入数据

目前本软件支持导入的文件类型包括 4 种，点击主页中**从单个文件载入数据**，如下图所示:



我们将不断扩充软件所能载入的数据类型，并可根据客户需求，个性化扩展被载入数据类型。

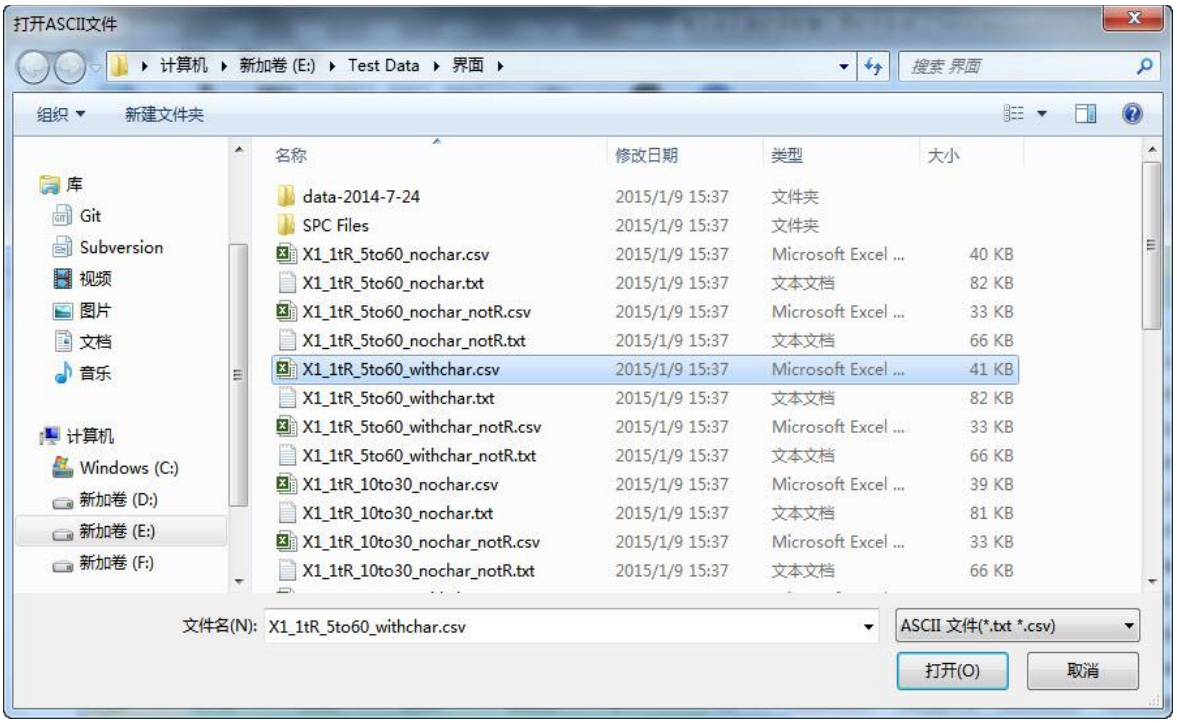
8.1.1.1. 载入 ASCII 文件

ASCII 文件是指仅含用标准 ASCII 字符集编码的字符和数据构成的文本文件，只含有字母、数字和常见的符号。字处理文件、批处理文件和源语言程序等文本文件通常都是 ASCII 文件。

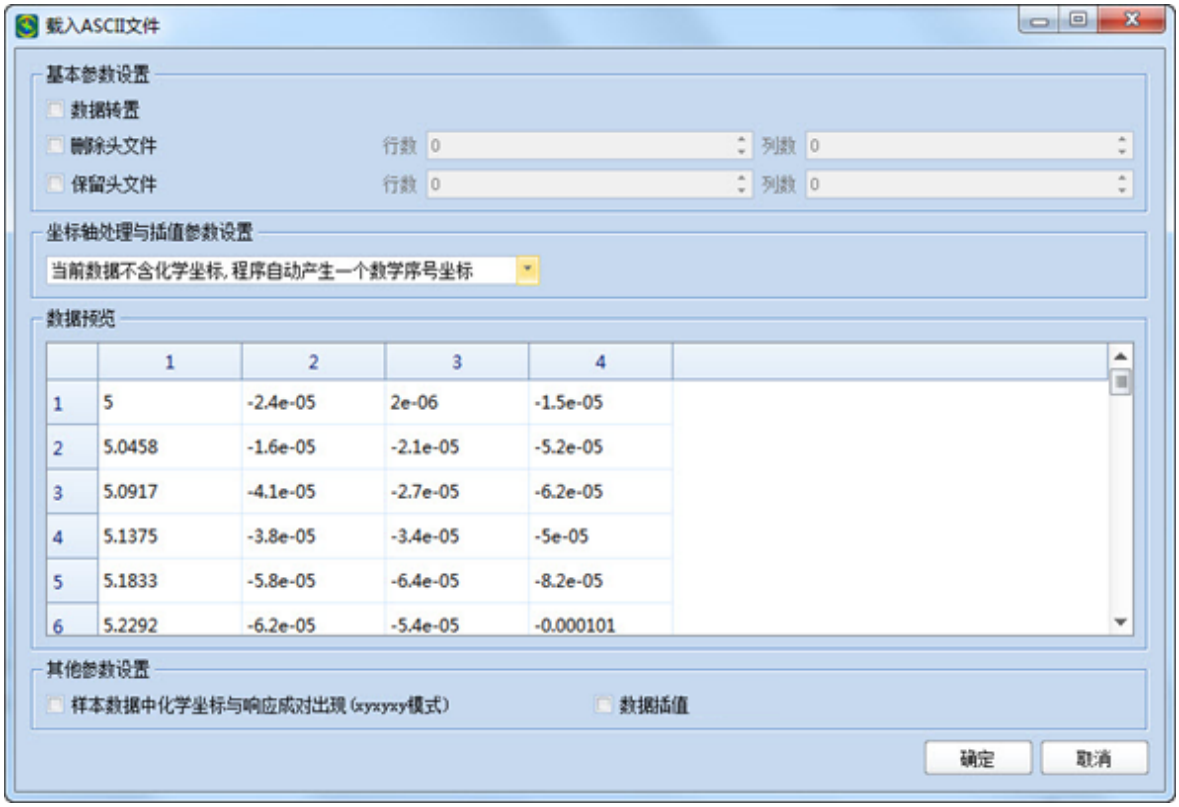
通常地，化学或生物数据均可表达或转换成 ASCII 文件，比如绝大多数科学仪器均可将原始数据格式转换成该类型格式，从而导入系统中进行分析处理。

操作步骤:

步骤 1: 点击**主页** -> **从单个文件载入数据** -> **ASCII 文件**，弹出如下对话框:



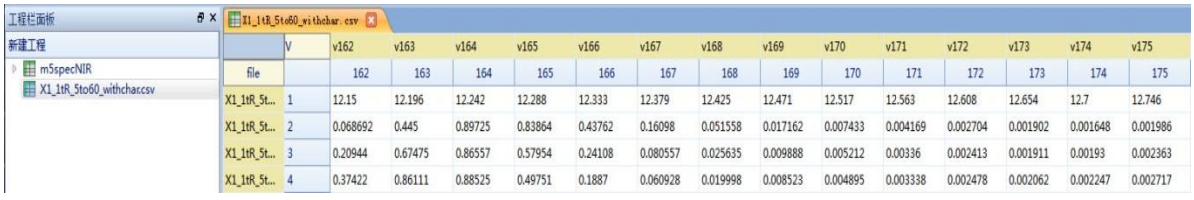
步骤 2: 选择需要载入的 ASCII 文件，点击**取消**，将取消操作并关闭对话框；点击**打开**则弹出如下对话框:



上图中的对话框可分为四部分，即基本参数设置，坐标轴处理与插值参数设置，数据预览，其他参数设置。对每个部分的具体操作及含义请参考本章节之从文件夹载入数据。

特别需要提到的是，其他参数设置中数据插值处理复选框，仅对化学坐标有效，其作用是将化学坐标插成等间隔的形式。

步骤 3: 点击**确定**，即开始数据载入，结果如下图。点击**取消**，则取消操作并关闭对话框。



数据载入完成后，作为一个新的基本数据节点存在于项目导航栏中。若需同时载入

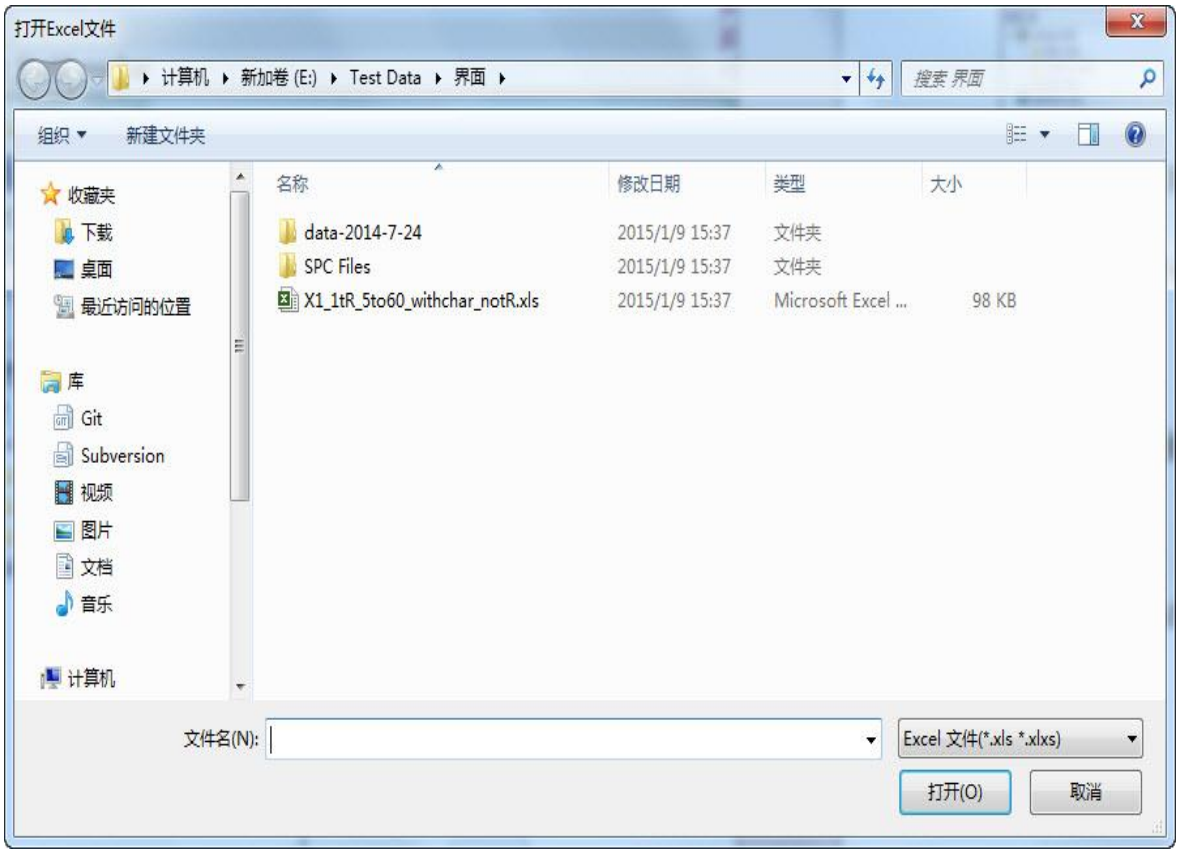
多个文件中的数据，则可使用 8.1.2.文件夹批载入数据功能，或自行先将多个数据文件整理合并后再导入。

8.1.1.2. 载入 Excel 文件

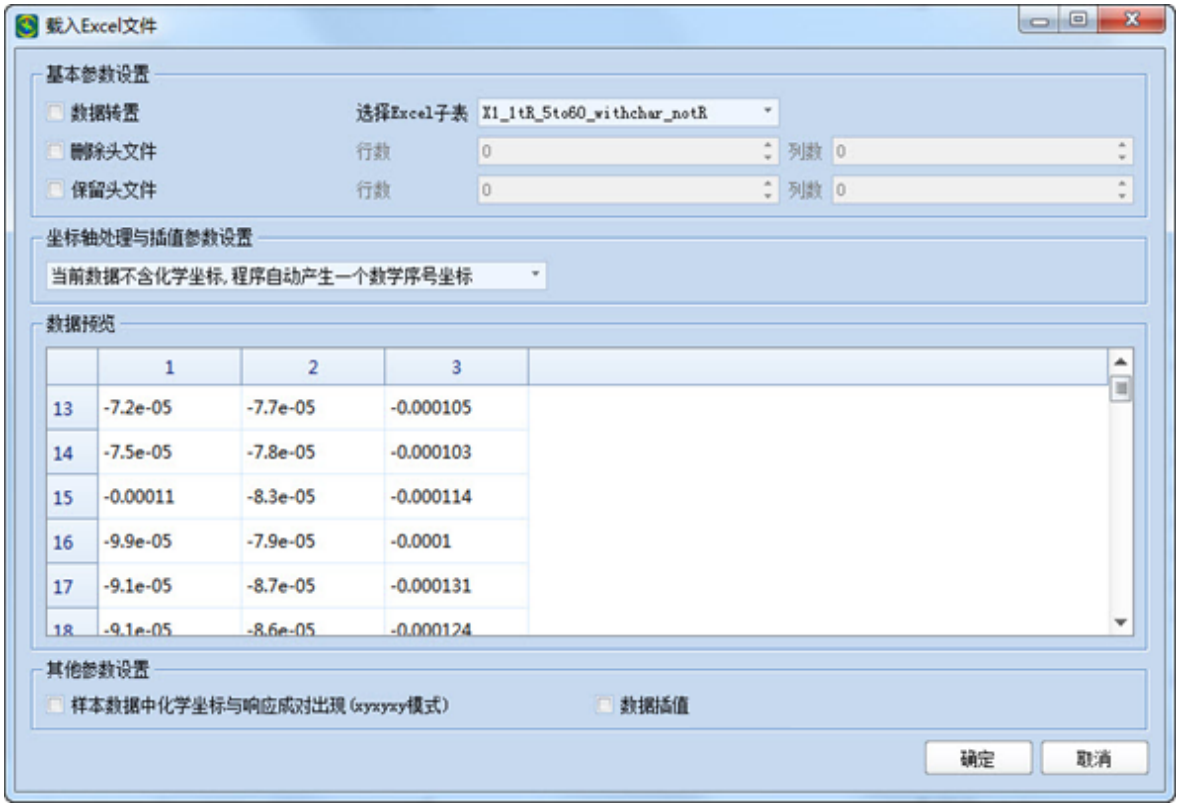
Excel 文件以表格的形式呈现，亦是本软件所处理数据的主要展现形式之一，支持.xls 和 xlsx 二种文件后缀格式。

操作步骤:

步骤 1: 点击**主页** -> **从单个文件载入数据** -> **Excel 文件**，弹出如下对话框:



步骤 2: 选择要载入的 Excel 文件，点击**取消**，将取消操作并关闭对话框；点击**打开**则弹出如下对话框:



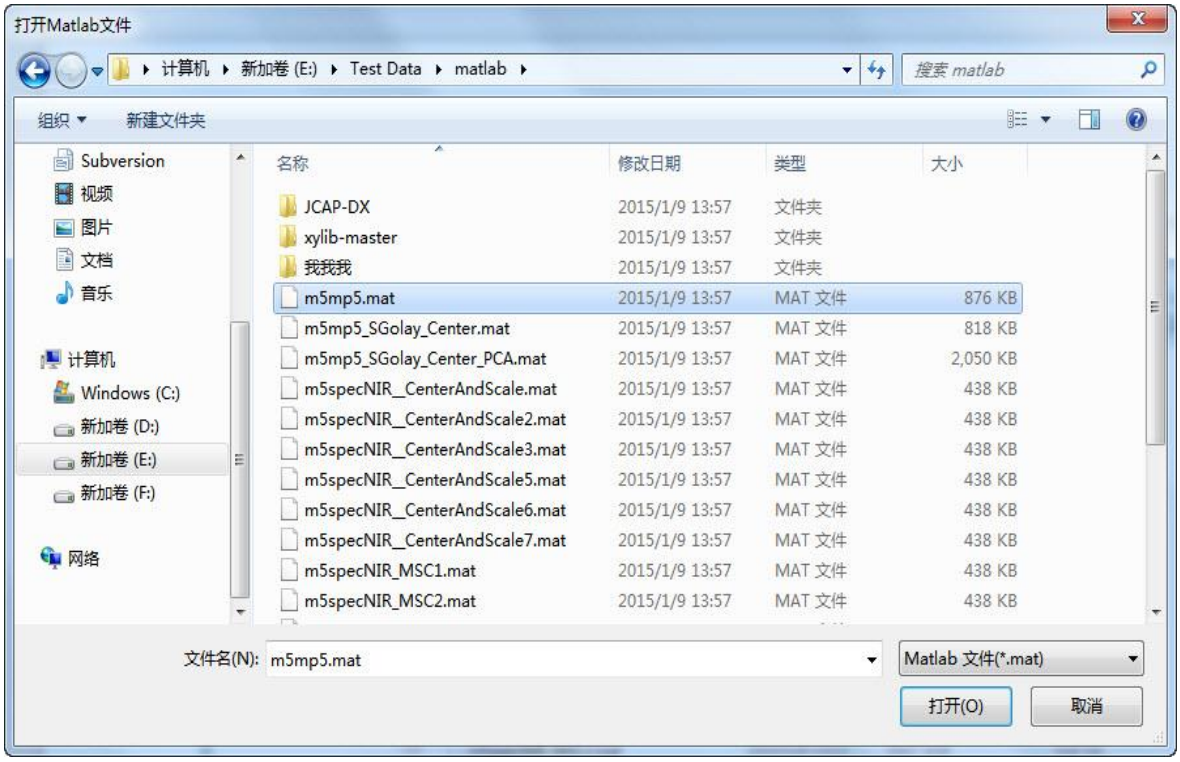
其后的操作步骤与 8.1.1.1.雷同，差异在于针对 Excel 文件，用户可选择性加载其子表。其他信息亦与 8.1.1.相同。

8.1.1.3. 载入 Mat 文件

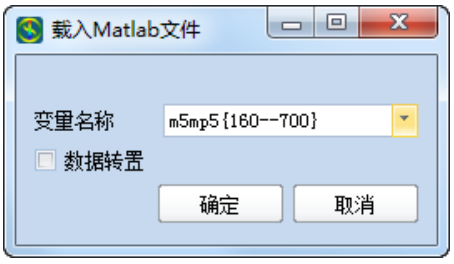
Mat 文件为 Matlab 自带的一种的数据存储格式，实因 Matlab 在数据分析中的地位，该文件格式具有一般性。

操作步骤:

步骤 1: 点击**主页** -> **从单个文件载入数据** -> **Matlab 文件**，弹出如下对话框:



步骤 2: 选择需要载入的 Mat 文件，点击**取消**，将取消操作并关闭对话框；点击**打开**则弹出如下对话框:



一个 Mat 文件中可包括多个数据变量，因此上图中提示用户选择需要载入的变量名称，同时显示变量长度大小。用户可通过勾选数据转置复选框，实现对所选数据矩阵的转置操作。其后的操作步骤与 8.1.1.1.雷同。

若数据已被整理成 Mat 文件，通常应已包含多个样本数据。

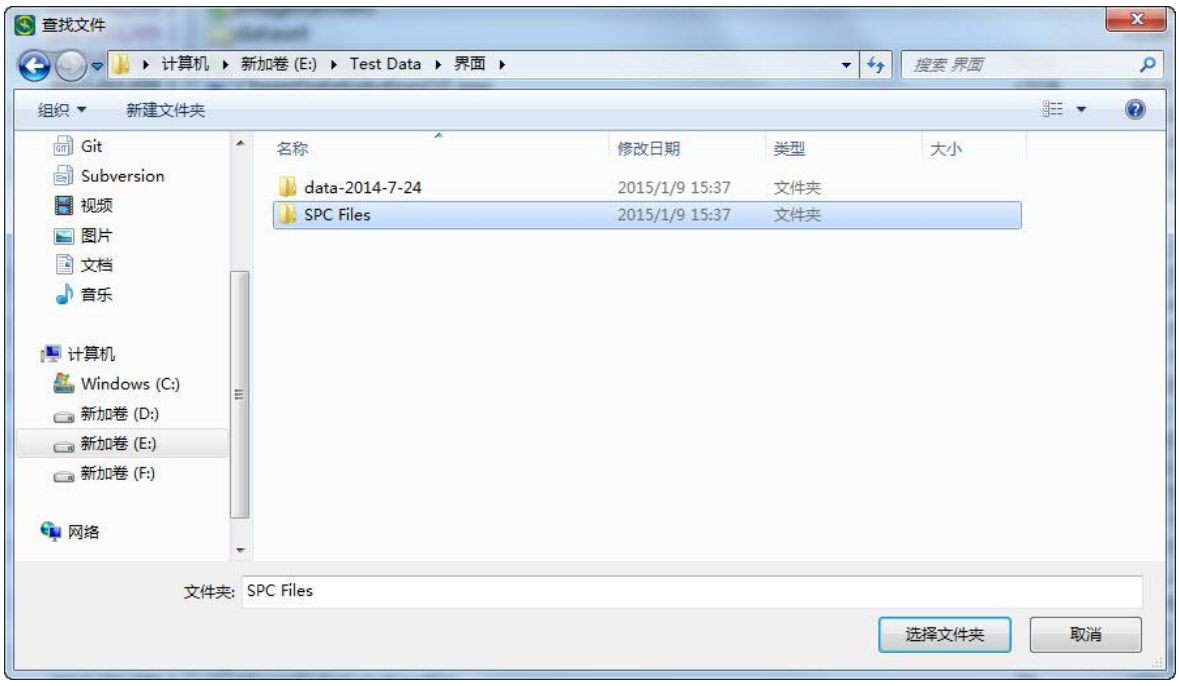
8.1.1.4. 载入 SPC 文件

SPC 是 Spectroscopy 的缩写，大部分光谱仪器所产生的数据，可转换为该种文件格式。

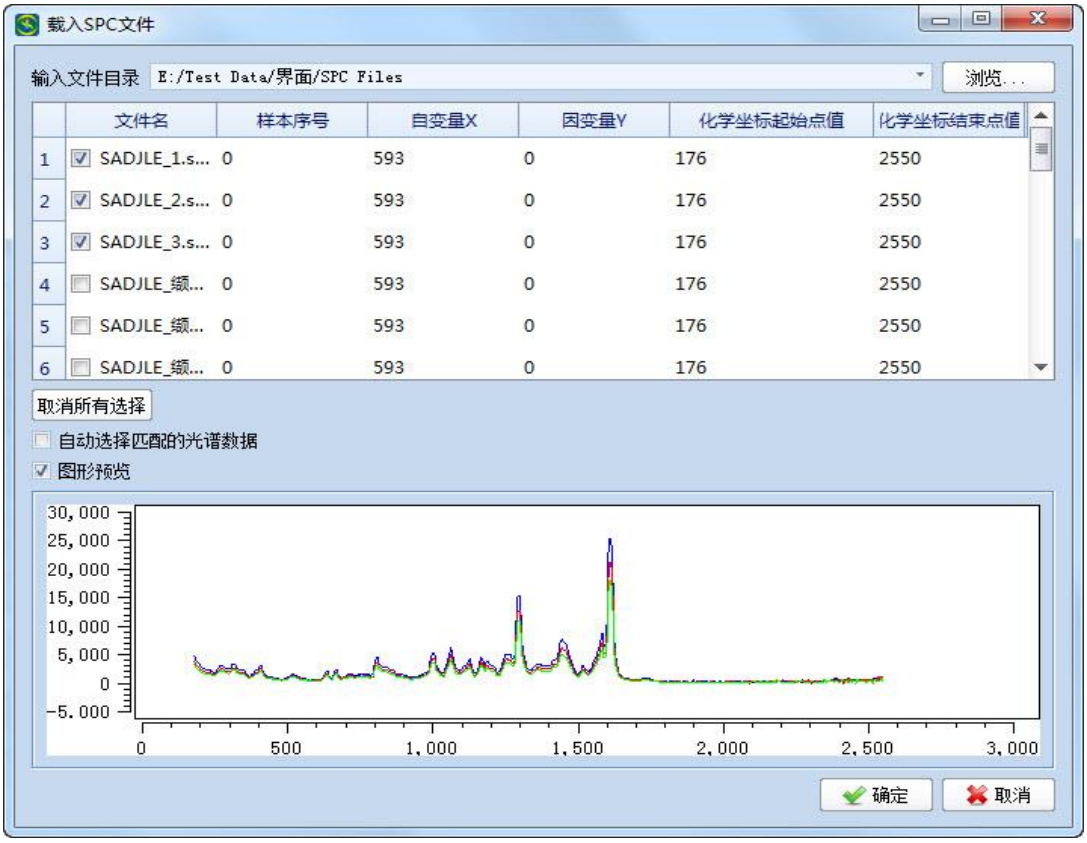
操作步骤:

步骤 1: 点击**主页** -> **从单个文件载入数据** -> **SPC 文件**，弹出如下对话框:

步骤 2: 用户可在输入文件目录框中，直接输入要载入的文件目录，亦可通过点击按钮 浏览... ，弹出如下对话框以选择 SPC 文件所在目录，如下图所示。



步骤 3: 选择需要载入的文件(通过选中文件名列表复选框选择)，选中**图形预览**复选框可对当前所选的文件数据进行预览，如下图所示。



选中自动选择匹配的光谱数据复选框，则系统将自动匹配符合条件的多个文件数据，并自动选中。所谓自动选择匹配，是指程序将化学坐标与最初被选中的数据文件一致的的其他文件一并找出来，并选择。若化学坐标不一致，则程序无法处理。单击取消所有选择按钮，即将所有文件名前的复选框置为非选中状态。其后的操作步骤与 8.1.1.1.雷同。

本处所述单个文件载入，实质单个文件类型的载入。事实上，在载入单个文件类型时，用户亦可通过多选，同时批量载入多个不同文件样本(Ctrl 和 Shift 键可用)。

8.1.2. 从文件夹批载入数据

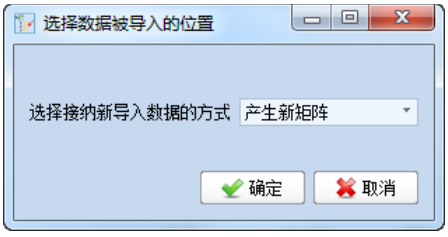
如前所述，从文件夹批载入数据是本软件的重要特色之一，可实现同时导入多个文件中的数据，即用户无需事先对数据进行任何预处理，只需依据程序的操作流程，便可完成文件

下多个数据文件的导入。

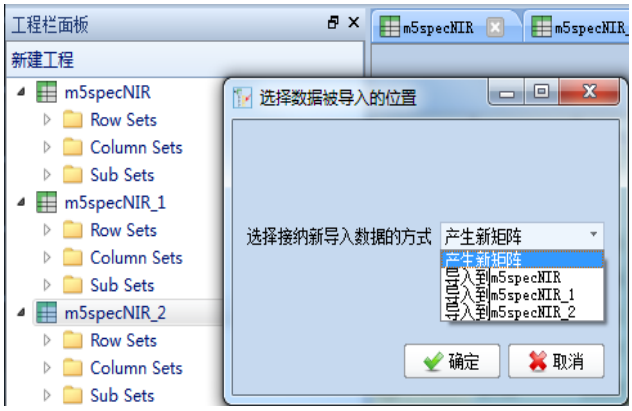
事实上，每个数据文件的数据结构极有可能不一致，比如数据长度的差异，是否含有化学坐标和字符等等(更多情形可参考 2.1.1.)。本软件针对数据的这些复杂情形，完美地提供整体解决方案，用户体验极佳。

操作步骤:

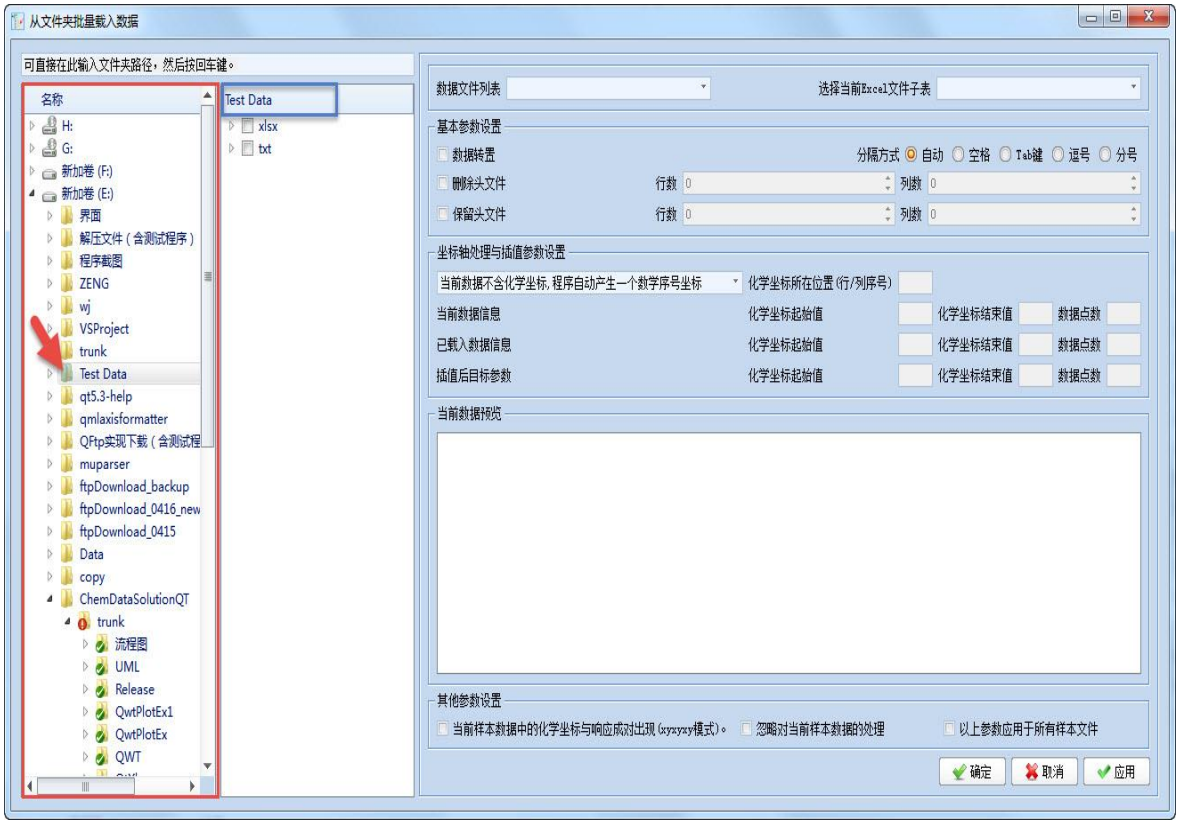
步骤 1: 点击**主页** -> **从文件夹批载入数据**，弹出如下对话框:



用户可从下拉列表中选择接纳新导入数据的方式，系统提供两种方式，即产生新矩阵和导入到工程中现有的数据中，下拉菜单如下图所示，其中的数据列表即为工程文件已经存在的基本数据表名称:



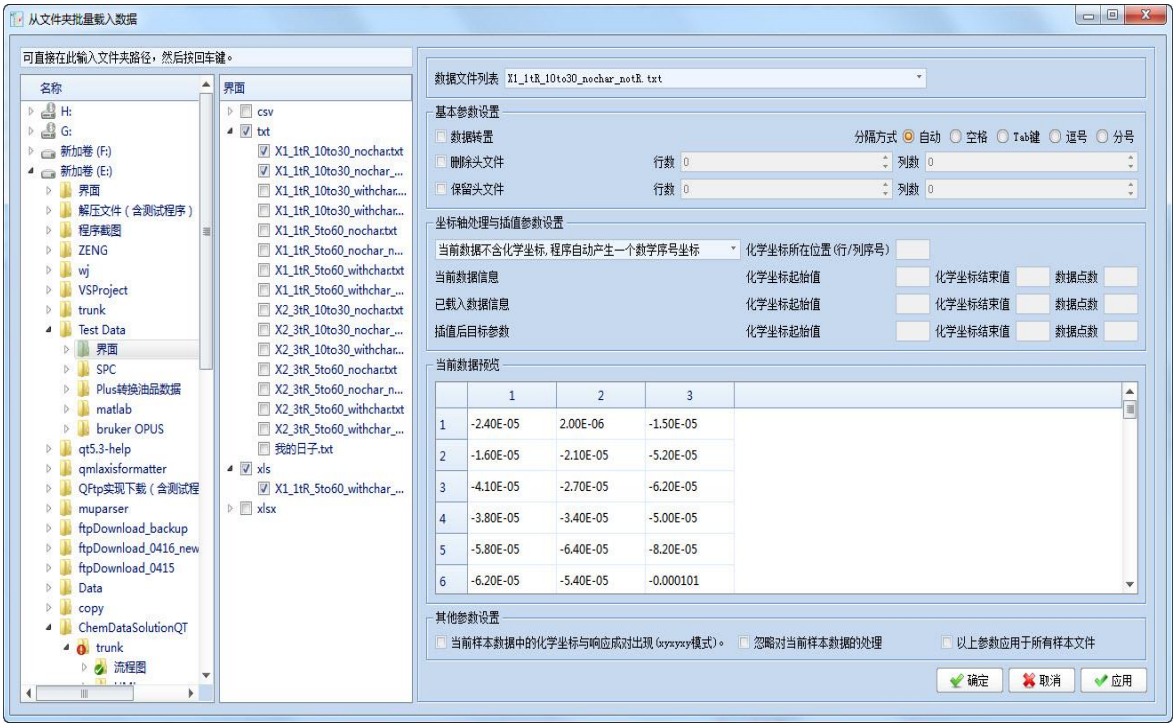
步骤 2: 点击**取消**，取消操作并关闭对话框；点击**确定**，则弹出如下对话框:



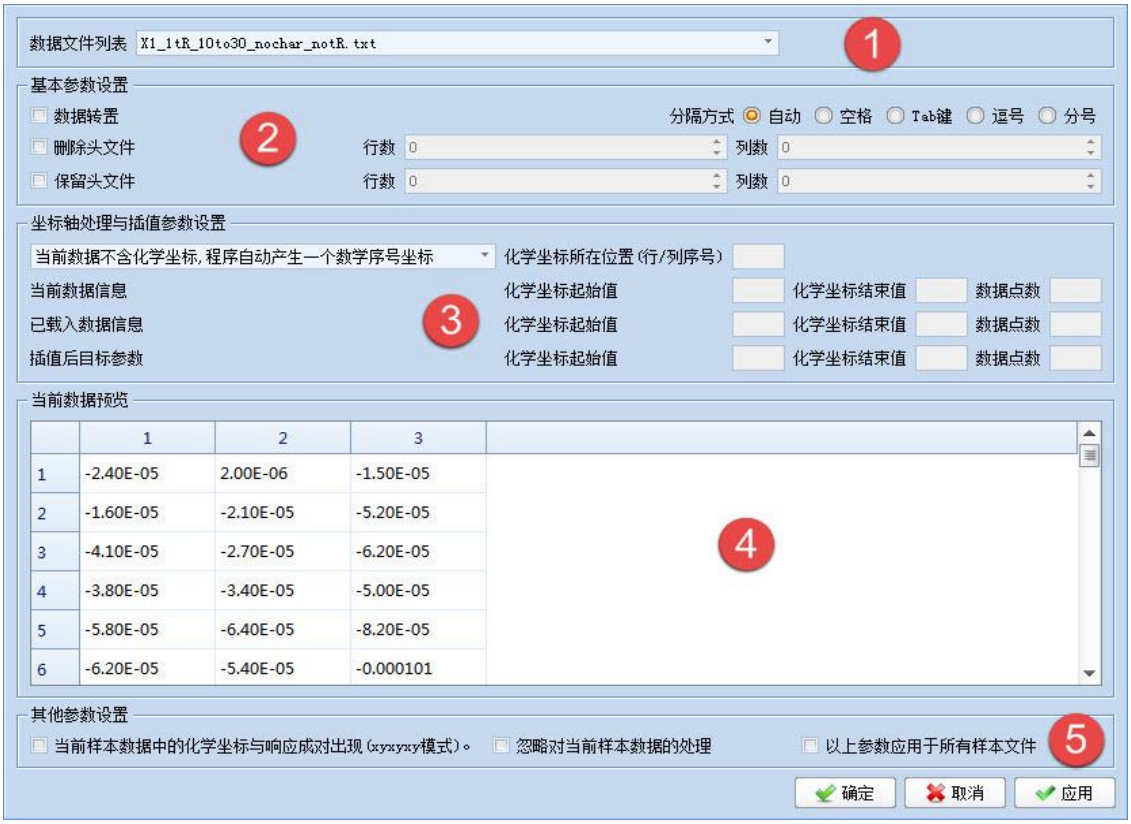
用户即可通过上图选择载入数据的文件夹，其中红色区域表示计算机文件系统；蓝色区域表示红色框中被选文件夹的名称(本例为 Test Data)。在该文件名的下方则列举出所有可载入到工程中的文件格式(本例为 xlsx 和 txt)，程序自动对这些文件进行归纳分类。

默认文件夹为用户在偏好设置中设置的路径。本软件支持批量载入 txt，CSV，xls 以及 xlsx 等文件格式。

步骤 3: 勾选文件类型前复选框，将目标文件加入到右侧数据文件列表中，如下图所示:

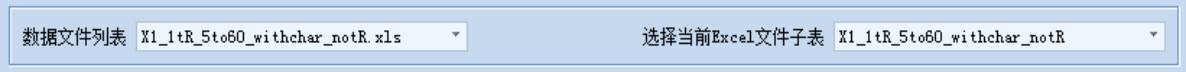


从上图可看出，用户可选择多个不同类型的文件，可部分选择某类型文件下的文件，这些数据文件均列表在右侧的数据文件列表中。将上图右侧部分定义为五个不同区域，如下图所示，再详细介绍各区域功能。



下面依次介绍各区域的功能。

1) 数据文件列表：显示被选文件夹中全部或部分被选择的数据文件，如下图所示。若当前文件为表格文件，则需进一步选择表格中的子表名称。对表格文件，默认状态自动选择第一个子表。



2) 基本参数设置：主要功能是将原始数据表格进行某些处理，以定义数据表格。

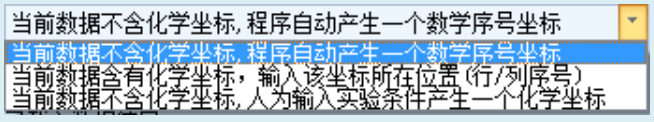
序号	功能名称	详细说明
1	数据转置	如前所述，本软件中数据矩阵的结构为行代表样本，列代表变量。而在实际数据文件中，数据结构可能出现相反的情况，即列表示样本，而行表示变量。为匹配软件的数据矩阵结构，需先将样本和变量的位置互换，即**数据转置**。用户勾选此功能，则数据在根据用户设置载入后，自动完成数据转置，放于工程导航栏中。
2	分隔方式	在不同数据文件中，数据间的分隔方式可能不同。用户可在**分隔方式**中选择待导入数据文件的数据分隔方式，程序可自动识别的分隔符类型有以下几种: ①自动(系统判断以最佳方式分隔)；②空格；③Tab 键；④逗号；⑤分号。
3	删除头文件	在数据文件中，除数据外，往往还有一些标注样本和/或变量属性的表头文字。当选中**删除头文件**时，可以通过指定删除的行数和/或列数来删除**相应数量**的表头字符。 ①表头文字可能为行表头，亦可能为列表头，或者行与列都有，用户根据需要进行选择。②删除表头操作可撤消，用户仅需将删除的行数和/或列数清零即可。
4	保留头文件	保留表头的作用在于在导入数据时，将样本和/或变量的属性一并添加到属性表

		中。用户可以自定义保留的行数和/或列数。 该功能的意义在于用户可继续使用原始数据中的样本或变量说明性信息。

3) 坐标轴处理与插值参数设置：主要功能是定义数据坐标轴，使得不同的数据样本可规范合并成数据表格，以便分析。

用户载入多个数据文件时，这些数据所含有坐标情形可能完全不同，从而导致数据无法合并，这些情形包括：①数据可能含有坐标轴，亦可能不包含坐标轴；②数据可能包含化学坐标轴，亦可能包含数学坐标轴(具体请参考 3.3.和 3.4.)；③数据所含有的化学坐标轴可能起始位置不一致，亦可能结束位置不一致(如色谱分析的起始和中止时间，光谱的检测波长范围，或质谱的 m/z 范围等。)；④数据可能添加到已经存在的基本数据表中，则与该数据的化学坐标不一致。

总之，本功能在于综合考虑各种情形，使得不同数据文件，在不同的情况下均能合理地得到合并，从而进行数据分析处理。

序号	功能名称	详细说明
1	数据有无化学坐标的下拉列表	下拉框中的内容是对当前数据文件是否包含化学坐标的一个选择，由用户根据数据本身的不同情形设定。可选择的项目包括:  下拉菜单可根据其描述具体选择。
2	化学坐标所在位置(行/列序号)	当用户选择**当前数据含有化学坐标**时，则**输入该坐标所在位置(行/列序号)**功能有效，即用户可输入具体的位置

		序号。
3	当前数据信息	当用户选择**当前数据不含化学坐标时**，则**人为输入实验条件产生一个化学坐标**功能有效。 该组参数用于设置待载入数据的信息。
4	已载入数据信息	当用户选择**接纳新导入数据的方式**为导入到已有数据时，该组参数用于显示已有基本数据表中数据的化学坐标起始值，结束值，以及数据点数信息。
5	插值后目标参数	当用户选择**当前数据含有化学坐标**，**输入该坐标所在位置(行/列序号)**或**当前数据不含化学坐标**，**人为输入实验条件产生一个化学坐标**时有效。该组参数是针对当前待导入数据和已载入数据的整体设置，用户可根据实际需求自定义该组值。 该组参数设置完成后，所有数据将依其做插值运算。

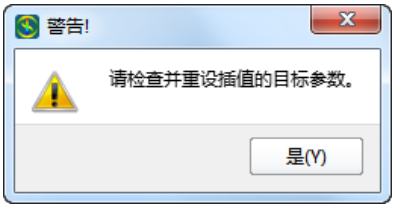
4) 当前数据预览：当前数据的预览表，根据参数设置而动态变化。

5) 其他参数设置：对不同数据文件的整体设置。

序号	**功能名称**	**详细说明**
1	当前样本数据中的化学坐标与响应成对出现(xyxyxy 模式)	设置当前样本数据中的化学坐标与响应值成对出现，即一个坐标数据对应一个响应数据。
2	忽略对当前样本数据的处理	跳过对当前的样本处理，即不会对当前样本数据做任何处理。

3	以上参数应用于所有样本文件	将对当前样本的参数，应用到数据文件列表中的所有数据文件。

步骤 4: 设置参数，若参数设置有误，系统将提示用户，如下图所示:



步骤 5: 点击**确定**或**应用**，则开始载入数据文件。点击**确定**则处理完成后关闭对话框，点击应用则可继续操作。点击取消，取消操作并关闭对话框。

用户很好地使用从文件夹载入数据功能，可实现对数据的快速批量分析，极大地节约时间。

8.1.3. 从数据库载入数据

暂略。

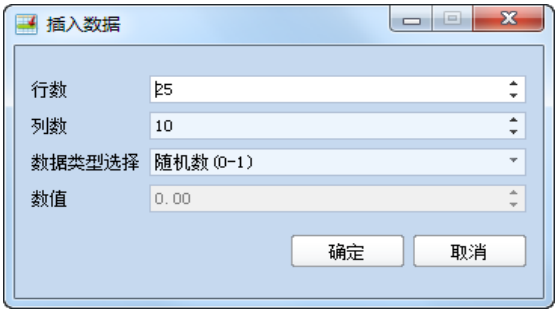
8.2. 插入数据

往工程导航栏中，插入自定义行、列和数值的数据矩阵。新插入的数据将作为新的基本数据显示。

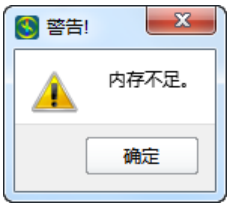
用户可通过此功能，实现为载入实际数据情况下的数据处理，即可采用插入的模拟数据，了解、学习和研究本软件提供的数据处理功能。

操作步骤:

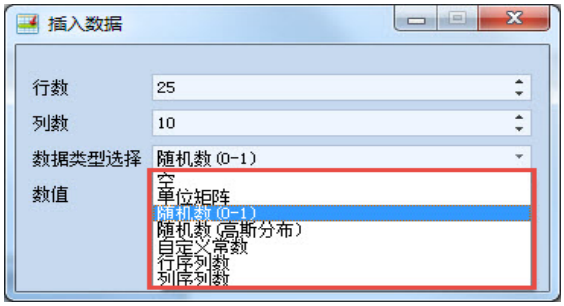
步骤 1: 点击**主页** -> **插入数据**，弹出如下对话框:



如果设置的行数或列数太大而导致内存不足时，系统将提示用户:



本软件所提供的可插入数据类型如下图所示:



下面对上图中所示的功能，一一做出介绍。

序号	数据类型	说明
1	空	产生以 nan 填充的数据表，如下图所示(示例): V Var_1 Var_2 Var_3 Var_4 Var_5 # 1 2 3 4 5 #_1 1 nan nan nan nan nan #_2 2 nan nan nan nan nan #_3 3 nan nan nan nan nan #_4 4 nan nan nan nan nan #_5 5 nan nan nan nan nan #_6 6 nan nan nan nan nan #_7 7 nan nan nan nan nan

<table>
<tr><td>2</td><td>单位矩阵</td><td>产生一个单位矩阵，如下图所示(示例):
<table>
<tr><td></td><td>V</td><td>Var_1</td><td>Var_2</td><td>Var_3</td><td>Var_4</td><td>Var_5</td></tr>
<tr><td>#</td><td></td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr>
<tr><td>#_1</td><td>1</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>#_2</td><td>2</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>#_3</td><td>3</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>#_4</td><td>4</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td></tr>
<tr><td>#_5</td><td>5</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td></tr>
<tr><td>#_6</td><td>6</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>#_7</td><td>7</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
</table></td></tr>
<tr><td>3</td><td>随机数(0-1)</td><td>产生以 0 到 1 之间随机数填充的数据表格，如下图所示(示例):
<table>
<tr><td></td><td>V</td><td>Var_1</td><td>Var_2</td><td>Var_3</td><td>Var_4</td><td>Var_5</td></tr>
<tr><td>#</td><td></td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr>
<tr><td>#_1</td><td>1</td><td>0.9676096...</td><td>0.4033077...</td><td>0.2359319...</td><td>0.4456959...</td><td>0.4241819...</td></tr>
<tr><td>#_2</td><td>2</td><td>0.5497027...</td><td>0.3445404...</td><td>0.0891464...</td><td>0.3829964...</td><td>0.3648084...</td></tr>
<tr><td>#_3</td><td>3</td><td>0.4624333...</td><td>0.9646739...</td><td>0.0695108...</td><td>0.4190656...</td><td>0.0413334...</td></tr>
<tr><td>#_4</td><td>4</td><td>0.8861662...</td><td>0.9393765...</td><td>0.8123440...</td><td>0.7589480...</td><td>0.2961684...</td></tr>
<tr><td>#_5</td><td>5</td><td>0.8799988...</td><td>0.7086120...</td><td>0.6383088...</td><td>0.4775804...</td><td>0.5110272...</td></tr>
<tr><td>#_6</td><td>6</td><td>0.1566446...</td><td>0.7177014...</td><td>0.1080253...</td><td>0.5564976...</td><td>0.5711566...</td></tr>
<tr><td>#_7</td><td>7</td><td>0.9981832...</td><td>0.5940795...</td><td>0.5812895...</td><td>0.8492613...</td><td>0.7373939...</td></tr>
</table></td></tr>
<tr><td>4</td><td>随机数(高斯分布)</td><td>产生高斯分布随机数数据矩阵，如下图所示(示例):
<table>
<tr><td></td><td>V</td><td>Var_1</td><td>Var_2</td><td>Var_3</td><td>Var_4</td><td>Var_5</td></tr>
<tr><td>#</td><td></td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr>
<tr><td>#_1</td><td>1</td><td>-0.224470...</td><td>-0.712047...</td><td>1.1798406...</td><td>-0.214726...</td><td>-0.395780...</td></tr>
<tr><td>#_2</td><td>2</td><td>-0.211586...</td><td>0.6428712...</td><td>1.4396413...</td><td>0.4698953...</td><td>-1.351634...</td></tr>
<tr><td>#_3</td><td>3</td><td>-0.575043...</td><td>-0.270286...</td><td>-0.521163...</td><td>-0.439738...</td><td>1.0382255...</td></tr>
<tr><td>#_4</td><td>4</td><td>0.5135690...</td><td>-0.608431...</td><td>0.5730389...</td><td>-0.316726...</td><td>0.6347756...</td></tr>
<tr><td>#_5</td><td>5</td><td>0.4577976...</td><td>1.5099991...</td><td>1.6715690...</td><td>0.2524684...</td><td>-0.977262...</td></tr>
<tr><td>#_6</td><td>6</td><td>0.4406758...</td><td>0.8788714...</td><td>-1.700597...</td><td>0.5891809...</td><td>-1.174246...</td></tr>
<tr><td>#_7</td><td>7</td><td>-0.330484...</td><td>-0.430252...</td><td>0.0240024...</td><td>-0.552066...</td><td>-0.151499...</td></tr>
</table></td></tr>
<tr><td>5</td><td>自定义常数</td><td>产生以用户自定义常数填充的数据矩阵，如下图所示(示例):
<table>
<tr><td></td><td>V</td><td>Var_1</td><td>Var_2</td><td>Var_3</td><td>Var_4</td><td>Var_5</td></tr>
<tr><td>#</td><td></td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr>
<tr><td>#_1</td><td>1</td><td>123</td><td>123</td><td>123</td><td>123</td><td>123</td></tr>
<tr><td>#_2</td><td>2</td><td>123</td><td>123</td><td>123</td><td>123</td><td>123</td></tr>
<tr><td>#_3</td><td>3</td><td>123</td><td>123</td><td>123</td><td>123</td><td>123</td></tr>
<tr><td>#_4</td><td>4</td><td>123</td><td>123</td><td>123</td><td>123</td><td>123</td></tr>
<tr><td>#_5</td><td>5</td><td>123</td><td>123</td><td>123</td><td>123</td><td>123</td></tr>
<tr><td>#_6</td><td>6</td><td>123</td><td>123</td><td>123</td><td>123</td><td>123</td></tr>
<tr><td>#_7</td><td>7</td><td>123</td><td>123</td><td>123</td><td>123</td><td>123</td></tr>
</table></td></tr>
<tr><td>6</td><td>行序列数</td><td>产生以行序列数填充的数据矩阵，如下图所示(示例):</td></tr>
</table>

<table>
<tr><td></td><td></td><td>
<table>
<tr><th></th><th>V</th><th>Var_1</th><th>Var_2</th><th>Var_3</th><th>Var_4</th><th>Var_5</th></tr>
<tr><th>#</th><th></th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th></tr>
<tr><td>#_1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td></tr>
<tr><td>#_2</td><td>2</td><td>2</td><td>2</td><td>2</td><td>2</td><td>2</td></tr>
<tr><td>#_3</td><td>3</td><td>3</td><td>3</td><td>3</td><td>3</td><td>3</td></tr>
<tr><td>#_4</td><td>4</td><td>4</td><td>4</td><td>4</td><td>4</td><td>4</td></tr>
<tr><td>#_5</td><td>5</td><td>5</td><td>5</td><td>5</td><td>5</td><td>5</td></tr>
<tr><td>#_6</td><td>6</td><td>6</td><td>6</td><td>6</td><td>6</td><td>6</td></tr>
<tr><td>#_7</td><td>7</td><td>7</td><td>7</td><td>7</td><td>7</td><td>7</td></tr>
</table>
</td></tr>
<tr><td>7</td><td>列序列数</td><td>产生以列序列数填充的数据矩阵，如下图所示(示例):
<table>
<tr><th></th><th>V</th><th>Var_1</th><th>Var_2</th><th>Var_3</th><th>Var_4</th><th>Var_5</th></tr>
<tr><th>#</th><th></th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th></tr>
<tr><td>#_1</td><td>1</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr>
<tr><td>#_2</td><td>2</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr>
<tr><td>#_3</td><td>3</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr>
<tr><td>#_4</td><td>4</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr>
<tr><td>#_5</td><td>5</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr>
<tr><td>#_6</td><td>6</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr>
<tr><td>#_7</td><td>7</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr>
</table>
</td></tr>
</table>

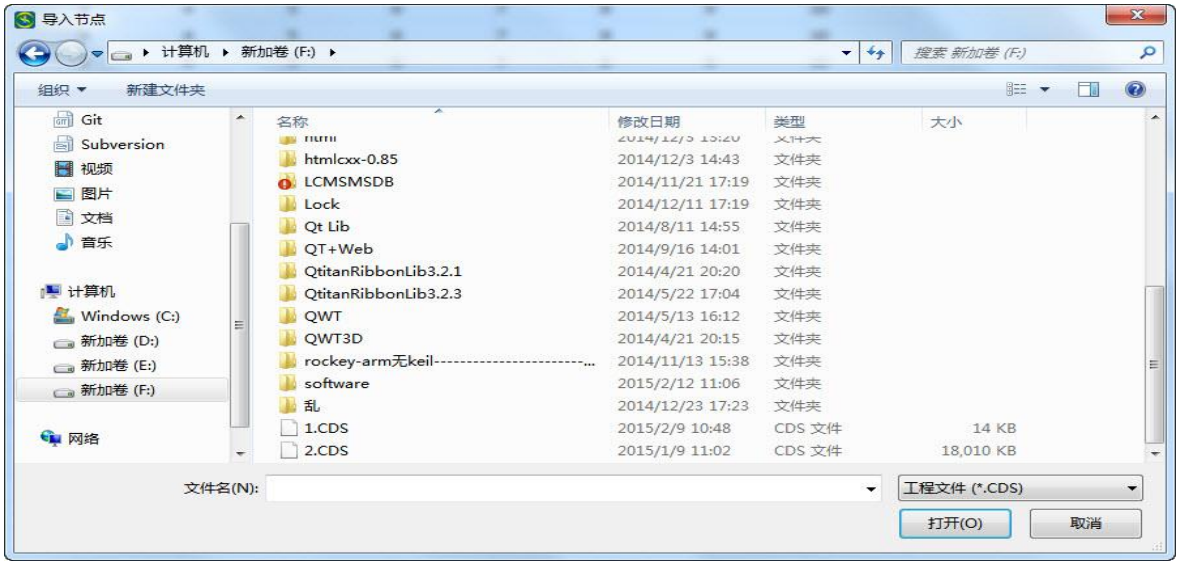
步骤 2: 点击**确定**，即开始插入数据，点击**取消**，取消操作并关闭对话框。

8.3. 导入节点

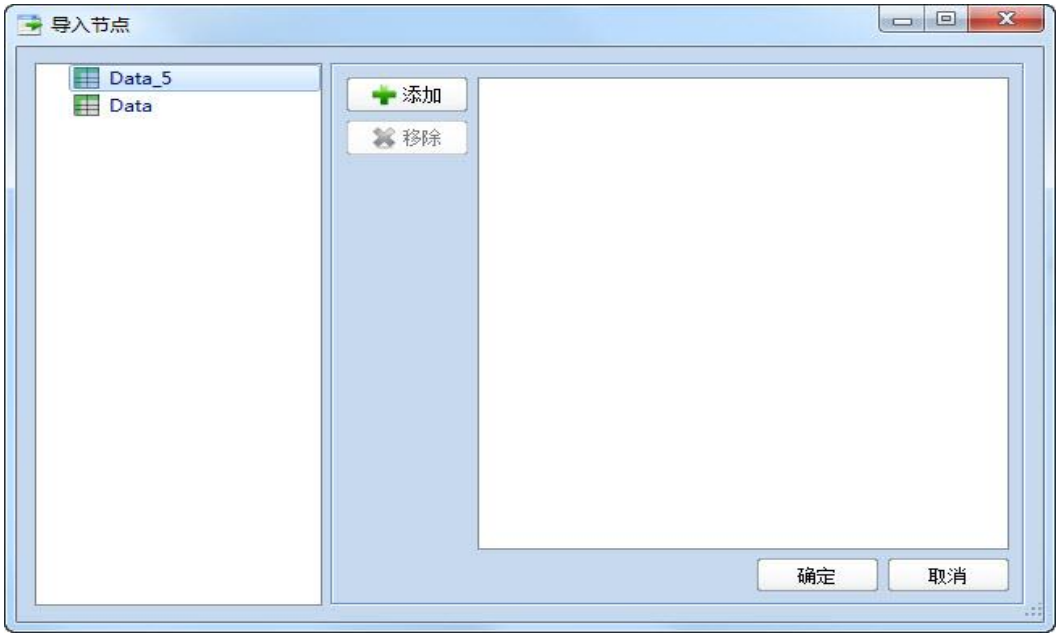
导入用户已经保存的一个或多个节点到当前工程中。用户可通过此功能导入已建立的模型到当前工程中，从而使用该模型分析新的数据，完成对新数据的验证与预测等。

操作步骤:

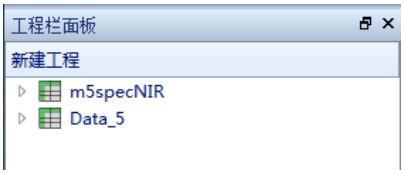
步骤 1: 点击**主页** -> **导入节点**，弹出如下对话框:



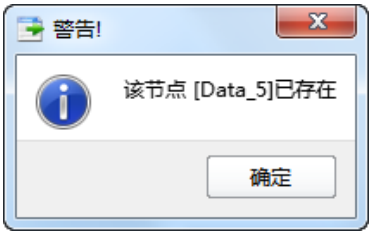
步骤 2: 选择一个工程文件(后缀为.CDS)，点击**打开**，弹出如下对话框，以列表形式显示该工程中所包含的所有节点，以供用户选择，如本例中的 Data_5 和 Data。



步骤 3: 用户添加需要导入的节点，点击**确定**，即可将被选节点导入到当前工程中。导入成功后便可在当前工程导航栏中看到新导入的节点，如下图所示(被导入节点为 Data_5):



若当前工程中已经含有同名字的节点，则新节点不能再被导入，显示如下图所示对话框以提示用户；点击**取消**，则取消操作并关闭对话框。



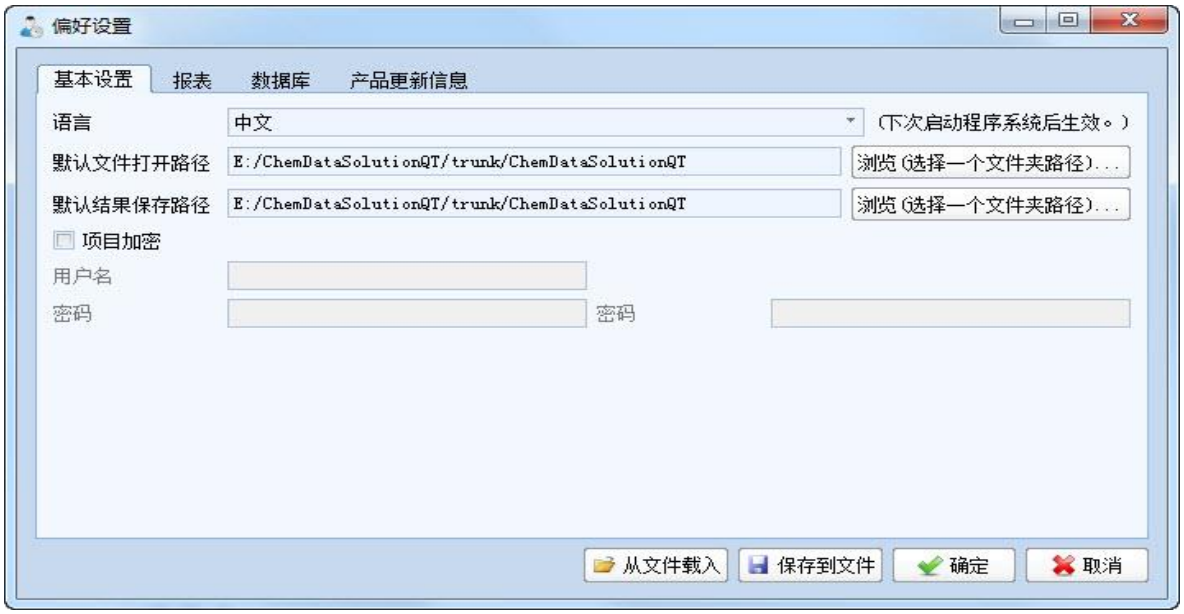
8.4. 设置

8.4.1. 偏好设置

使用该功能，用户可根据个人使用习惯和偏好对系统的工作环境进行设置，在实际使用中则自动出现用户设置的内容，当然用户亦可继续修改。

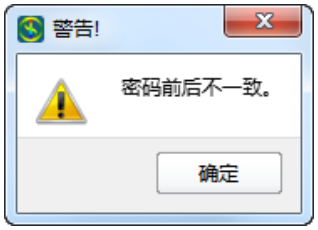
操作步骤:

步骤 1: 点击**主页** -> **偏好设置**，弹出如下对话框:



该界面共有四个标签页，下面依次介绍各个标签页的内容。

1) 基本设置: 对工程中常用和基本的使用方式进行设置。特别地，用户可对项目加密，但用户前后二次输入的须一致，否则，系统将给出错误提示，如下图所示:



2) 报表：本软件提供便捷的报表功能。用户可通过如下图所示的界面，实现报表中表头信息的预定义，即通过此界面定义的报表头文字，将作为报表产生时的默认信息。

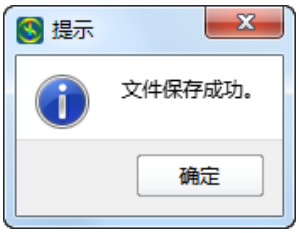
3) 数据库：本软件提供数据库管理功能。用户可通过基本数据表与数据库中数据的交互导入与导出，实现丰富的数据处理和比较功能，详情请参见 2.1.3.。

用户通过如下界面所保存的信息，将作为往数据库中添加样本的默认说明性信息。当然，用户亦可根据实际情况编辑和修改。

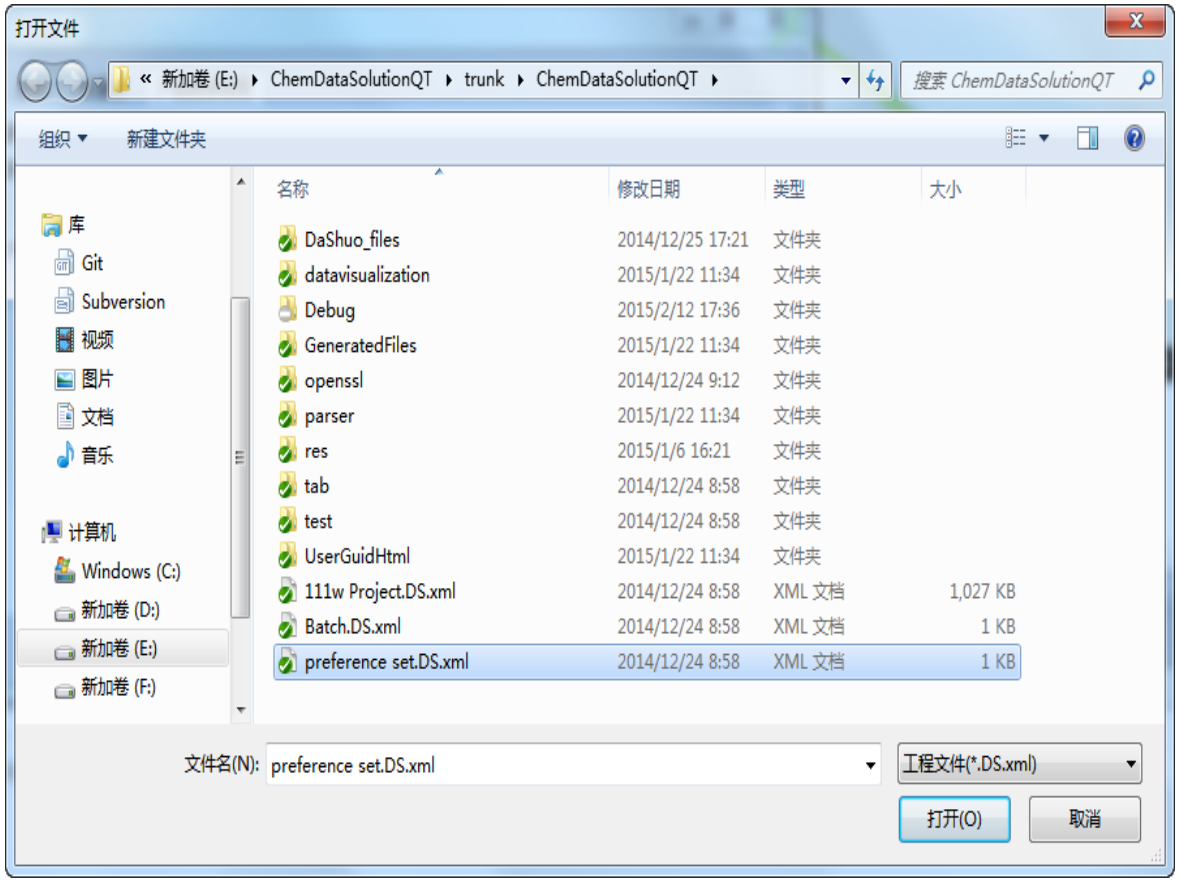
4) 产品更新信息：本部分实现本公司及产品与用户的交互，以帮助用户更好地了解本公司及产品的信息。

步骤 2：参数设置完成后，点击**确定**即可完成偏好设置。点击**取消**，则取消操作并关闭对话框。

本软件提供将当前设置保存到文件的功能，以方便下次使用可从文件直接载入这些设置，而不必手动重新输入。点击**保存到文件**即可将当前设置保存到文件中。保存成功后将提示用户，如下图所示。保存文件的格式为.xml。



点击**从文件载入**，则弹出如下对话框:



选择一个偏好设置文件，点击**打开**即可把文件中的设置内容填充到图所示的界面上。

8.4.2. 参数设置

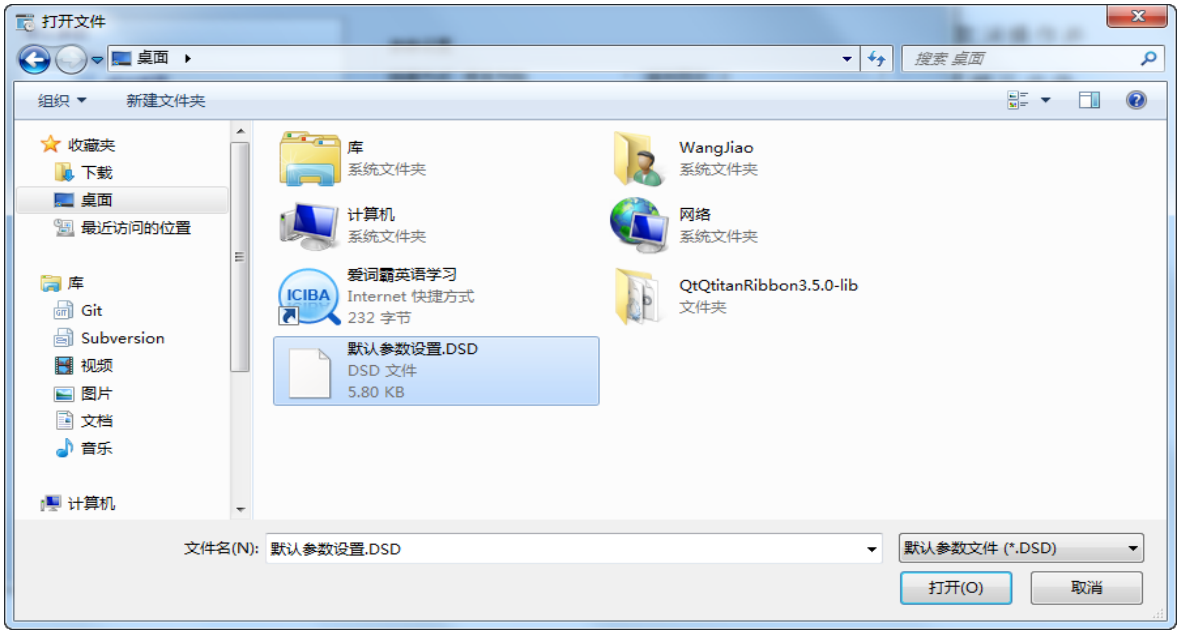
参数设置是本软件提高用户用户体验的另一特色，实现预处理、变量选择和建模等数据处理方法中参数的集中设置，用户仅需在该功能页面中对相应的方法参数进行预设置，则在实际使用这些数据处理方法时，这些参数设置将自动作为默认值填充到参数设置对话框中，以方便用户在每次执行数据处理方法时，均需设置参数的麻烦。

操作步骤:

步骤 1: 点击**主页** -> **参数设置**，弹出如下对话框:

单击左侧方法节点，右侧便显示该方法参数，用户可修改这些参数值。

步骤 2: 参数设置完成后，点击**确定**或**应用**即可完成默认参数的设置。点击**取消**，则取消操作并关闭对话框。接下来的操作与 8.4.1.雷同，保存到文件的格式为.DSD。点击**从文件载入**则弹出如下对话框:



选择一个默认参数设置文件，点击**打开**即可把文件中的参数设置内容填充到界面上。

8.5. 算法流(批方法)

复杂多变量数据处理，可归纳为一个典型的分析流程：即导入待分析数据 -> 数据预处理(质量提高) -> 变量选择 -> 建模分析 -> 预测/验证 -> 生成报表。实因流程的复杂性，特别是流程中不同方法，方法中不同参数的变化，以及被分析数据的变化等，均极大地增加数据分析处理的时间和困难度，给用户使用软件带来很大不便。

本软件的算法流(批方法)功能实现将数据分析流程整合起来，实现一键处理，并在此基础上可实现同时多模型处理，同步建模、预测与验证，达致智慧型数据分析。

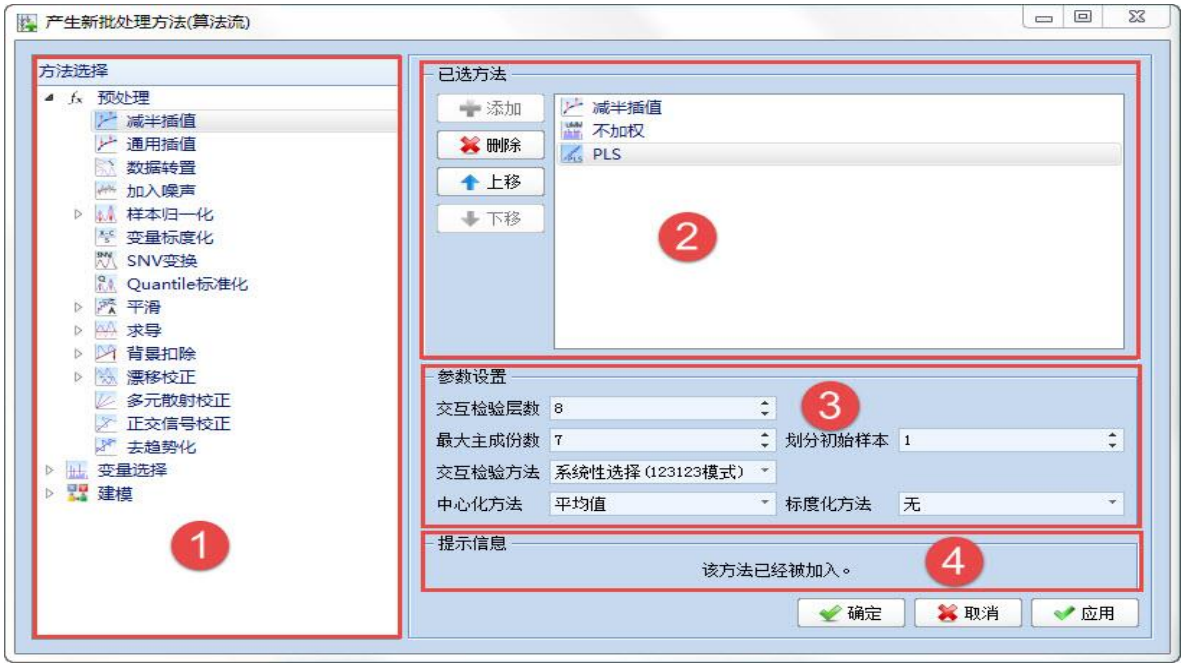
算法流(批方法)的详细介绍，可参考 2.1.1.。用户可自定义不同样式的算法流，本小节将详细介绍该功能的具体操作。

8.5.1. 新建批

新建一个算法流(批方法)流程。

操作步骤:

步骤 1: 点击**主页** -> **新建批**，弹出如下对话框:



如上图所示，新建批界面可分为四个主要部分:

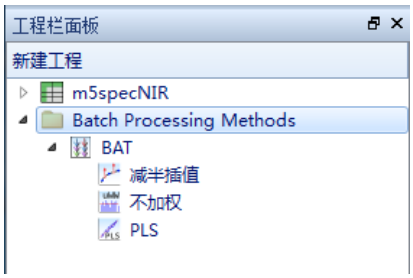
1) 方法选择: 左侧列表包含可被加入算法流中的批处理方法。与第四章中所述的用户界面比较，该方法列表仅排除需要人工手动干预的方法，详见 4.2.4.至 4.2.7.。

2) 已选方法: 根据用户数据处理具体需求，在第 1 部分所示的列表栏中选择相应的数据处理方法，点击添加按钮将被选方法加入到批处理流程框中。点击删除按钮即可移除批处理流程框中的选中方法；通过上移/下移按钮可重排批处理的方法顺序。

3) 参数设置: 参数设置组合框中显示当前被选中的数据处理方法对应参数。参数的初始状态为 8.4.2.中设置的默认参数，用户可以根据需要修改或调整。

4) 提示信息:信息框中显示批处理流程中当前被选中数据处理方法的提示信息。比如若用户点击添加已经存在于算法流中的方法，则提示信息显示“该方法已经被加入。”内容。

步骤 2: 点击**确定**或**应用**即可新建一个批处理流程，而点击**应用**则并不关闭当前界面，可继续操作使用。新建算法流成功后，可在工程导航栏中看到新建的算法流节点，如下图所示:



点击**取消**，则取消操作并关闭对话框。关于节点的详细信息，可参考第五章节点文件夹与节点的管理。

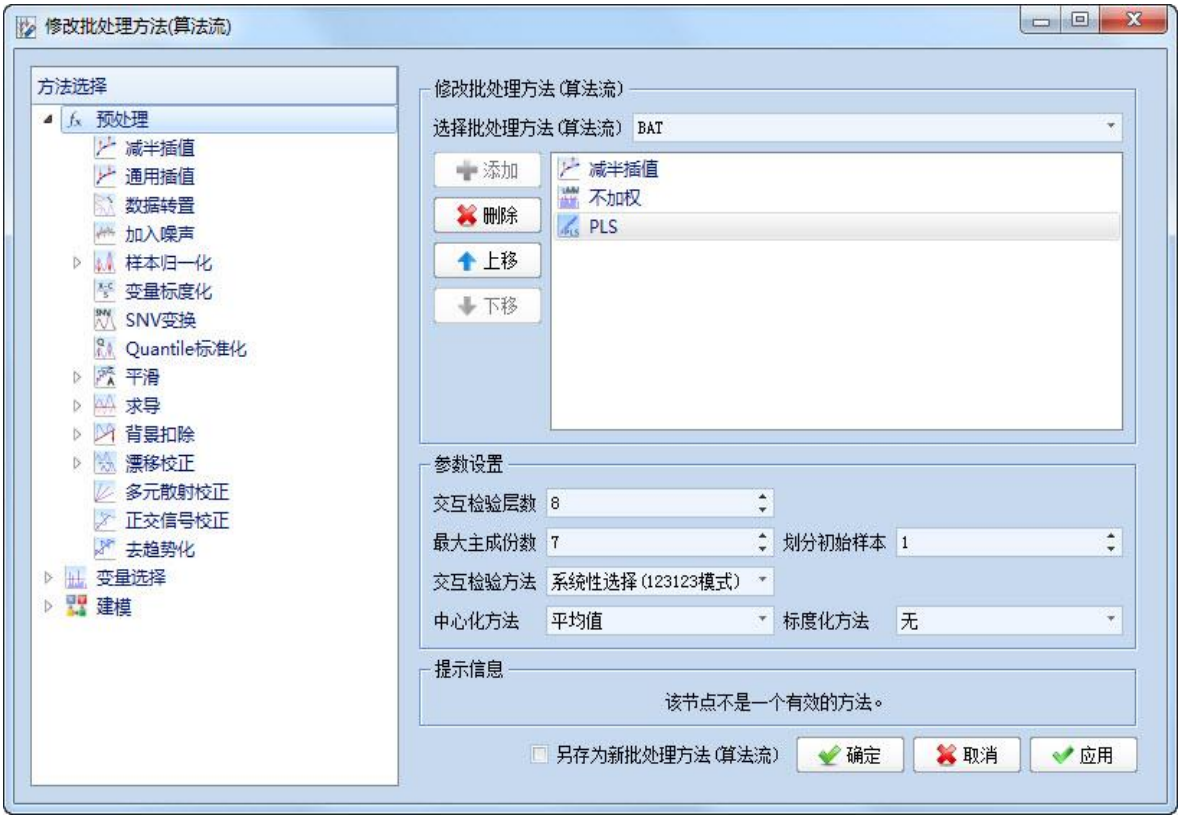
8.5.2. 修改批

通过 8.5.1.所建立的算法流，用户可进行进一步修改，即再次添加或删除算法流中的方法，

调节方法的运算顺序，或者修改方法参数再次运行等。

操作步骤:

步骤 1: 点击**主页** -> **修改批**，弹出如下对话框:



上图中所示界面与新建批对话框类似，不同之处在于修改批对话框提供一个下拉式列表，列举所有算法流方法，用户可任意选择和查看，程序将被选算法流中所包含的方法罗列出来。

修改批对话框亦提供是否另存为新批处理方法(算法流)复选框，选中该复选框，则被修改的算法流将作为新的批节点添加到工程导航栏中，而非替代修改前的批处理节点。

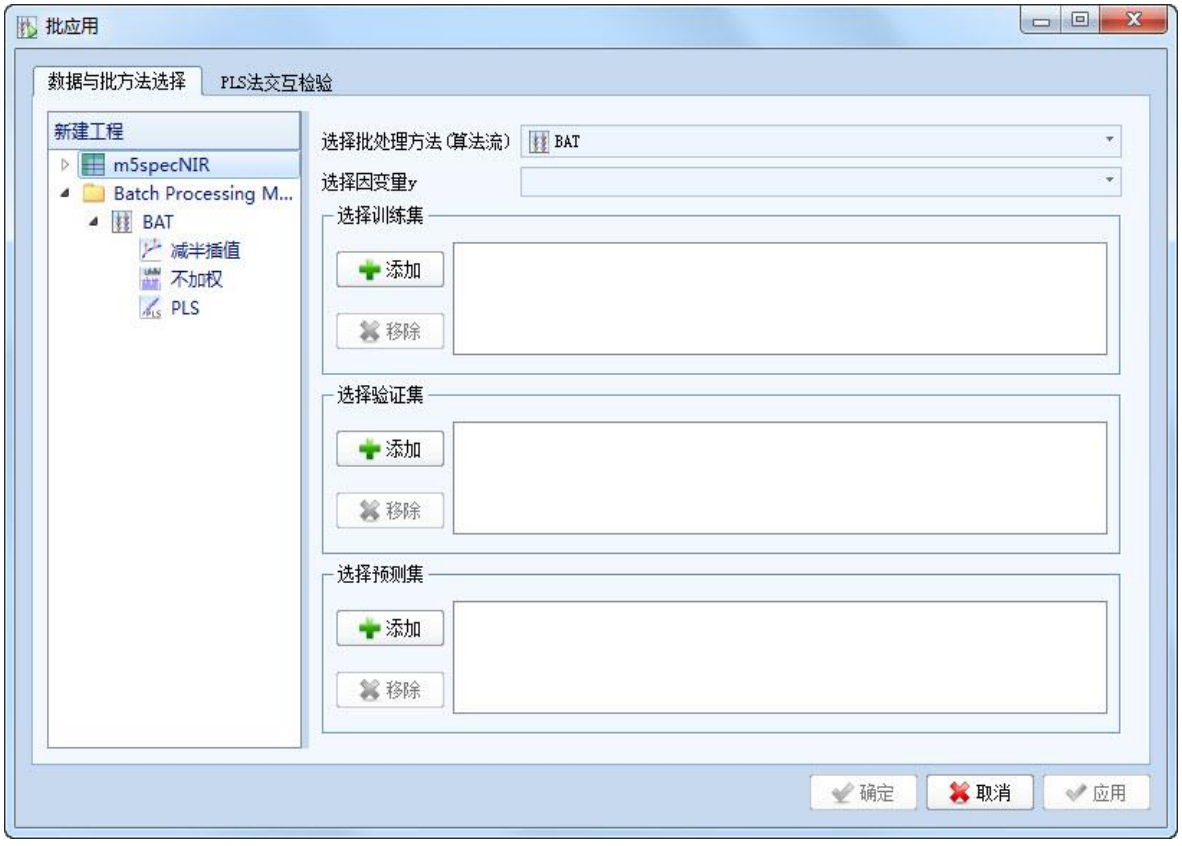
步骤 2: 点击**确定**或**应用**，即可修改当前被选算法流，点击**应用**则可继续在此界面操作。点击**取消**，则取消操作并关闭对话框。

8.5.3. 应用批

应用批处理流程：将 8.5.1.或 8.5.2.所得到的算法流作用于需要分析处理的数据。当然，被分析的数据可以是训练集、校正集、验证集和预测集等，本软件提供同时分析的解决方案，一次性得到分析结果。

操作步骤:

步骤 1: 点击**主页** -> **应用批**，弹出如下对话框:

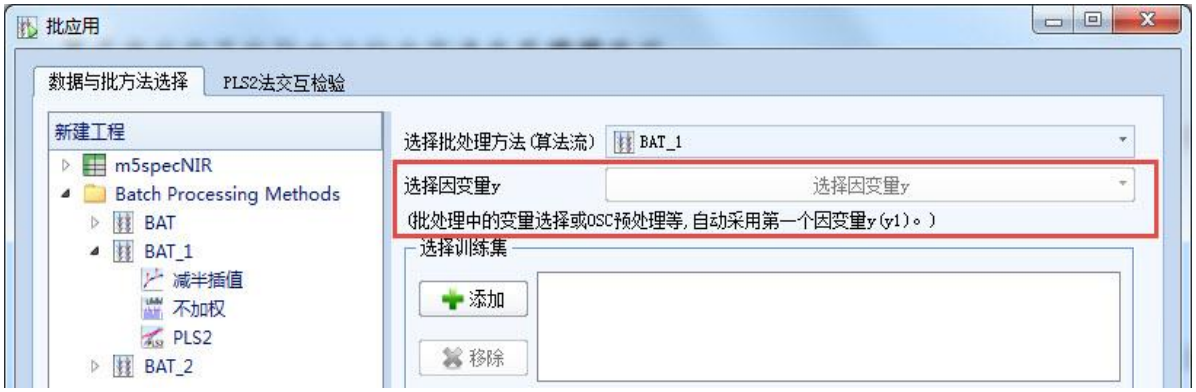


关于应用批有以下几点需要说明:

1) 系统将把所有已经建立好的批处理方法自动读取到选择批处理方法(算法流)的下拉框中。若用户所选择的批处理方法中含有涉及交互检验的方法，则对话框中自动增加该方法的交互检验分页，如下图所示:

更多有关交互检验方法的内容请查看**建模**章节。

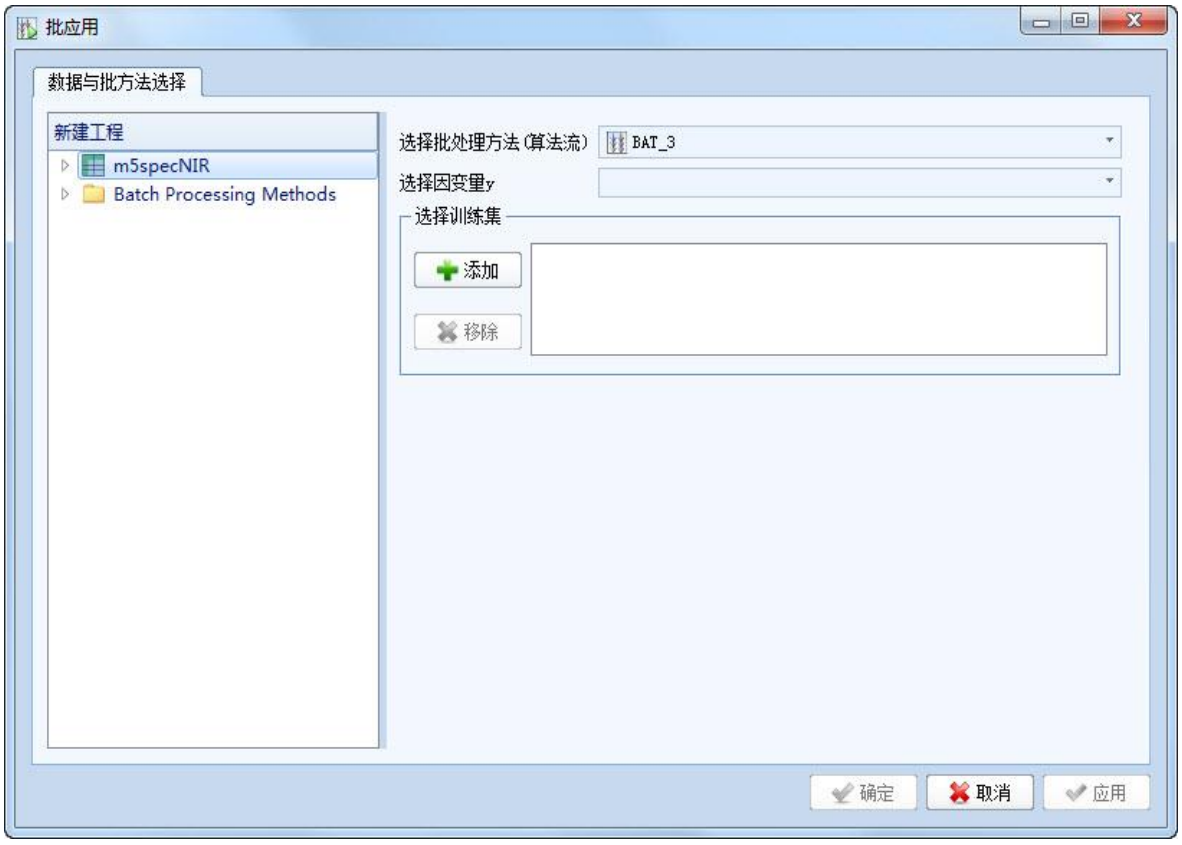
2) 若批处理方法中(本例为 BAT_1),有涉及两个或两个以上因变量 y 的方法(如 PLS2)，则对话框提供选择因变量 y 的控件，如下图所示:



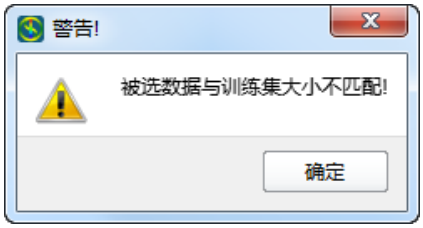
若批处理方法中(本例为 BAT)，不涉及需要两个或两个以上因变量 y 的方法，但涉及需要一个因变量 y (如 PLS)，则对话框提供仅选一个因变量 y 的控件，如下图所示:

若批处理方法中(本例为 BAT_2)，不涉及需要因变量 y 的方法，则以上两种控件均不显示，如下图所示:

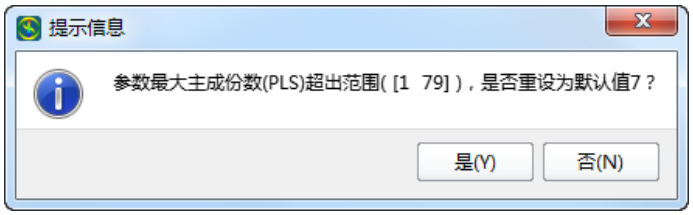
3) 若批处理方法中不涉及验证集，则**选择验证集**界面不显示；若批处理方法中不涉及预测集，则**选择预测集**界面同样不显示，如下图所示:



4) 若涉及验证集或预测集的选择，则需保证所选训练集与预测集中的变量数一致。否则，程序将给出如下图所示的错误提示。



步骤 2: 点击**确定**或**应用**按钮，即可应用所选批处理方法，点击**应用**则可继续在此界面操作。点击**取消**，则取消操作并关闭对话框。若批处理方法中有参数超出范围，则系统将给出提示信息，并推荐最优值给用户，如下图所示:



点击**是**，则程序将超出范围的参数自动设为系统最优值，并以修改后的值参与运算；点击**否**，则不进行修改，但无法进行数据批处理。

8.6. 报表

8.6.1. 产生新报表

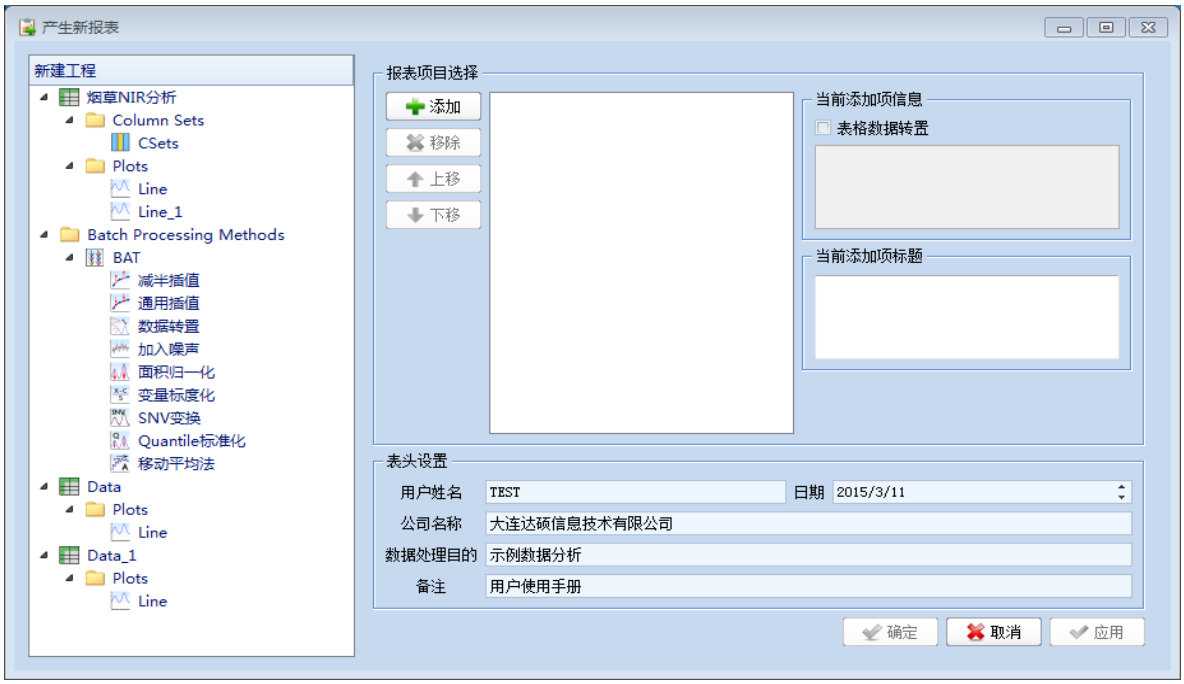
根据用户自定义内容，创建一个新的报表文件。

创建报表时，程序将显示主界面中的左侧工程导航栏，将可加入到报表中的内容完整罗列，用户可快速往报表中添加需要的内容，并加入说明性信息。若用户已经选择性添加导航栏中的项目，产生我的收藏节点文件夹，则在产生报表时，快速添加内容。

若用户在 8.4.1.中添加了报表头的信息，则此时在界面中将完整显示默认信息。用户亦可自定义表头内容的功能，以修改**用户姓名**，**日期**，**公司名称**，**数据处理目的**，**备注**等项目的默认信息。

操作步骤:

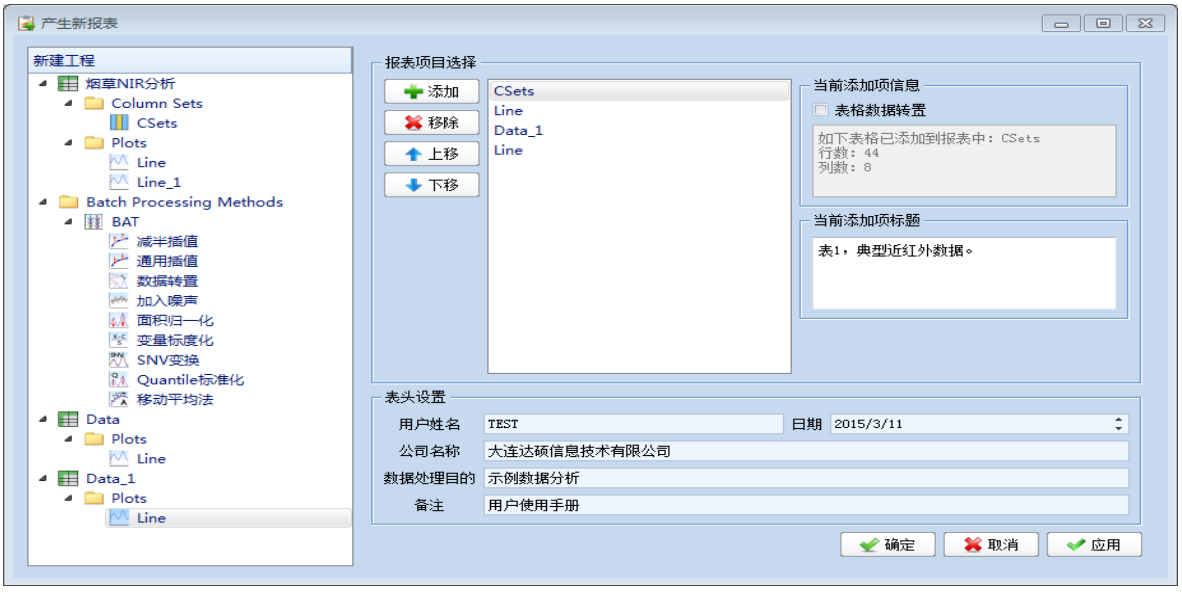
步骤 1: 点击**主页** -> **产生新报表**，弹出如下对话框:



用户可通过**添加**按钮，选择需要加入到报表中的项目，亦可通过移除按钮删除已经被添加的项目，并使用上移/下移按钮调节项目在报表中的顺序。此外，亦可对当前项增加标题(在编辑框当前添加项标题中输入标题内容即可)。

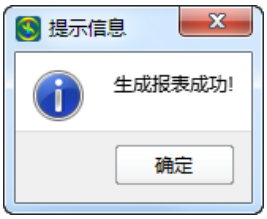
若当前项目为表格，则可对其进行转置操作(勾选表格数据转置复选框)，以便更好地将数据表格呈现在报表中。

图中添加需要产生到报表中的内容后，界面如下图所示。

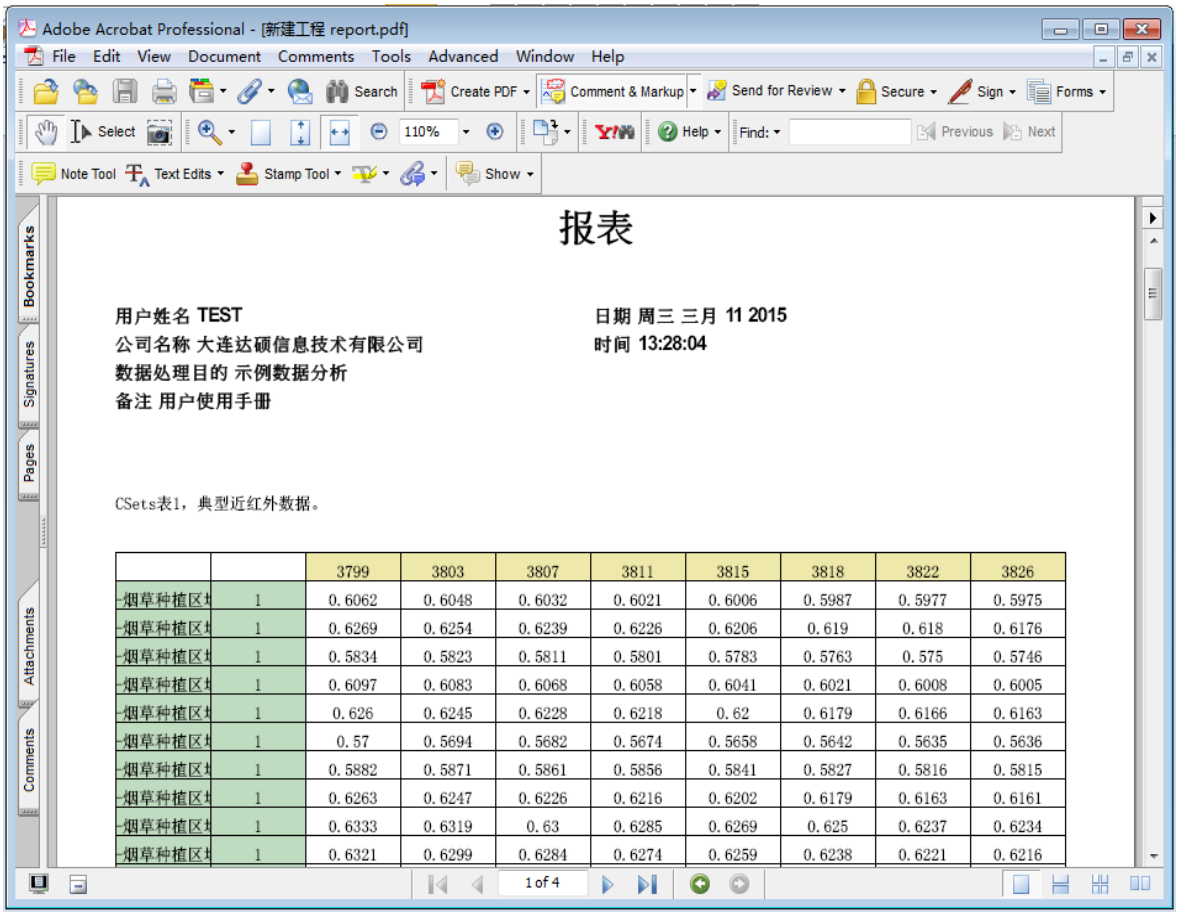


用户在选择添加到报表中的内容时，当选择一个项目后，程序将自动跳转到下一个可加入到报表中的内容，以供用户选择。用户选择添加项目时，右侧信息框中将显示该项目的详细信息，比如被添加表格数据的尺寸大小等。

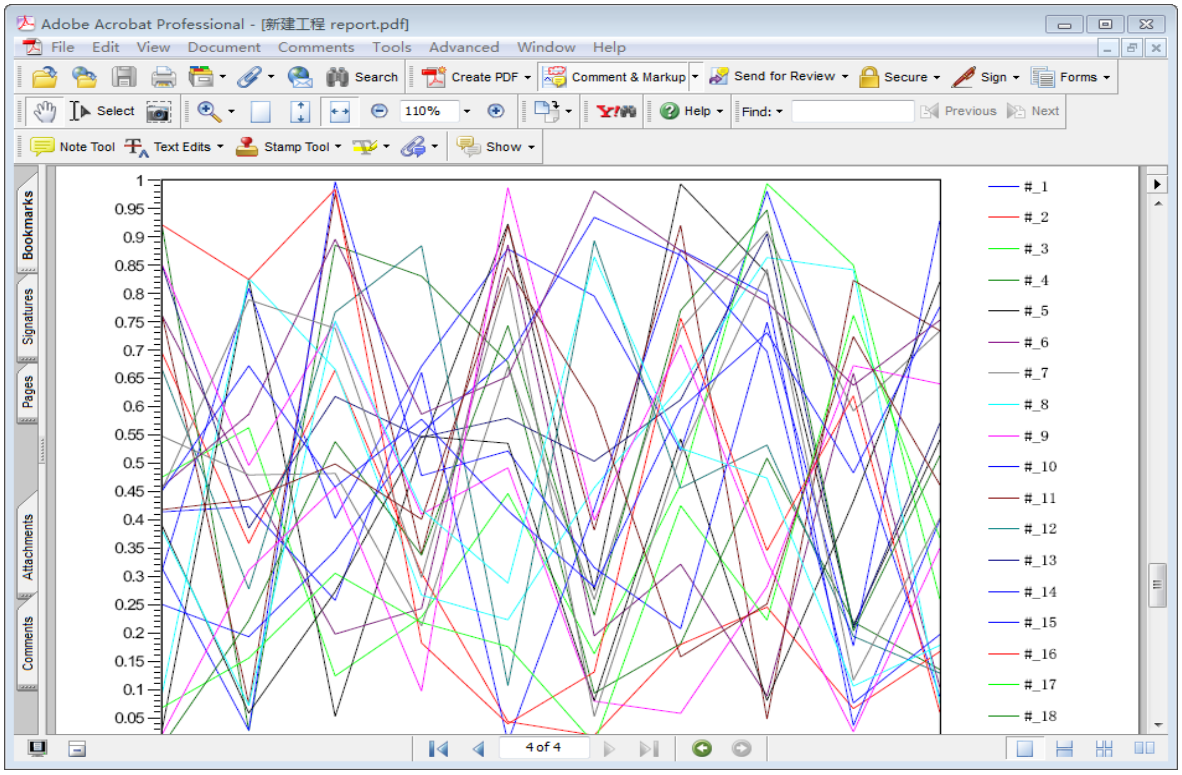
步骤 2：点击**确定**或**应用**即可开始报表，点击**应用**还可继续在此界面操作。点击**取消**，则取消操作并关闭对话框。报表成功弹出如下对话框:



可看到以上示例的报表结果如下二图所示:



		3799	3803	3807	3811	3815	3818	3822	3826
-烟草种植区块	1	0.6062	0.6048	0.6032	0.6021	0.6006	0.5987	0.5977	0.5975
-烟草种植区块	1	0.6269	0.6254	0.6239	0.6226	0.6206	0.619	0.618	0.6176
-烟草种植区块	1	0.5834	0.5823	0.5811	0.5801	0.5783	0.5763	0.575	0.5746
-烟草种植区块	1	0.6097	0.6083	0.6068	0.6058	0.6041	0.6021	0.6008	0.6005
-烟草种植区块	1	0.626	0.6245	0.6228	0.6218	0.62	0.6179	0.6166	0.6163
-烟草种植区块	1	0.57	0.5694	0.5682	0.5674	0.5658	0.5642	0.5635	0.5636
-烟草种植区块	1	0.5882	0.5871	0.5861	0.5856	0.5841	0.5827	0.5816	0.5815
-烟草种植区块	1	0.6263	0.6247	0.6226	0.6216	0.6202	0.6179	0.6163	0.6161
-烟草种植区块	1	0.6333	0.6319	0.63	0.6285	0.6269	0.625	0.6237	0.6234
-烟草种植区块	1	0.6321	0.6299	0.6284	0.6274	0.6259	0.6238	0.6221	0.6216



报表中输出的数据表格，包括样本和变量的说明性信息，非常清晰，查看方便。若数据表格列数太大而无法在一个页面中呈现，则报表产生程序将自动合理切割合适列数的数据，余下的列数据则拼接在后面，同时亦包括样本行的信息。

PDF 文件质量很高，适合论文或出版需要。

8.6.2. 修改报表

暂略。

第九章　图形

可视化图形是数据最直观的表达形式，显然好的图形表达比一连串数据更能有效地传递信息，在复杂数据处理中尤其具有极其重要的意义。一个数据内部，或者多个数据之间往往含有多种复杂关系，但不通过合理的表达，则无法有效体现出来。可视化图形则可在详尽的数据处理前，帮助粗略地理解数据结构，反映数据内部信息和相关关系，有时甚至能达到数据处理同样的效果。

本软件提供如下 8 种绘图方式:

- 曲线图
- 散点图
- 条形堆积图
- 填充图
- 棒状图
- 三维散点图
- 三维表面图
- 用户自定义

9.1. 简述

本软件所述的绘图对象包括基本数据表、产生于基本数据表的行划分、列划分和子数据、以及模型所产生的结果数据等。点击工程导航栏中的数据节点，选择全部或部分数据，如下图所示，再点击图形菜单中不同的绘图类型即可绘制图形。

	V	Var_1	Var_2	Var_3	Var_4	Var_5	Var_6	Var_7	Var_8
#		1	2	3	4	5	6	7	8
#_1	1	0.7629989...	0.6301714...	0.4943015...	0.6987806...	0.1809307...	0.3440078...	0.1903862...	0.047687
#_2	2	0.0441041...	0.9582762...	0.4818951...	0.8317578...	0.8835146...	0.4644531...	0.0611743...	0.423934
#_3	3	0.6575660...	0.8903929...	0.2476369...	0.0649610...	0.4293488...	0.8215772...	0.0559543...	0.351817
#_4	4	0.8555595...	0.1755212...	0.5231968...	0.6332991...	0.3940505...	0.7006020...	0.0431841...	0.848581
#_5	5	0.0596269...	0.5435857...	0.8530550...	0.2753491...	0.0062676...	0.4586467...	0.7549812...	0.056438
#_6	6	0.6868263...	0.7189269...	0.0187414...	0.5059959...	0.9956493...	0.6706878...	0.2751116...	0.392291
#_7	7	0.3396667...	0.8593484...	0.2427031...	0.0148053...	0.2366924...	0.0157735...	0.0649599...	0.089842
#_8	8	0.3387515...	0.3939670...	0.4645341...	0.7600320...	0.6101246...	0.3818818...	0.8917418...	0.836641
#_9	9	0.8973058...	0.5060350...	0.8387629...	0.3840065...	0.3908252...	0.0083547...	0.4391691...	0.327497
#_10	10	0.1385039...	0.9869648...	0.2824059...	0.0867371...	0.4320113...	0.5236251...	0.1259793...	0.628029

9.1.1. 行/列优先绘图

指依据数据矩阵的行或列绘图，详见 3.8.。

9.1.2. 内/外部作图

指绘图时如何选择横坐标，包括选择化学或数学坐标，以及数据矩阵或响应矩阵中的数据二种方式。详见 3.9.和 3.10.。

9.1.3. 选择图形 X 轴

外部作图时，所选的被绘图数据作为 Y 轴，X 轴则由固定行或固定列的数字序号，或者化学坐标提供。

9.1.4. 选择因变量 y

当绘制散点图(二维或三维)，且选择内部作图时，用户可选择因变量 y，所得到的图形将根据 y 的不同类别显示不同形状，以在分类分析中区分样本组别。

9.2. 数据绘图

本软件所能绘制的图形类型及其意义，前面已有详细介绍，相关内容请参见 4.2.3.。

9.2.1. 曲线图

曲线图以曲线的上升或下降来表示统计数量的增减变化，曲线变化幅度越大，则数量关系变化亦越大。与棒状图相比，曲线图不仅可以表示数量的多少，亦可反映同一样本在不同时间内的发展变化情况。

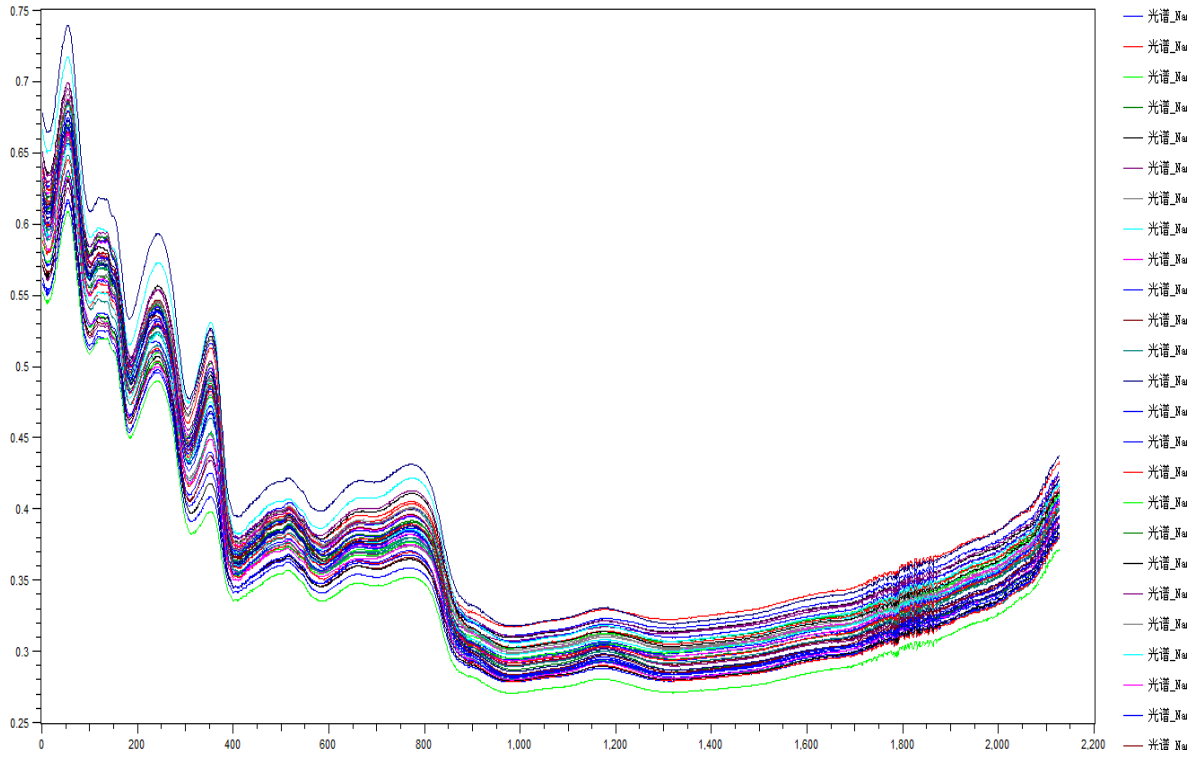
操作步骤：

步骤 1：点击数据节点，选择数据中的全部或部分数据。

步骤 2：点击图形菜单，选择曲线图，弹出如下对话框：

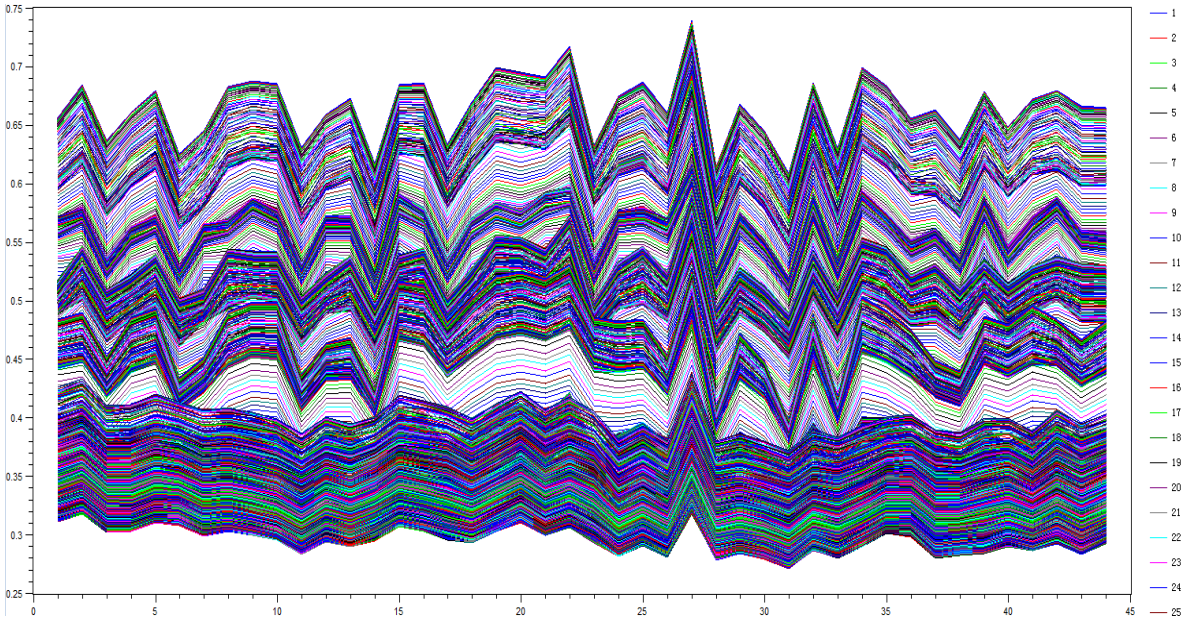
步骤 3：选择上图对话框中的可选项，绘图优先性可从行优先和列优先中选择，前者的 X 坐标轴可选数学序号或变量化学坐标(如近红外波长)，而后者的 X 坐标轴则可选数学序号或样本属性(如类别值或某一化合物含量)。点击**确定**得到对应的图形，点击**取消**则关闭窗口，不进行任何操作。

若用户选择行优先绘图，X 坐标轴选择数学序号，则得到如下图所示的图形：

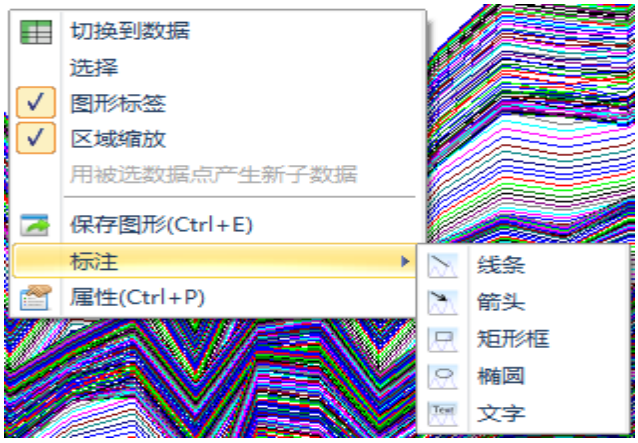


很显然，上图显示以近红外量测数据，实因本软件定义行为样本，列为变量，则图中每一曲线表示某一样本在所有量测变量下所得到的值，不同样本则以不同颜色区分，共有 44 条曲线，每条曲线均含 2127 个点。

然而对同一数据，若选择列优先绘图，X 坐标轴选择数学序号，则得到如下图所示的图形：



该图中每一曲线均表示一个变量在不同样本中变化情况，共有 2127 条曲线，每条曲线均含 44 个点。从上二图可以看出，尽管使用同一数据，所得到的图形完全不同。绘制得到图形后，在图形的任一区域内，点击鼠标右键，可得如下图所示的图形：



下表归纳图形右键菜单功能：

序号	属性名称	图标	说明
1	切换到数据		实现图形与数据的界面互换，即点击可进入数据页面。

2	选择	无	点此可选择图形中的曲线(样本或变量)，且 Ctrl 键可用，此时图形菜单下的选择功能同时被激活，即  。
3	图形标签	无	默认状态为激活，点此则关闭图形标签，即  。
4	区域缩放	无	默认状态为激活，即可放大或缩小图形，点此关闭图形缩放。
5	用被选数据点产生新子数据	无	即用户使用选择功能，从图形中选择数据后(样本或变量)，此功能可帮助用户根据图形选择合适数据。
6	保存图形		将当前图形保存为 PDF 格式文件。
7	标注	无	图形标注功能的右键快捷键。
8	属性		图形属性，点此可进入图形属性修改界面。

上述功能中，图形属性修改部分尤为重要，下面做一详细介绍。

图形属性修改功能分为二个部分，其中图形基本属性是不同绘图类型均涉及到的内容，先对此做整体介绍，不同图形的特色性属性修改，则介绍各类图形时单独加以说明。

9.2.1.1. 图形基本属性

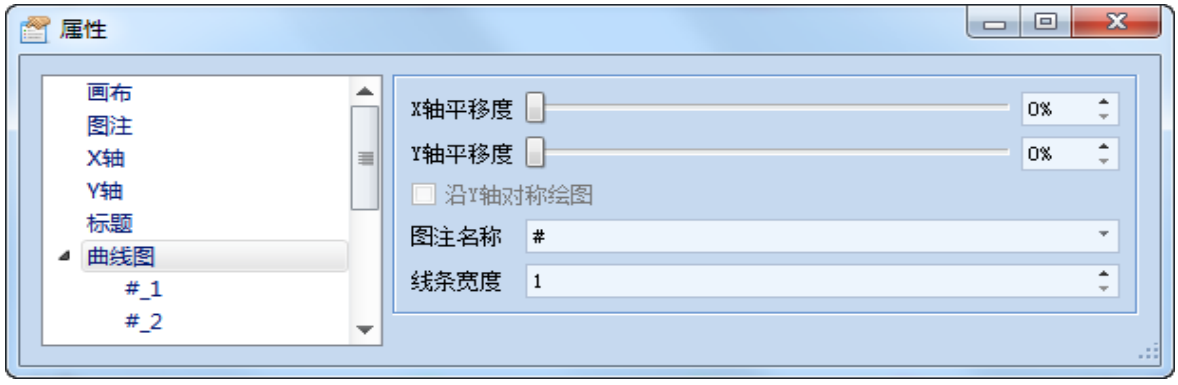
二维绘图时，属性对话框的固定内容如下图所示：

如下表则详述各部分的功能，以及可修改的属性等内容。

序号	属性类型	说明
1	画布	控制和修改图形边框颜色、背景颜色、一级网格和二级网格属性。当作图方式为曲线图时，可设置反走样。
2	标签	控制和修改图注的位置、隐藏/显示设置。
3	X 轴	激活 X 轴(默认激活)、添加说明文字、修改字体。
4	y 轴	激活 X 轴(默认激活)、添加说明文字、修改字体。
5	标题	添加图形标题、修改字体属性。

属性修改：本软件提供完整图形属性修改功能，用户通过使用这些功能，可以得到满足各种不同需要的美观图形。

如下图显示曲线图的整体属性，即针对所有曲线的属性修改。

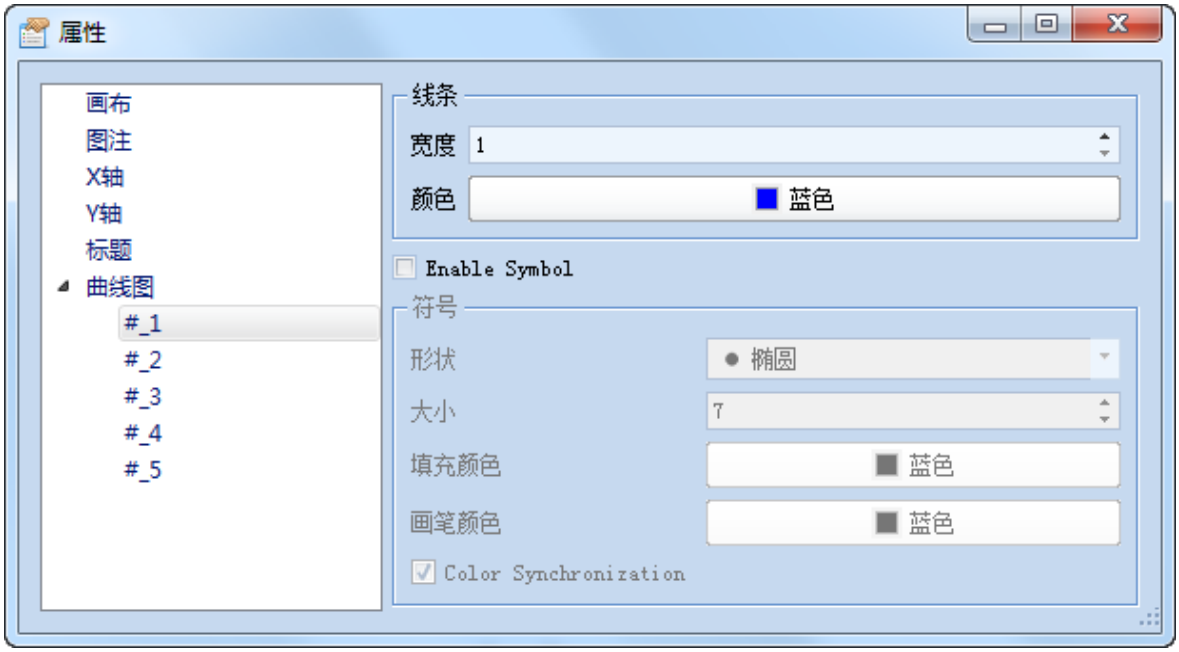


下表则详细上述属性的修改内容。

序号	属性名称	说明
1	X 轴平移度	除固定第一条曲线外，可选择沿 X 轴平移图形中不同曲线。

2	Y 轴平移度	除固定第一条曲线外，可选择沿 Y 轴平移图形中不同曲线。
3	沿 Y 轴对称绘图	当只有两条曲线时，则额外增加沿 Y 轴对称绘图的勾选框，使第二条线的 Y 值会变成相反数，倒绘在 X 轴的下方，以更好地比较二条曲线。
4	图注名称	图注名称下拉列表可修改图注旁的说明文字。
5	线条宽度	可整体改变界面上所有的曲线宽度。
6	线条框	选中单个线条，可修改其宽度和颜色。
7	符号可用	可给单个线条添加符号。
8	符号框	当符号可用时，可修改其形状、大小、填充颜色、画笔颜色，亦可使画笔颜色和填充颜色不同步，以单独修改颜色。

针对每条曲线，提供如下图所示可被修改的属性。



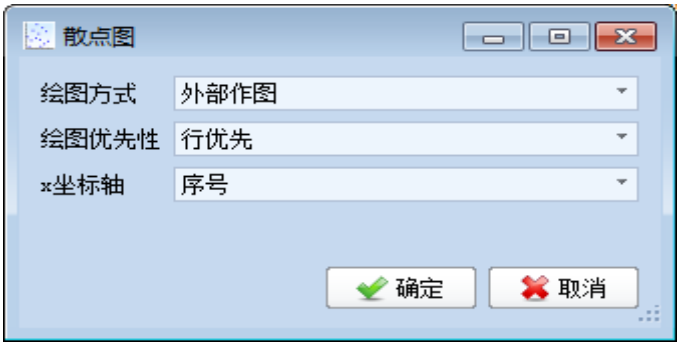
其他曲线的属性修改完全一样。

9.2.2. 散点图

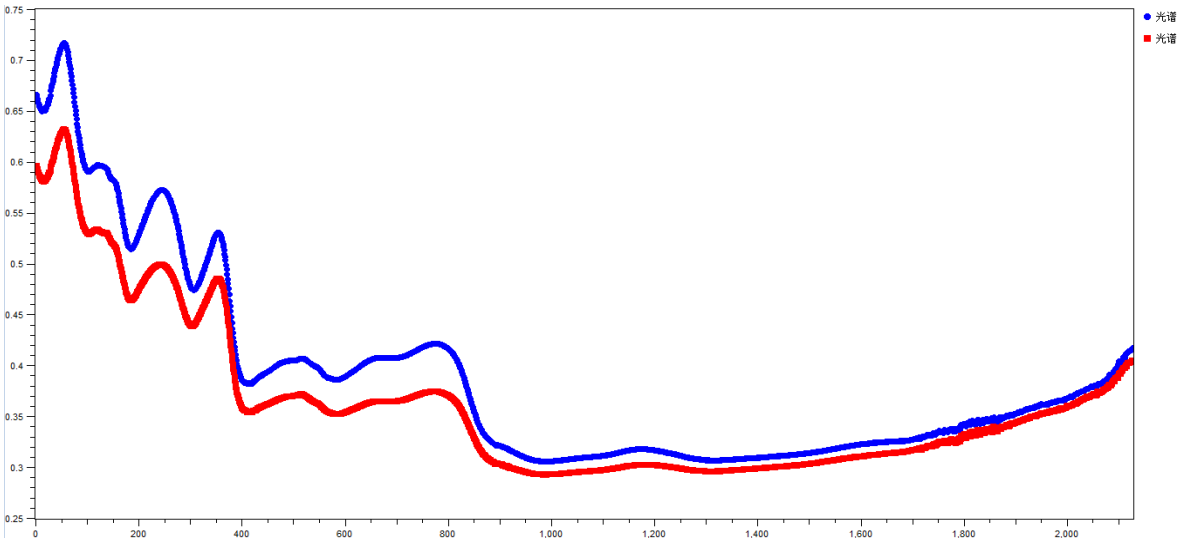
散点图与与前述曲线图类似，其差异在于以点的形式表示不同数值，而不是将相邻点之间用直线连接。散点图的另一个差异之处则是其同时包含内部绘图与外部绘图二种方式，详情请参见 3.9.和 3.10.。

操作步骤：

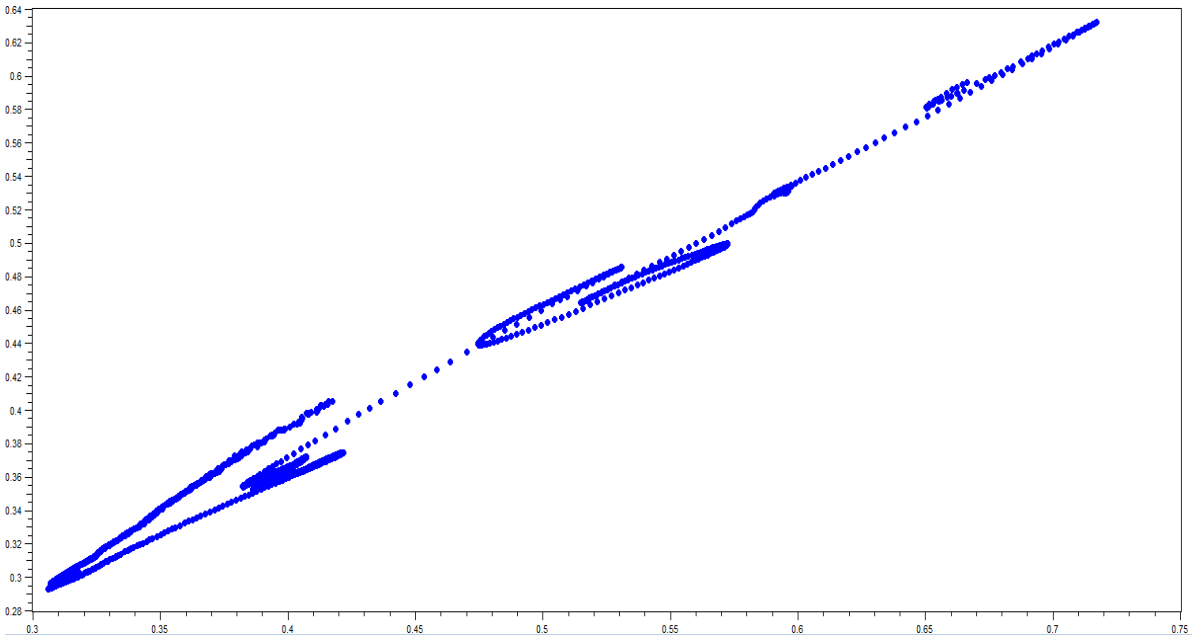
绘图步骤与 9.2.1.雷同，差异在于选择数据，点击绘制散点图后，出现如下图所示的对话框：



如前所述，散点图多了内部绘图方式，以一个包含二列的数据矩阵为例，若选择外部绘图，行优先，以及数学序号选项，则得到如下图所示的图形(选择基本数据表数据中的二行)。

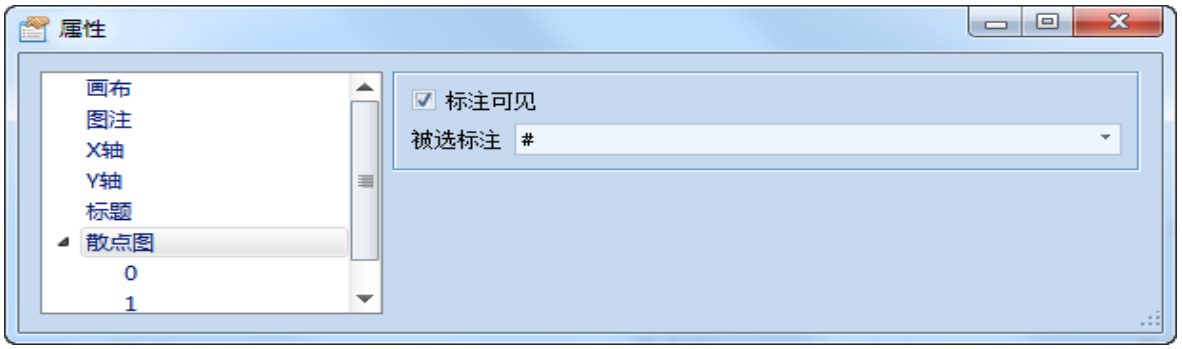


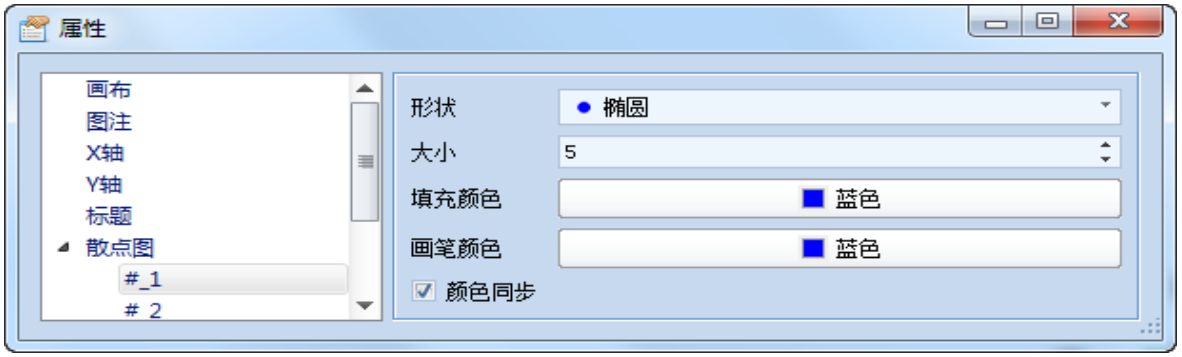
若选择内部绘图，行优先，以及数学序号选项，则得到如下图所示的图形(选择基本数据表数据中的二行)。



很显然，上述二图差异明显，表达的意义亦完全不同。前者表示二个样本的量测数据，而后者则将二个样本的变量量测值分别作为 X 和 Y 轴，表示他们的量测值变化关系。

属性修改：除 9.2.1.1.所述基本属性外，散点的属性修改还包括如下二图所示的内容，分别对应整个散点图共有属性的修改，以及单组散点图的属性修改。





上述属性分别介绍于下表：

序号	属性名称	说明
1	标注可见	在每个散点旁加注说明文字，默认状态为不勾选。
2	被选标注	当标注可见勾选时，可从下拉列表中选择散点说明文字的类别。
3	形状	修改散点的形状。
4	大小	修改散点的大小。
5	填充颜色	修改散点的填充颜色。
6	画笔颜色	修改散点的边框颜色。
7	颜色同步	使散点边框颜色与填充颜色同步，默认状态为勾选。

在解决分类问题时，往往采用散点图，以散点的不同类型或颜色表示不同类别。

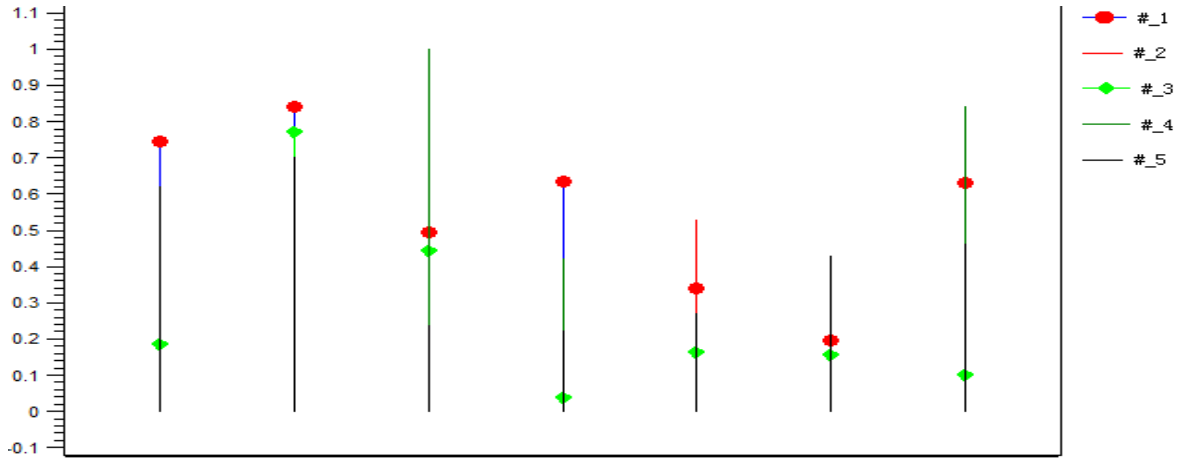
9.2.3. 条形堆积图

条形堆积图是棒状图的变种，可很直观地比较多组数据间数值的变化关系。

操作步骤：

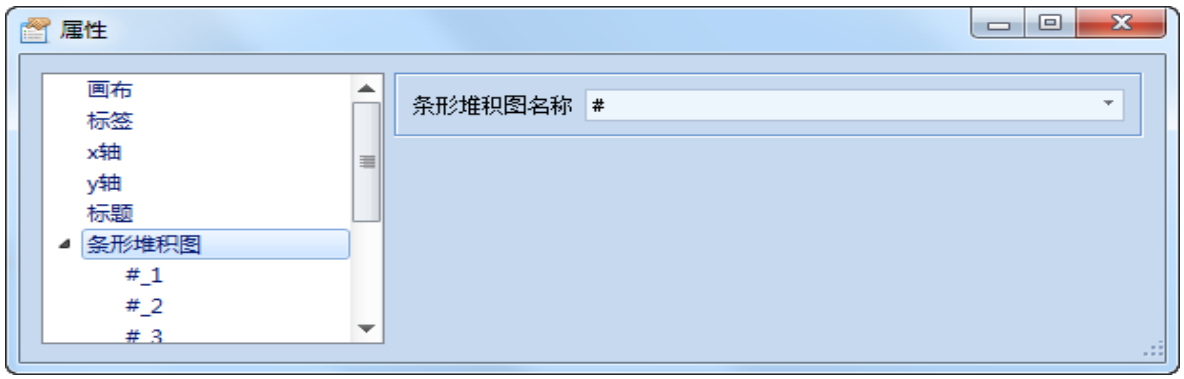
条形堆积图的操作步骤与 9.2.1.雷同，包括选择数据，点击绘图后出现的对话框。

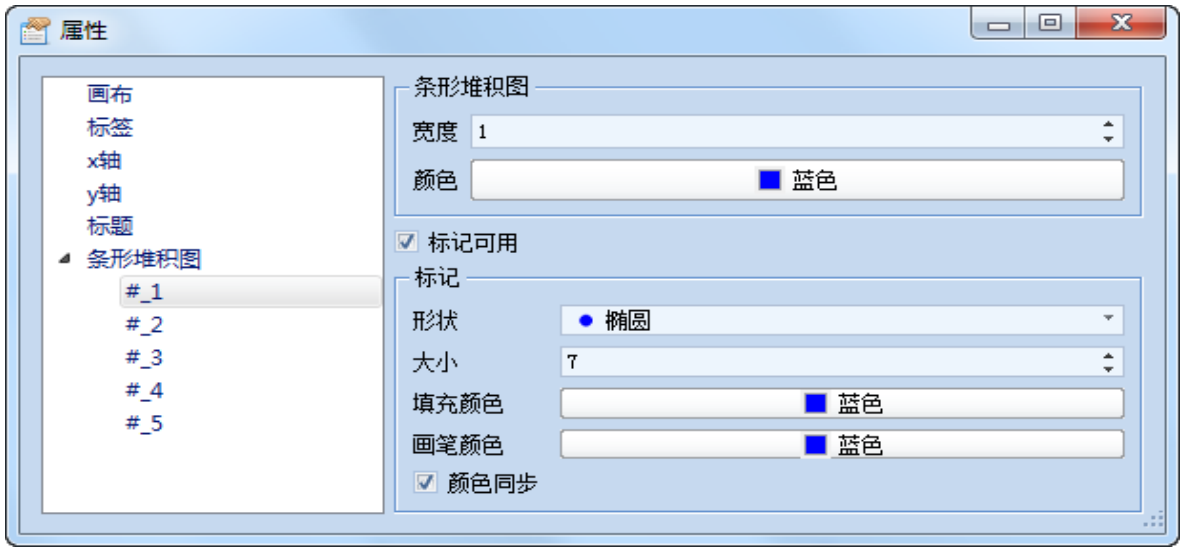
典型的绘图结果如下图所示，同时修改图形属性加注标记。



若上图为行优先绘图得到的结果，则选择了 7 行数据，每行选择 5 个变量数据；若上图为列优先绘图得到的结果，则选了 7 列数据，每列选择 5 个样本数据。

属性修改：除 9.2.1.1.所述基本属性外，条形堆积图的属性修改还包括如下二图所示的内容，分别对应整个图形的共有属性修改，以及单组图的属性修改。



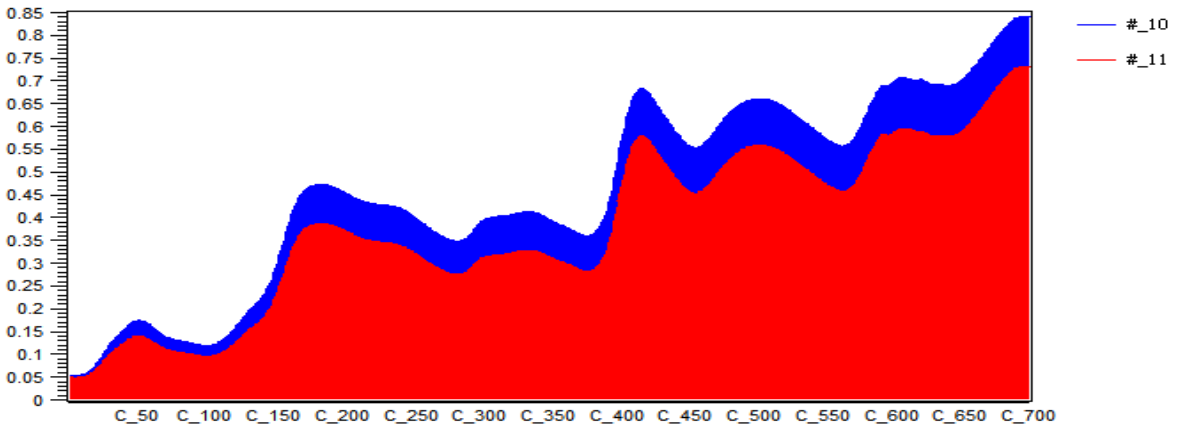


上述属性分别介绍于下表：

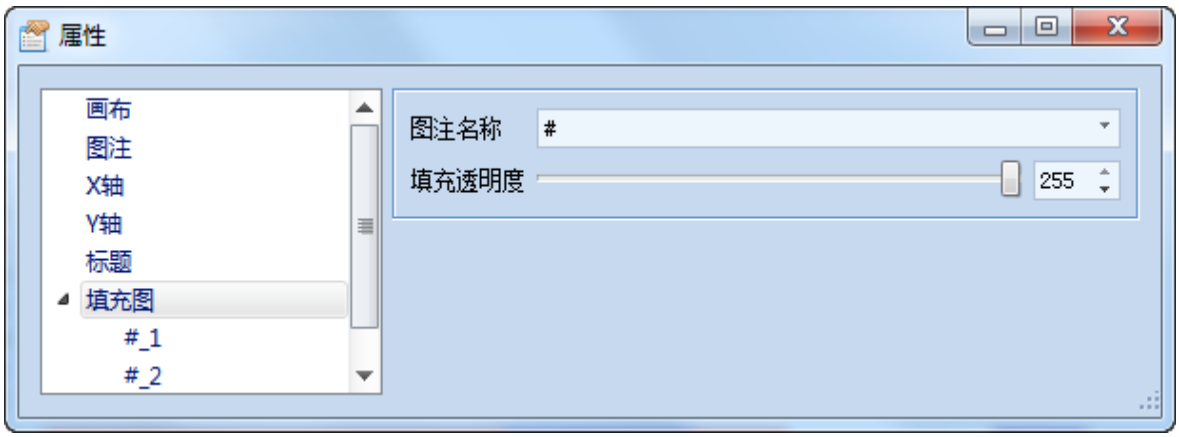
序号	属性名称	说明
1	图注名称	修改图注旁的说明性文字。
2	宽度	修改线条宽度。
3	颜色	修改线条颜色。
4	形状	修改条形堆积图上散点形状。
5	大小	修改条形堆积图上散点大小。
6	填充颜色	修改散点的填充颜色。
7	画笔颜色	修改散点的边框颜色。
8	颜色同步	使散点边框颜色与填充颜色同步，默认状态为勾选。

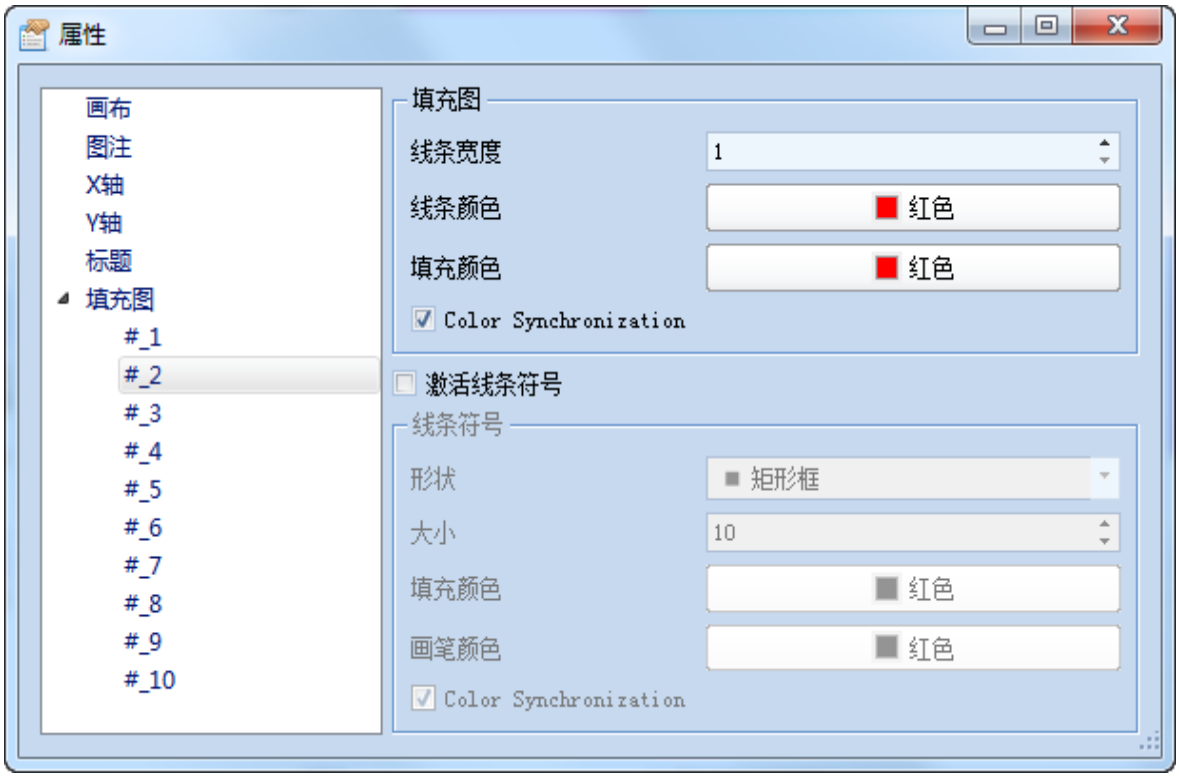
9.2.4. 填充图

填充图是在 9.2.1.曲线图的基础上，以不同颜色填充曲线下面积后所得到的图形。填充图中若前后多个曲线下面积重叠在一起，则前面的填充图会遮挡后面的图形，此时可通过属性对话框调节透明度以实现图形优化。



绘图步骤与 9.2.1.完全相同，差异在于属性修改部分。属性修改：除 9.2.1.1.所述基本属性外，条形堆积图的属性修改还包括如下二图所示的内容，分别对应整个填充图的共有属性修改，以及单个图形的属性修改。





上述属性分别介绍于下表，相对而言，填充图的属性更加丰富。

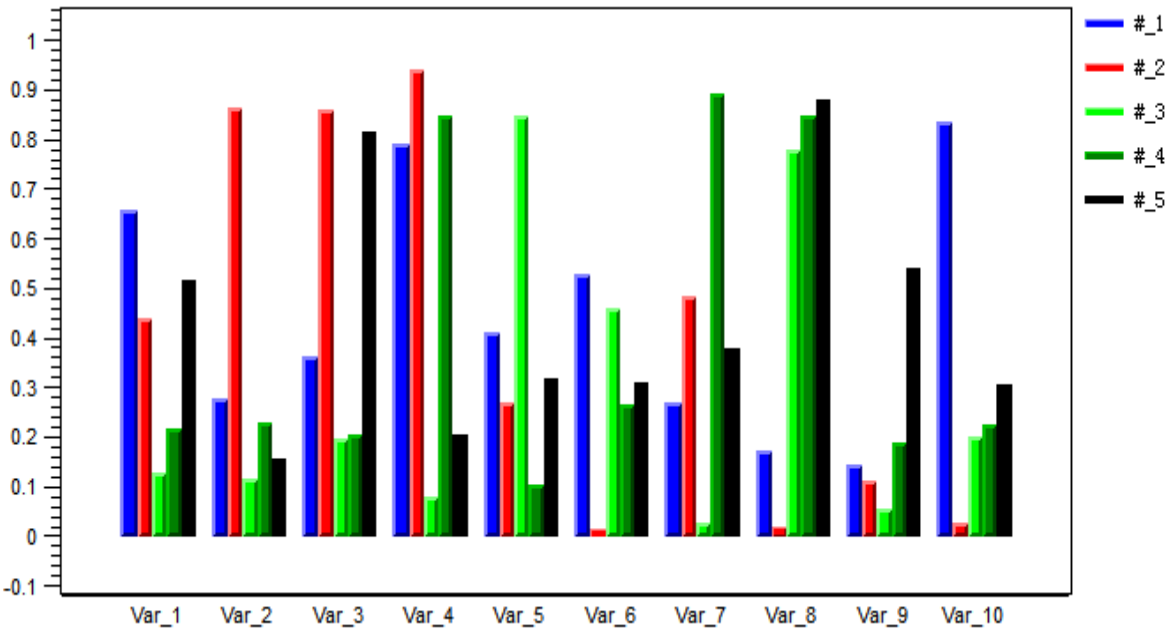
序号	属性名称	说明
1	图注名称	修改图注旁的说明文字。

2	填充透明度	修改填充颜色的透明度，默认状态为不透明。
3	线条宽度	修改线条宽度。
4	线条颜色	修改线条颜色。
5	填充颜色	修改填充颜色。
6	颜色同步	使线条颜色与填充颜色同步，默认状态为同步。
7	激活线条符号	使线条添加符号，默认状态为不添加。
8	形状	修改符号的形状。
9	大小	修改符号的大小。
10	填充颜色	修改符号的填充颜色。
11	画笔颜色	修改符号的边框颜色。
12	颜色同步	使符号颜色与画笔颜色同步，默认状态为勾选。

9.2.5. 棒状图

棒状图亦称条形统计图，以单位长度表示一定数量，根据数量多少绘制长短不同的棒状直条，再把这些直条按一定顺序排列起来，从棒状图中可很容易看出数量的变化，比如本软件中所述的样本或变量，或者模型结果等。

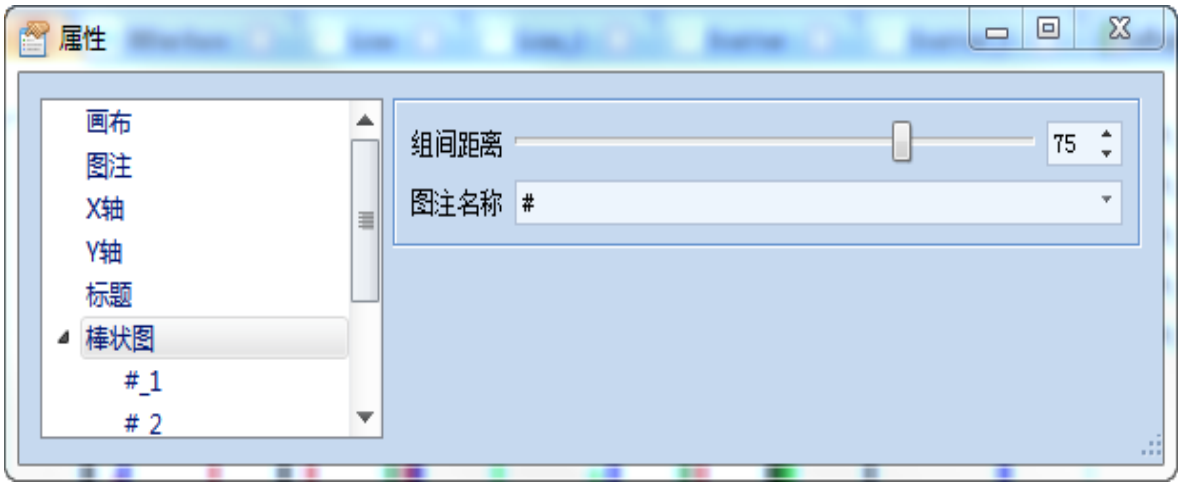
如下图为一典型棒状图结果，若为列优先的结果，在本软件中表示 10 个不同变量在 5 个不同样本中的变化情况；若为行优先的结果，则表示 10 个不同变量在 5 个不同样本中的变化情况。



棒状图以分组的形式，清晰比较不同样本中变量的变化情况，或不同变量在样本中的变化。棒状图条形堆积图类似，其差异在于前者将同组内的样本或变量数据并排而非堆积排列，同时线的初始宽度更大。

绘图步骤与 9.2.1.完全相同，差异在于属性修改部分。

属性修改：除 9.2.1.1.所述基本属性外，棒状图的属性修改还包括如下二图所示的内容，分别对应整个棒状图的共有属性修改，以及棒状图组的属性修改。



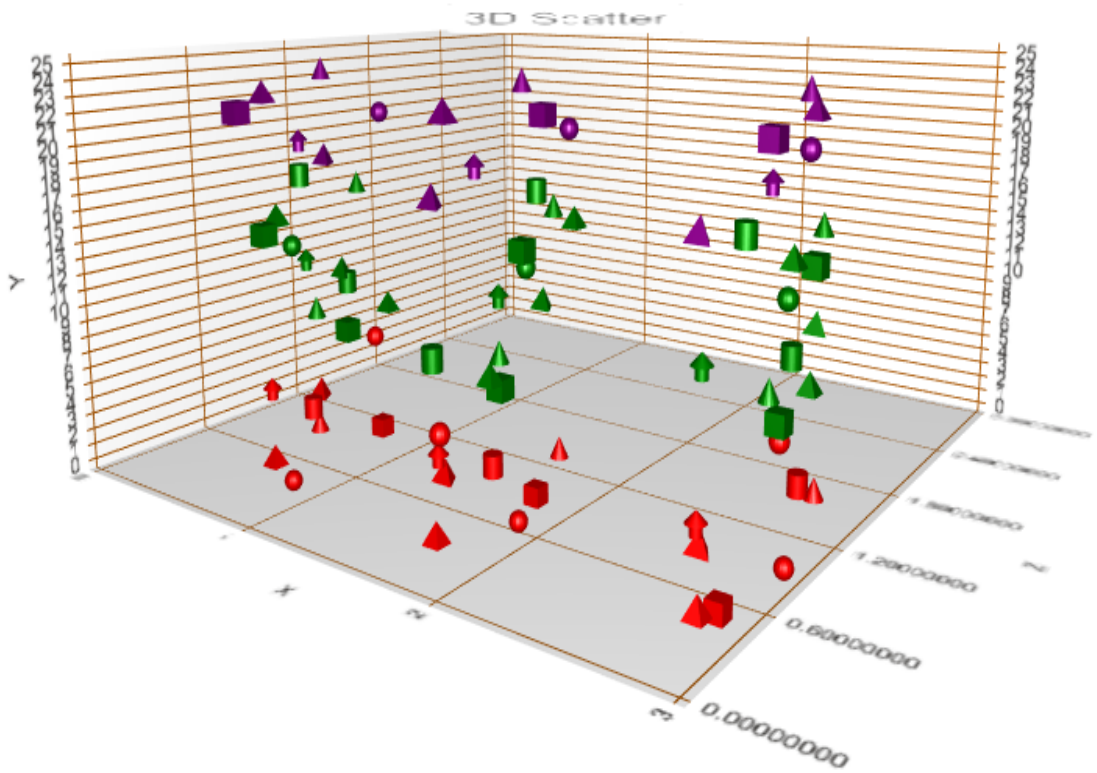
通过上述属性对话框可修改棒状图的组间距离，以及图注旁的说明性文字等。上述属性分别介绍于下表：

序号	属性名称	说明
1	组间距离	修改组间的显示距离。
2	图注名称	修改图注名称。
3	Bar 颜色	修改单个棒状的颜色。
4	填充图案	修改画布的填充类型。
5	填充颜色	修改填充颜色。

9.2.6. 三维散点图

三维散点图与 9.2.2.雷同，差异在于在二维散点图的基础上，增加了一个数据维度，以散点的形式更好地表征数据。

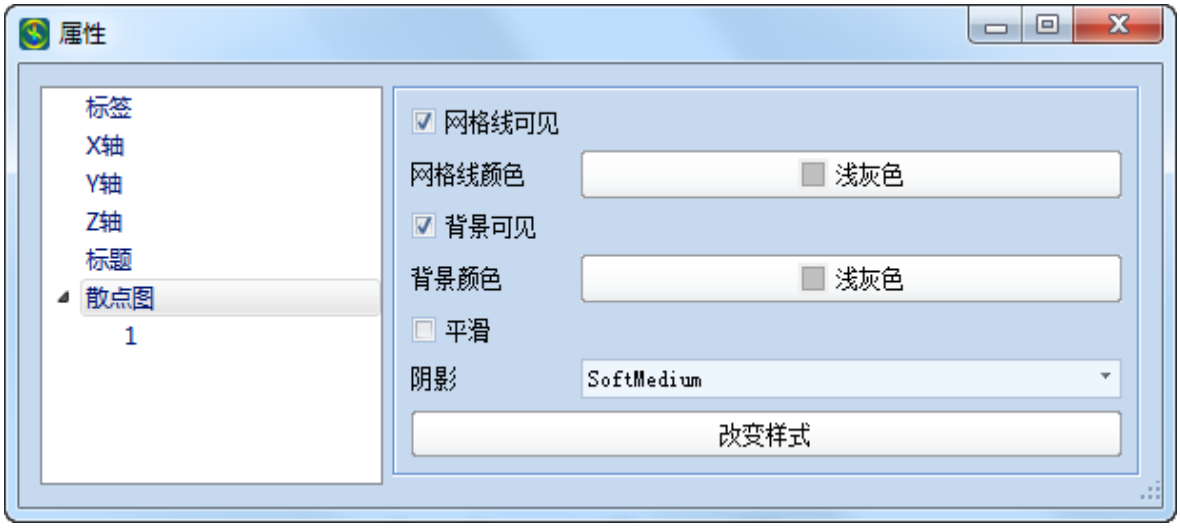
如下图是一个典型的三维散点图，不同三点均可以形状和颜色区分。

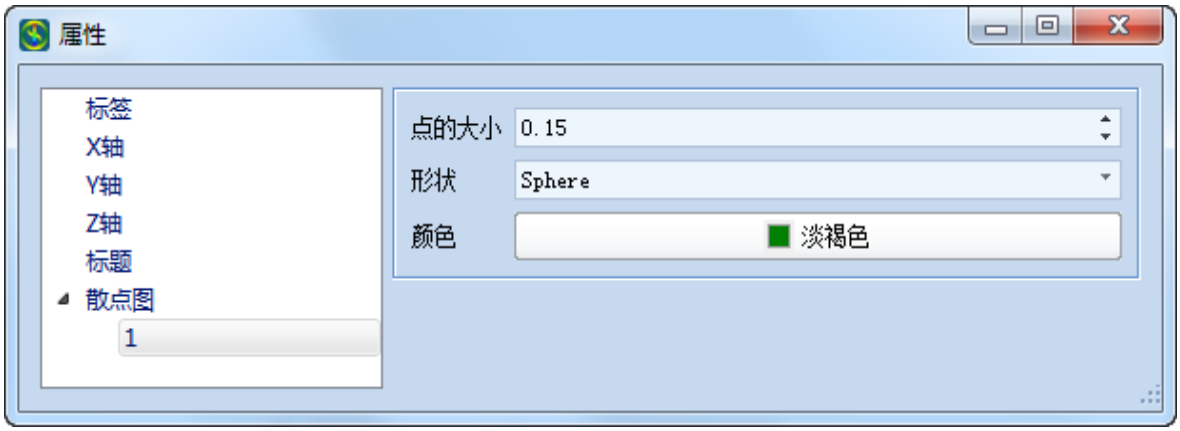


操作步骤：

绘图步骤与 9.2.2.雷同，区别在于选择数据后，点击绘图菜单后出现的参数选择对话框有所区别，如下图所示。实因三维散点图涉及多一维数据，其坐标轴的选择亦多了一项，其余部分与 9.2.2.相同。

属性修改：除 9.2.1.1.所述基本属性外，三维散点图的属性修改还包括如下二图所示的内容，分别对应整个三维散点图的共有属性修改，以及单个三维三点组的属性修改。





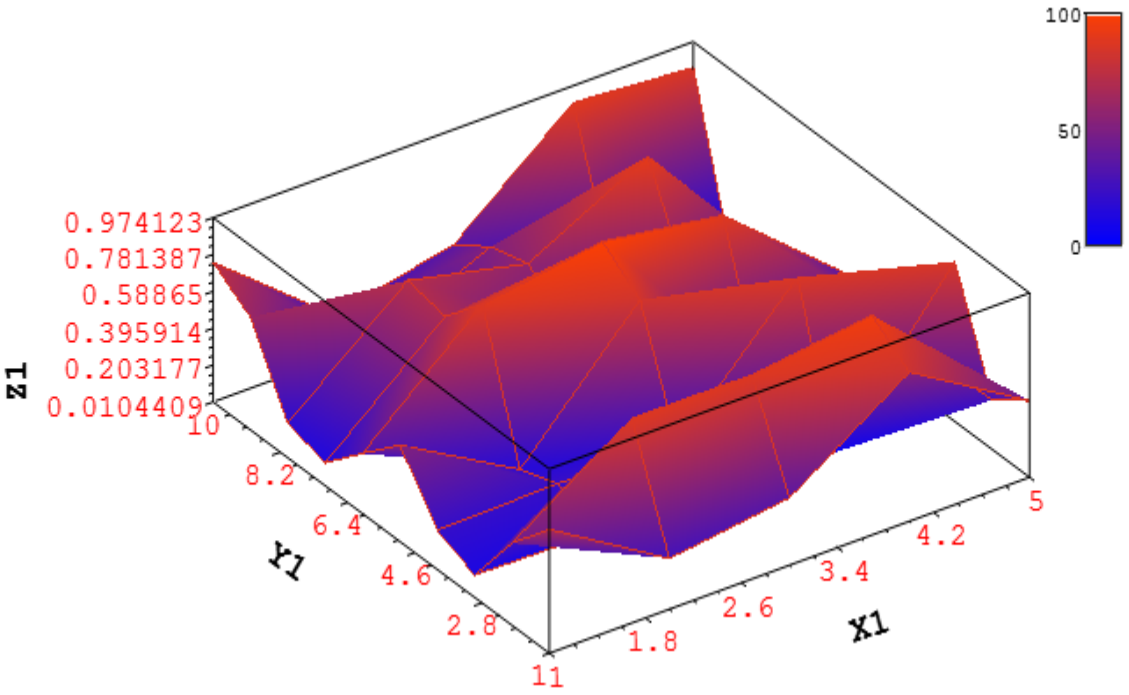
从上图中可以看出，用户可修改图形背景、网格以及坐标轴的显示类型(通过改变样式按钮)等。上述属性分别介绍于下表。

序号	属性名称	说明
1	网格可见	使网格可见。
2	网格颜色	修改网格颜色。
3	背景可见	使背景可见。
4	背景颜色	修改背景颜色。

5	平滑	是否平滑图形。
6	阴影	修改阴影类型。
7	改变样式	修改坐标轴的显示样式。
8	点的大小	修改点的大小。
9	形状	修改点的形状。
10	颜色	修改点的颜色。

9.2.7. 三维表面图

如前所述，三维表面图图亦可表征丰富的数据信息，详见 4.2.3.1.。一个典型的图形如下所示。



操作步骤：

绘图步骤与 9.2.6.雷同，区别在于选择数据后，点击绘图菜单后出现的参数选择对话框有

所区别，如下图所示。其余部分与 9.2.6.相同。

属性修改：三维表面图的属性修改很丰富，此处暂略。

若用户选择少于 3 行或列的数据，将得不到合适的图形结果。

9.2.8. 用户自定义图形

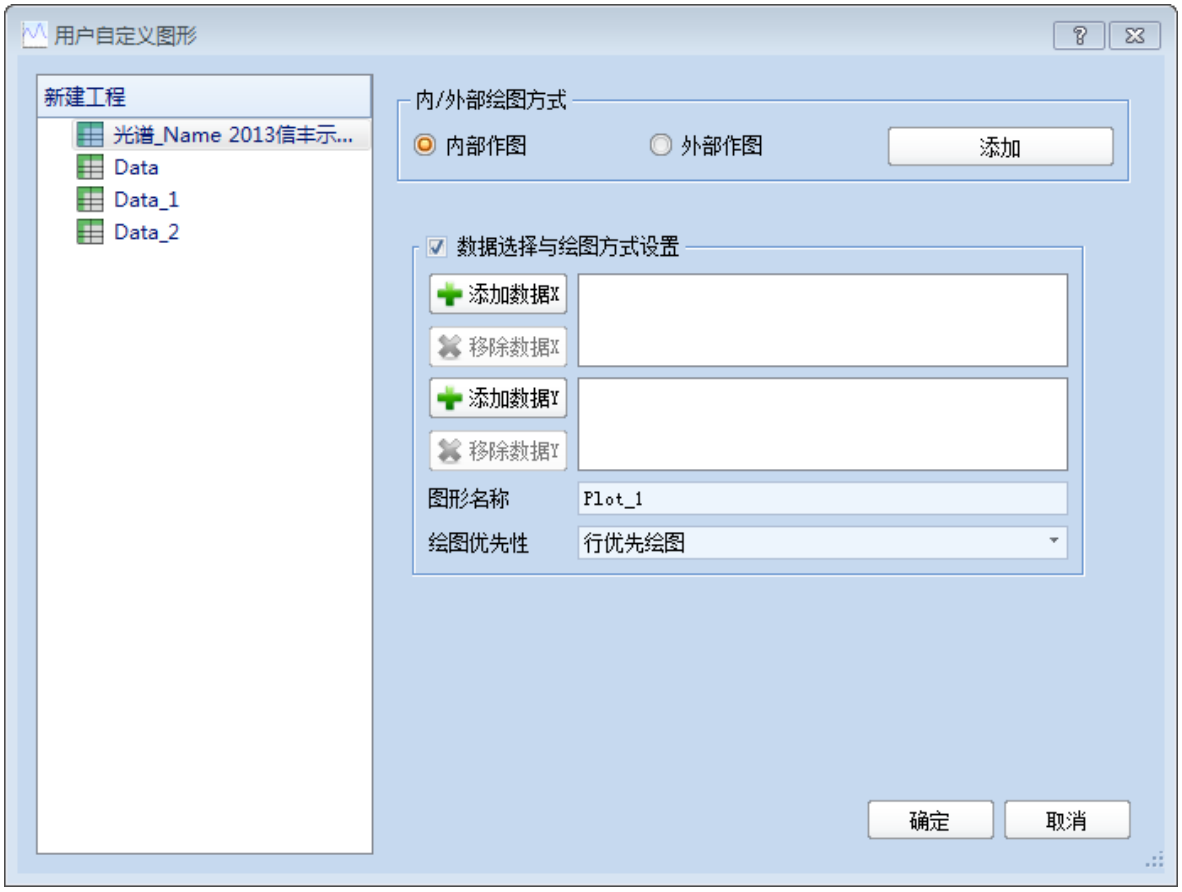
用户自定义绘图提供丰富的绘图选择，可根据用户的需要，同时对多个矩阵数据绘图，并可将其同时显示在同一图形内。

1) 用户可从工程导航栏中选择任意矩阵绘图；绘图方式亦分为内部和外部二种方式；在内部和外部绘图界面下，其可选项和界面不同。
2) 可以将多个不同数据绘制在同一个图上，以方便比较，增大信息量。
3) 选择外部作图时，可切换不同的显示方式，包括曲线图，散点图，条形堆积图，填充图，棒状图。

操作步骤：

初始操作步骤与上述绘图方式不同，自定义绘图无需事先选择数据，直接点击图形下的对应菜单即可(如下图所示)，数据选择等则在新出现的界面中完成。

- **第一种情形：面中选择内部作图，则界面如下图所示。**

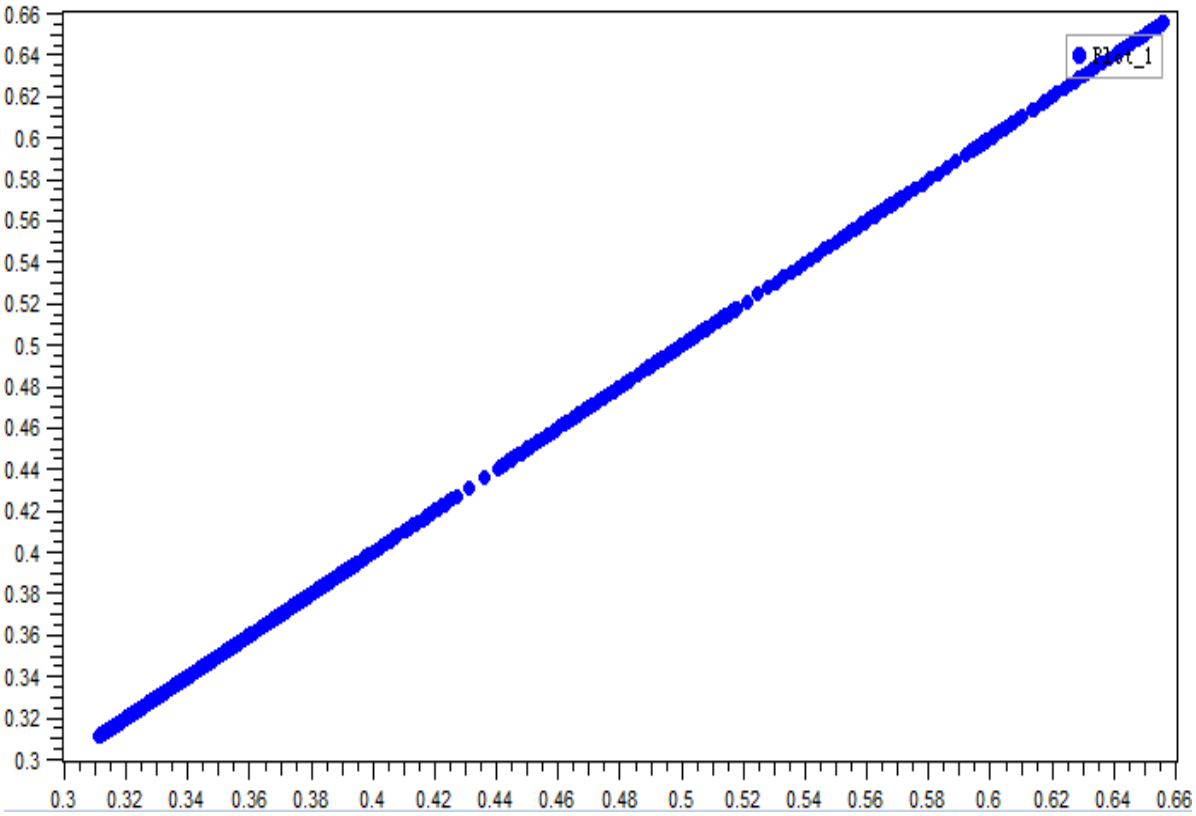


与其他绘图方式比较，数据选择部分多了一项，即需同时添加二个数据。上述二维和三维散点图所述的内部绘图，为一个数据内部的行(样本)或列(变量)间的绘图；而本处所述的内部绘图，则是二个不同数据中，以不同行(样本)或列(变量)分别作为横坐标，或纵坐标所绘制的图形。

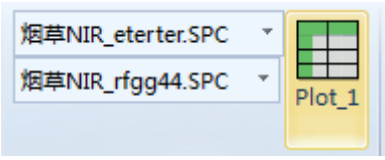
添加数据 X，该数据中的某一行或列作为绘图时的横坐标，初始状态下自动选择第一行，用户可在得到图形后，在图形菜单功能中任意选择其他行或列。添加数据 Y，意义与数据 X 雷同，差异在于该数据作为绘图时的纵坐标。实因 X 与 Y 为成对使用行或列数据，其长度必须相等。添加数据 X 后，与其长度不匹配的数据，将不可被加入到被选数据中。

图形名称则为显示在图形菜单功能的图形名称，以实现快捷达到图形表格。绘图优先性则与上述介绍内容相同，实现行或列优先绘图。

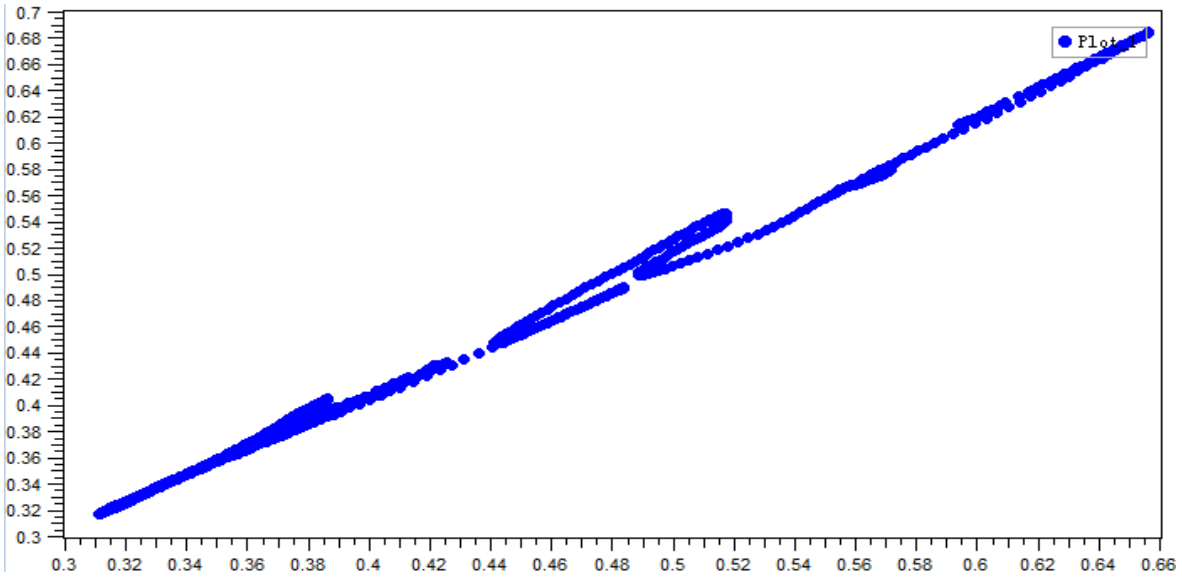
点击取消后，则不进行任何操作，直接返回；若点击确定，则出现如下图所示的结果：



上图为添加数据 **X** 和添加数据 **Y** 时，选择同一数据得到的结果。实因其为同一数据，因此所得图形为一过原点，且每点至 X 轴和 Y 轴长度相等的直线。与其他非自定义图形相比，在图形菜单功能下，同时还增加如下图所示的功能。



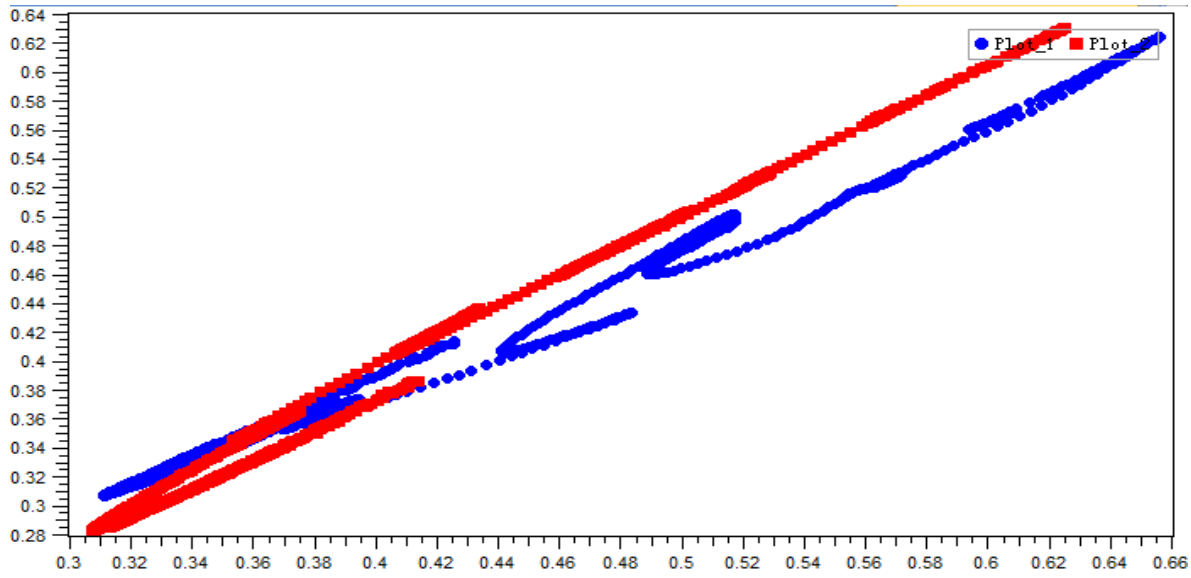
通过上述功能，用户可选择数据 **X** 和 **Y** 中任意行或列，作为所绘图形的 X 轴或 Y 轴，从而实现数据的高度灵活性，比如改变 **Y** 中纵坐标数据后，得到如下图。



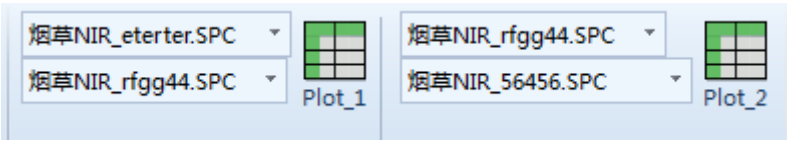
点击图中右侧的表格按钮，则进入绘制图形所得到的数据，即数据 **X** 和 **Y** 中所提取的绘图数据。当然，在数据表格页面，单击右键，并点击切换到图形功能，可再次返回到上述图形。

在图中，若点击内/外部绘图方式中的添加按钮，则得到如下图所示的界面。

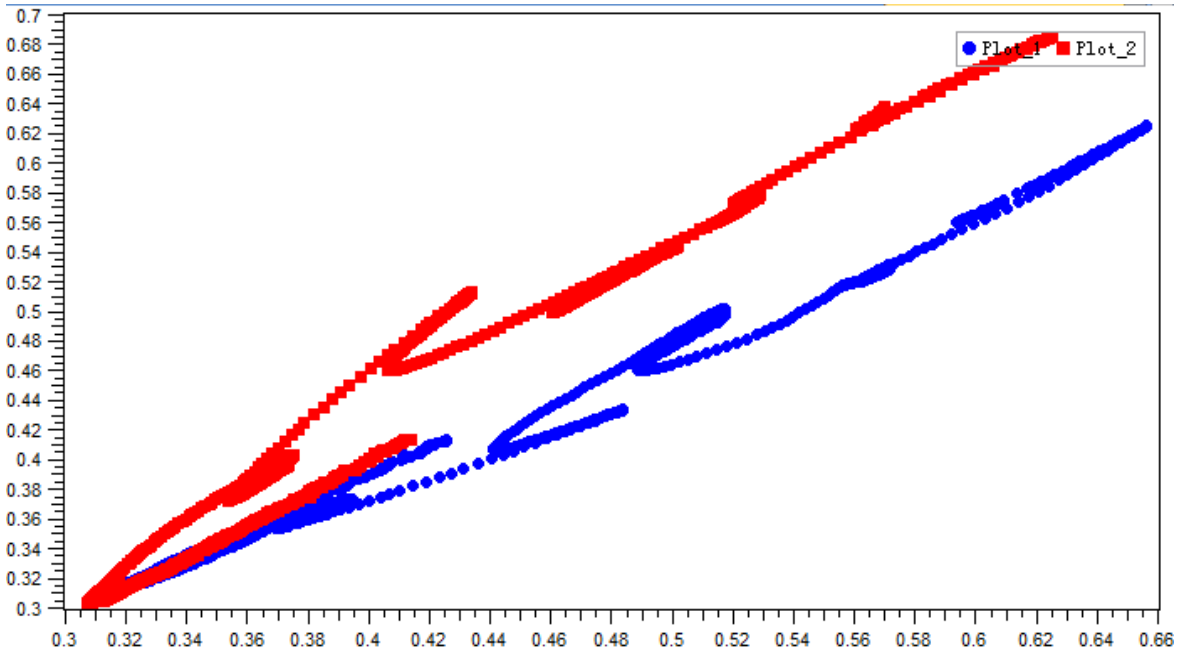
点击确定后，所得图形如下图所示。从图中可以看出，由第一个数据组中第一行数据所得的蓝色散点图和第二个数据组中第一行数据所得的红色散点图并列显示，实现不同数据的可视化比较。



此时，图形菜单功能中，同时添加 Plot_1 和 Plot_2 二个图形的可供选择项，如下图所示。



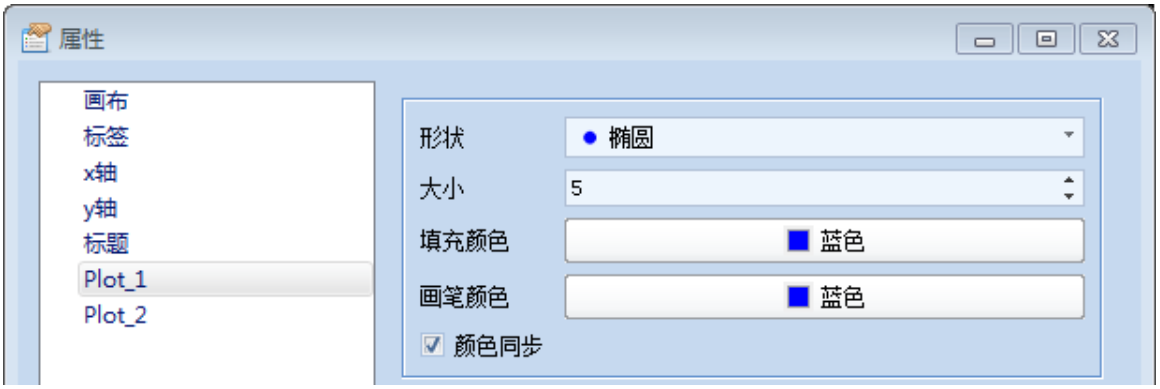
其操作使用与上述单组数据时相同，如修改其中的一个数据，图形变动如下。



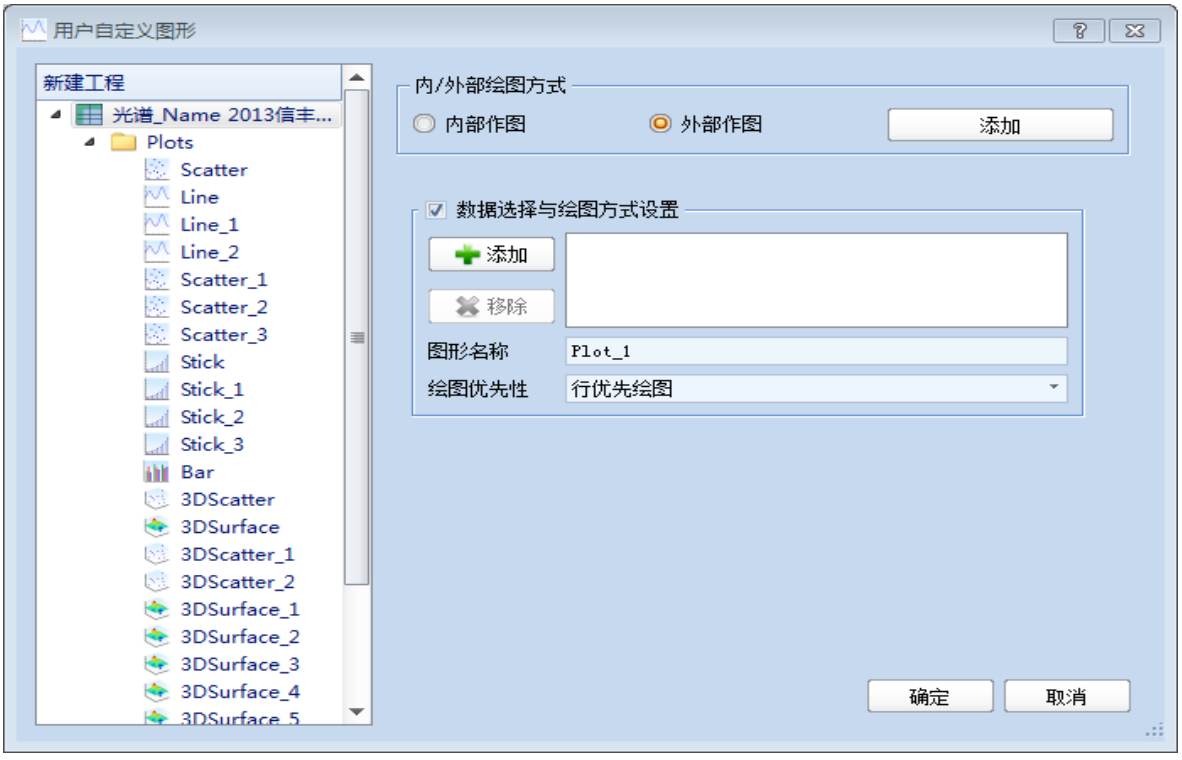
通过使用该按钮，可无限添加绘图数据。

属性修改：此种情形下的属性修改与前述一般图形雷同，在图形中任一位置单击右键，点击属性，便进入属性修改页面。

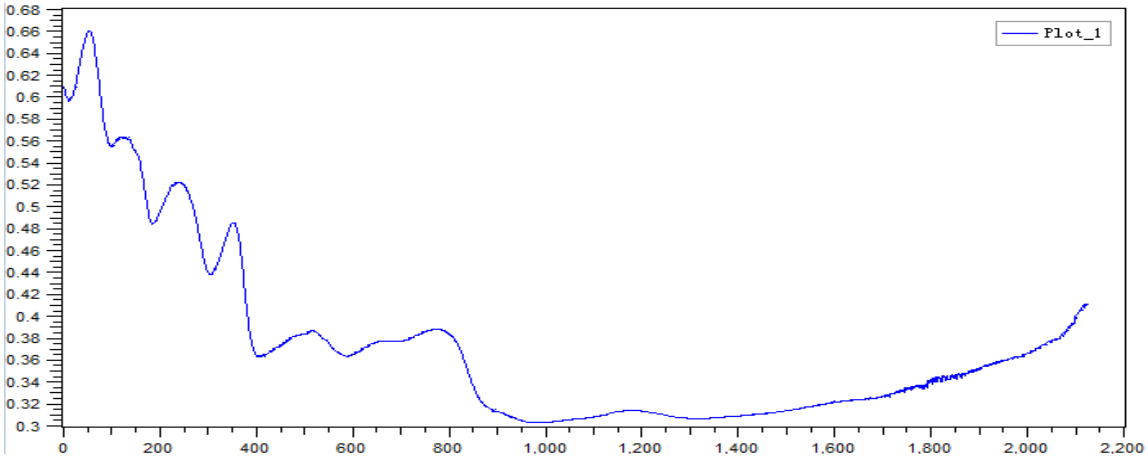
除基本属性修改外，可供修改的图形属性如下图所示，图中显示 Plot_1 散点图的当前属性，用户可做任意修改。



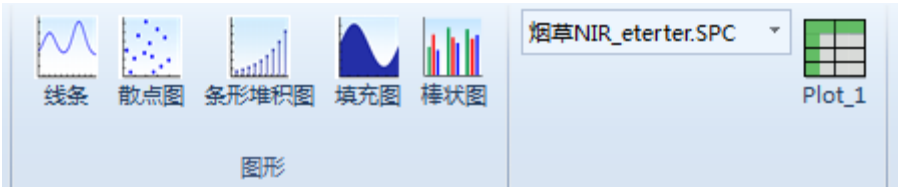
- **第二种情形：若在新出现的界面中选择外部作图，则界面如下图所示。**



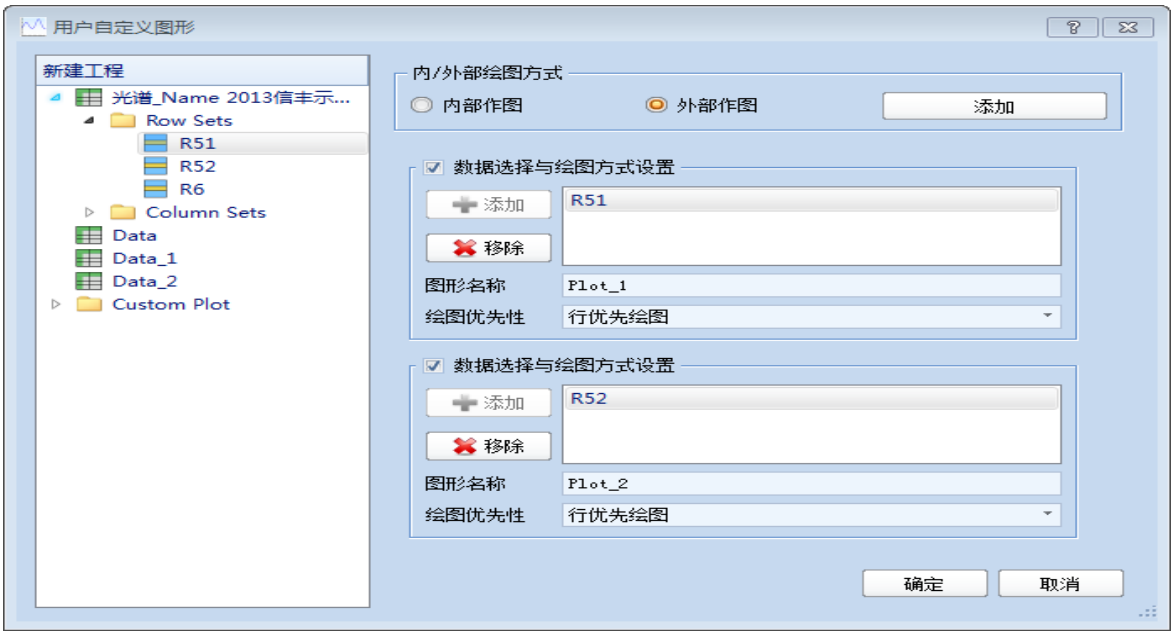
在外部绘图中，除仅需选择一个数据外，其余均与上述内部绘图雷同。所得图形则是所选数据中依照行或列优先所绘制的某一行或列曲线，如下图所示。



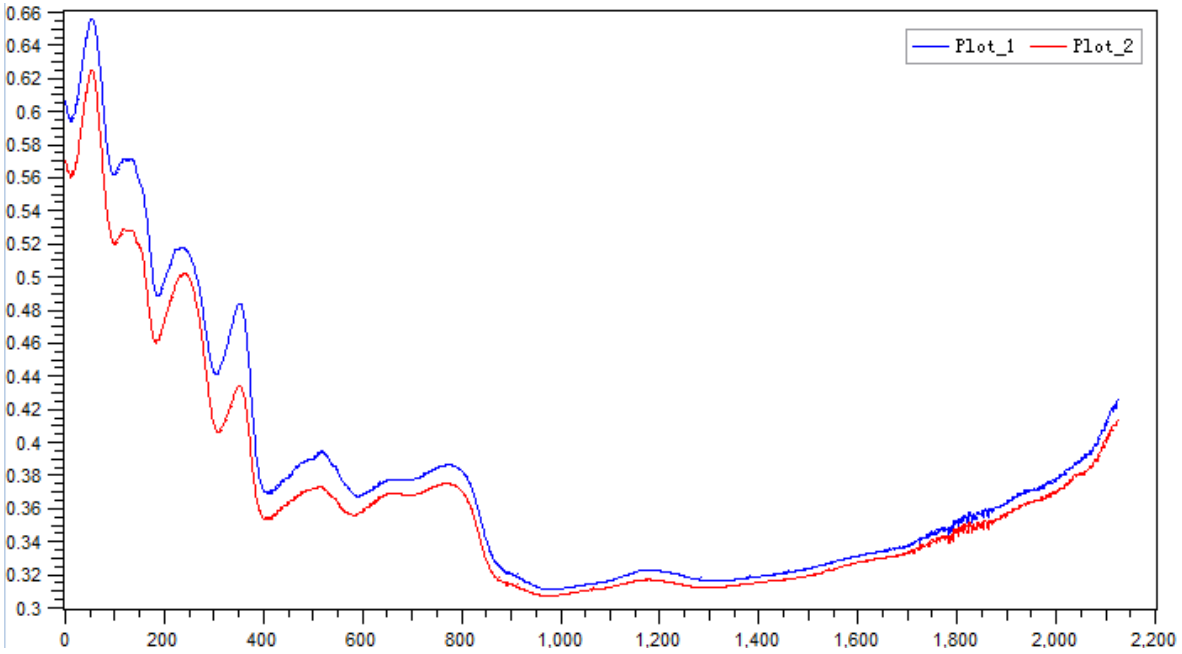
在图形菜单功能中，除图形标注和缩放等基本功能外，亦包括如下图所示的功能，用户可直接点击图形标签，快速将当前数据切换为不同的绘图类型，或选择不同的行或列绘图数据，或者切换到数据界面，在此不再敖述。



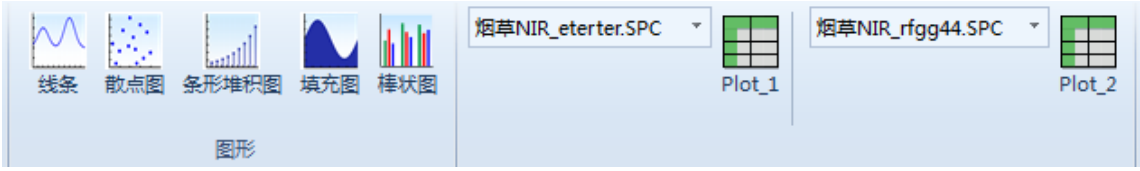
若在图中点击添加按钮，即同时绘制多个不同数据，其界面如下图所示。



如上选择数据后，得到如下图所示图形。

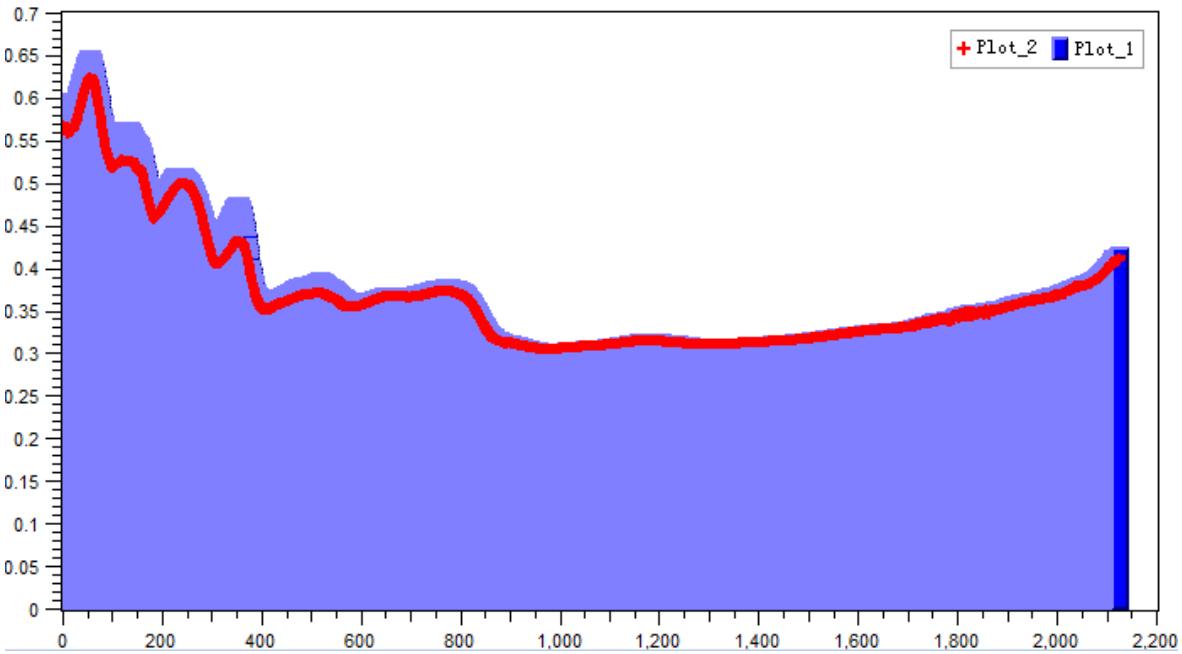


图形菜单功能中则添加如下图所示功能，其操作与前述雷同，其中对绘图类型的改变，将同时作用于图中的所有曲线。



属性修改：此种情形下的属性修改与前一种情况有所不同，进入修改修改页面后，出现如下图所示的图形。

从图中可以看出，通过此种方式所绘制的每条曲线，用户均可分别进行属性修改，从而实现图中不同数据的差异化绘图类型，以实现丰富的可视化表达。如下图所示：



从图中可以看出，分别以填充和散点图的绘图方式，很好地表达了二个不同数据。与此同时，图中所述的图形属性修改，亦同时根据图形的不同而动态变化，以更好地满足不同绘图类型的个性化情形。

自定义绘图从多个方面丰富了本软件可视化绘图的功能，包括数据个数，被选数据中用于实际绘图的数据，以及不同绘图类型的组合和表达方式。

9.3. 右键菜单

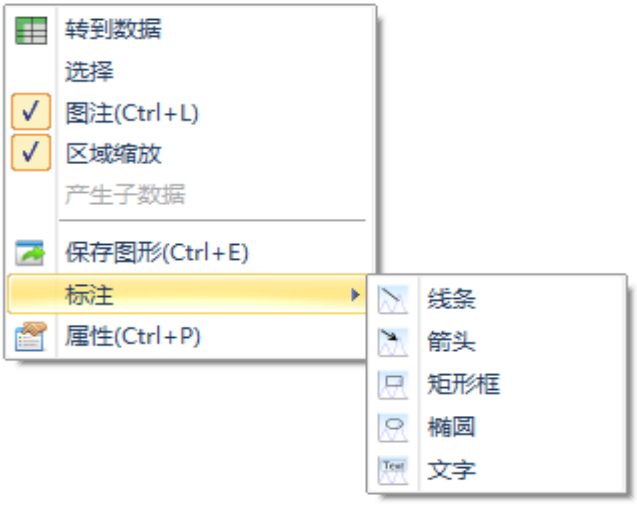
在上述方法所得到的图形中，点击右键可得到如下图所示的菜单功能。这些功能主要可分为如下三类:

1) 图形操作的快捷键：这些功能亦在其他界面出现，此处添加以使用户快速对图形进行某些操作，如选择、图注、区域缩放、标注等。

2) 图形操作的新功能：针对图形进行操作的功能，且在别的界面没有出现过，如转

到数据，保存图形等。

3) 图形属性及其修改：单击可进入图形属性编辑界面。



上述右键菜单功能，可总结为如下表。

序号	操作类型	图标	说明
1	转到数据		图形界面转到作图时所选择的数据界面。转到数据后，数据表中将标记图形中被标记的样本或变量(整行或整列)；从数据表中亦可使用转到图形功能返回。
2	选择	无	当处于选择状态时，可再点击不同标注功能以改变其位置或大小，亦可选择后转到属性对话框。
3	图注	无	显示或隐藏图注，默认状态为显示。图形上的图注亦可点击，此时界面上对应的线将处于选中状态。
4	区域缩放	无	与图形菜单工具栏上对应的功能相同，框选需缩放查看的区域。
5	产生子数据	无	当有样本或变量被标记时，可用被标记的数据产生新的子数据。
6	保存图形		导出当前界面中的图形为 PDF 格式文件。

7	标注	无	与工具栏上对应的功能相同，为当前界面添加标注。
8	属性		转到属性修改对话框。可修改界面上属性内容详见 9.2.中各部分的介绍。

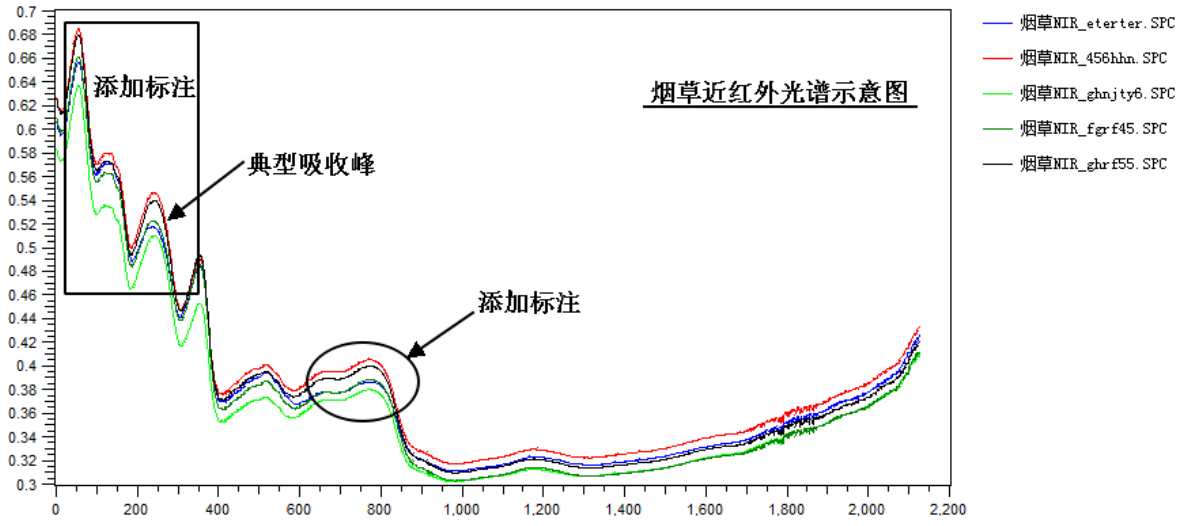
9.4. 工具栏操作

本软件提供丰富图形操作功能，可实现对图形的标注，缩放，以及图形中样本或变量的标记等。

9.4.1. 图形标注

图形标注是指往图中添加注释，以便更好地理解图形，或者用于图形结果的交流目的。本软件支持添加的图形标注包括如下五种：线条，箭头，矩形框，椭圆，以及文字，如下图所示。

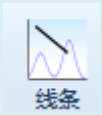
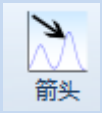
如下图则显示了一个实际的图形添加标注的例子。



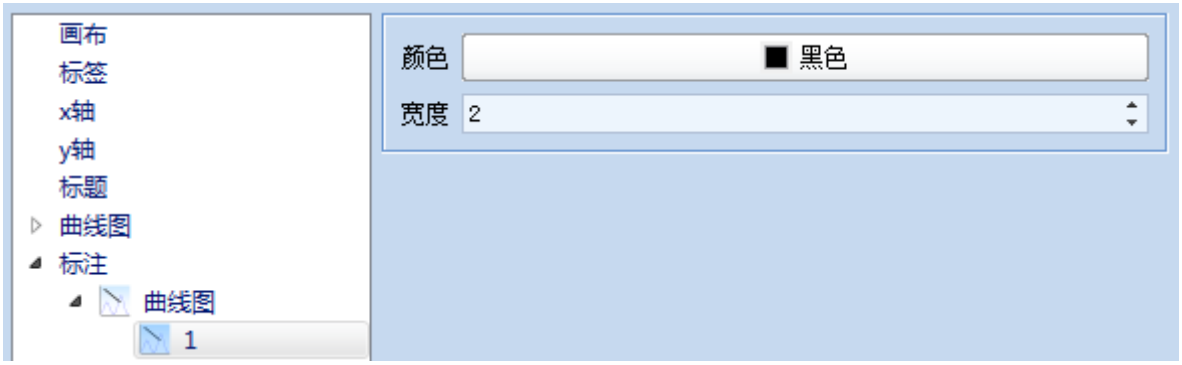
从上图可看出，本软件的图形标注功能非常完备，可满足绝大多数情况下的使用需要。

图形中添加的标注，可使用选择功能使其处于被选择状态，此时便可对其进行位置或形状的动态调整。当要删除标注时，可在界面上选中以便删除，亦可转至到属性对话框，再按下 Delete 键，一次删除多个标注。

上图中添加的标注，可进行丰富的属性修改，不同标注类型可具体被修改的内容，总结如下表所示。

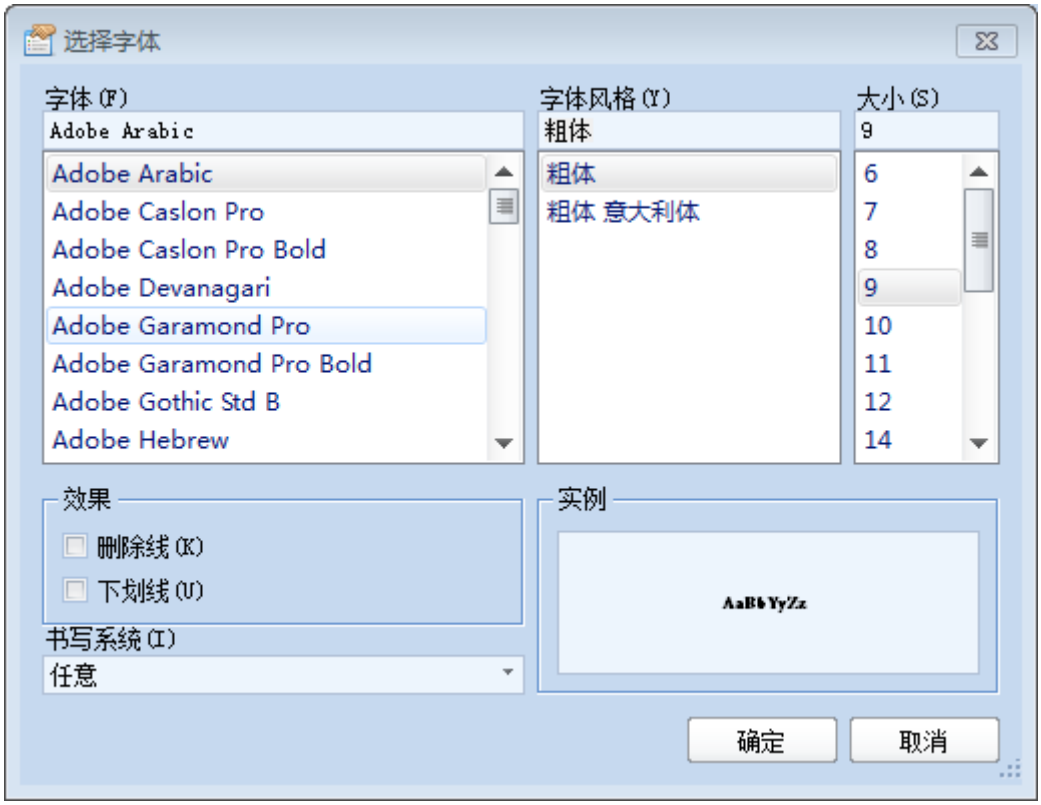
序号	标注类型	图标	可修改形状	可修改颜色	可修改位置
1	线条		是	是	是
2	箭头		是	是	是
3	矩形框		是	是	是
4	椭圆		是	是	是
5	文字		否	否	是

除文字外，其他标注的属性修改对话框如下图所示。

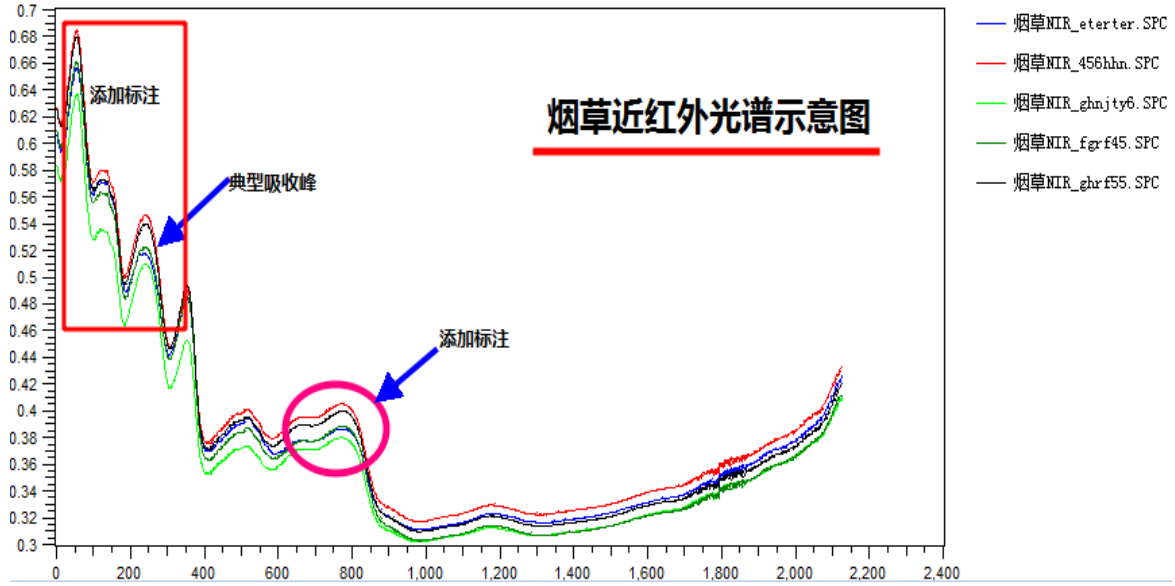


标注位置和大小的修改，则当其处于选中状态时，直接拖动和缩放即可。

文字属性修改对话框则如下图所示。



图中所示的图形，其标注进行部分属性修改后，可得到如下所示图形。



属性修改界面的操作，将即时反应在图形上。

9.4.2. 放大/缩小

本软件提供的图形缩放功能亦很丰富，具体如下图所示。用户点击进入图形界面，图形工具栏，包括缩放功能将即时出现在图形菜单功能下。

上图中所示的图形缩放功能，详解于如下表。

序号	操作类型	图标	说明
1	区域缩放		框选需要放大查看的图形区域。
2	缩放重置		恢复到原图大小，双击鼠标亦可恢复到原始大小。

3	后向缩放		撤销到当前放大查看的上一步缩放操作，重复使用此功能则回到图形的原始大小。
4	前向缩放		后向缩放的反操作，即恢复至前一步的缩放撤销操作。
5	图形平移		整体拖动图形画布。

若被缩放的图形中已添加标注，则使用缩放功能时同时对标注进行缩放。

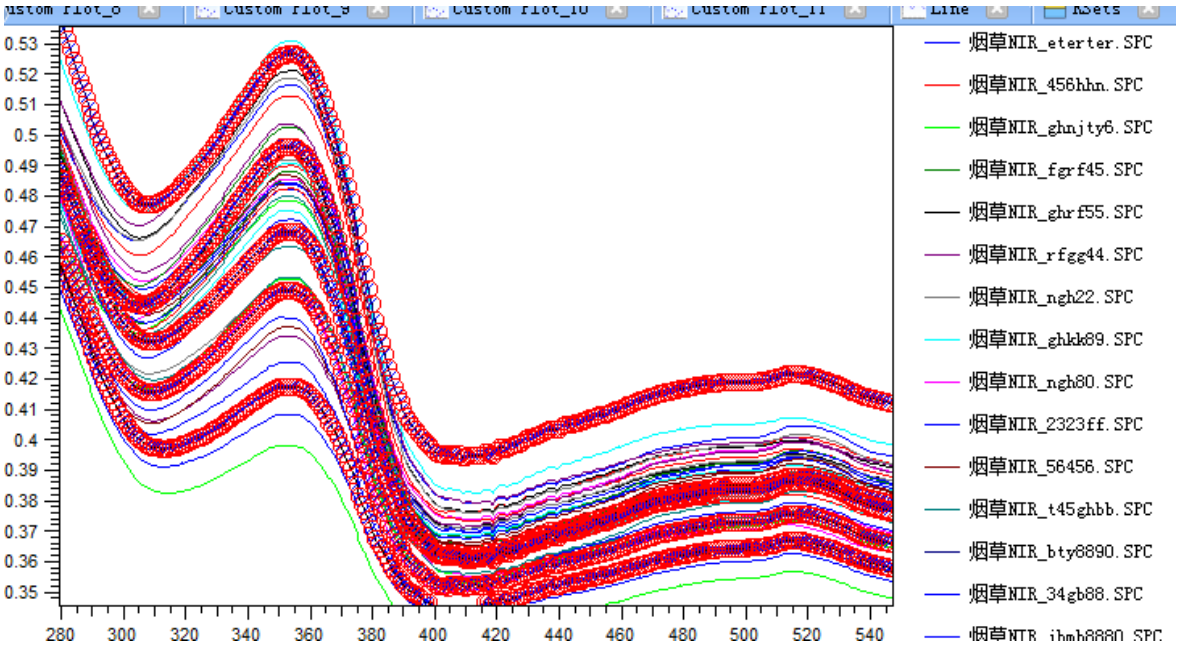
9.4.3. 图形标记

图形标记是指通过鼠标选中图形中的数据点(样本或变量)，并以添加圆圈的形式将被选数据标记出来。进入图形后，在图形菜单功能中，包括如下图所示的标记功能。

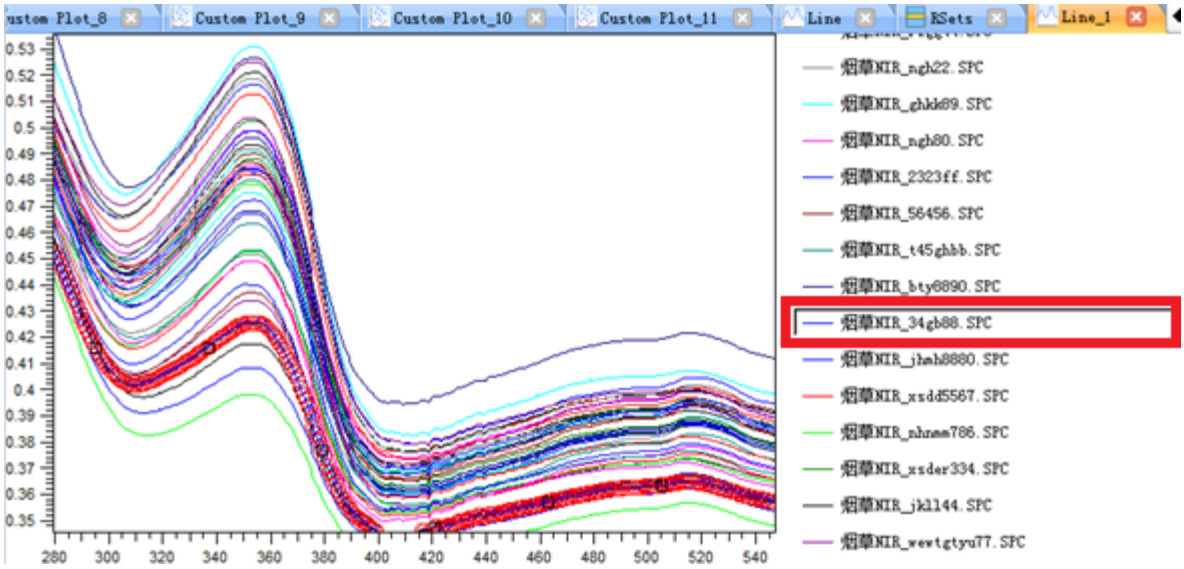
用户可根据实际需要，首先可在下拉式选项中选择标记样本或标记变量，选择完成后，其后的标记则仅根据该选择进行。

本软件规定行为样本，列为变量，而在绘制图形时，亦可选择行优先或列优先绘图。显然用户所得到的图形，每一条曲线可能是样本(某样本的所有变量)，亦可能是变量(某变量的所有样本)。

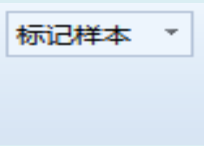
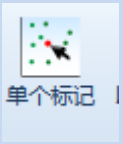
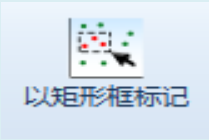
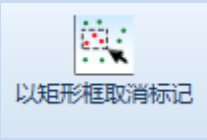
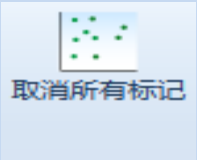
如下图所示，即为图形标记的示例情形。



此外，在产生上述图形后，图形右侧同时显示其图注，用户点击图中线条后，其对应的图注亦处于选中状态；反之通过点击图注，亦可实现图形中线条的选择。如下图所示，图形圆圈所选线条与被选中的图注具有一一对应关系。



上图中的图形标记功能，具体描述为如下表。

序号	操作类型	图标	说明
1	标记样本/标记变量		选择标记样本或变量，当所选项改变时，当前界面已有标记将会被清除，且不可恢复。
2	单个标记		可通过鼠标在图中点击待标记的项来(图形)以选中目标。
3	以矩形框标记		可同时选中多个分组，亦可通过拖动光标定义矩形左上角和右下角以确定矩形范围。
4	以矩形框取消标记		与以矩形标记操作的反操作，被框选的样本或变量将被取消标记。
5	反向标记		当图中已有标记时，使用该选项则自动标记未被标记的项目(图形)，清除标记已被标记的项目(图形)。
6	取消所有标记		清除所有已经做出的图形标记。

对图形标记，另一个重要的意义在于以可视化的方式，实现图中数据的样本选取或变量选择，类似于 6.3.1.所述创建子数据，只是基于图形的子数据产生方式，可更直观方便地达到目的。

图形中标记样本或变量后，右键菜单中用被选数据点产生新的子数据功能即进入可用状态，点击所得结果与 6.3.1.所述相同。

模型结果图形中，图形标记功能更可使用标记或未被被标记的样本或变量产生子数据，或重新建模。

第十章　预处理

本章主要介绍本软件数据处理方法的使用和操作步骤，以及图形结果的解释。使用数据预处理方法，可从数据样本或变量方向对数据进行某个转换，以提高数据本身的质量，或者模型结果的准确度与泛化能力。

10.1. 整体介绍

本软件集成丰富的数据预处理方法，其功能在 2.1.7.中已做初步介绍，本章则详细介绍各具体方法的使用步骤以及结果解释等，以期用户可更好地使用方法并理解其结果。

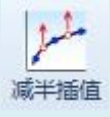
本章所叙数据预处理方法可减弱或去除数据中干扰或噪声信息的影响，减少模型的复杂度，并提高数据或模型的可解释性，比如导数光谱可减少基线偏移及倾斜效应，并突出光谱间的微小差异；散射校则常正用于漫反射光谱的分析中，以减少光散射或光路长度导致的差异。

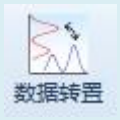
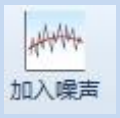
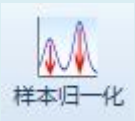
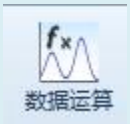
数据预处理方法的内容介绍详情请参见第十章。

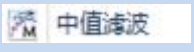
10.1.1. 概述

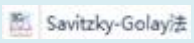
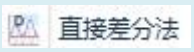
详细介绍数据预处理方法的具体使用步骤前，在如下表中初步说明各方法的基本情况。

表 数据预处理方法初步说明。

序号	功能名称	图标	说明
1	减半差值		根据减半因子的设置，从数据样本或变量方向减少原始数据量。

2	通用插值		根据用户设定的坐标名称，从数据样本方向，以设定任意插值间距的方式，对数据进行插值操作。
3	数据转置		如上所述，本软件约定行方向为样本，列方向为变量。数据转置则是转换数据的行与列方向。
4	加入噪声		根据设定参数，往数据中加入噪声。
5	样本归一化		从样本方向对数据进行归一化处理。
6	变量标度化		从变量方向对数据进行标度化处理。
7	SNV 变换		标准正态变量变换，从样本方向对数据进行中心化和标度化处理。广泛用于光谱数据处理，可去除光谱散射效应。
8	Quantile 标准化		从行方向对数据进行标准化处理，使所有样本的经验分布一致，可很有效地去除数据样本间的背景差异(假定分布一致)。
9	数据计算		从数据样本或变量方向，基于各种数学运算规则产生新的样本或变量(亦可替换旧的样本或变量)。

10	**平滑**		
	移动平均法		计算用户自定义范围内数据的数学均值，用以替换目标数据点。
	高斯滤波		以加权移动平均法的方式，计算用户自定义范围内的数据，用以替换目标数据点。
	中值滤波		获得用户自定义范围内数据的中位数，用以替换目标数据点。
	Savitzky-Golay 平滑		采集目标数据点以及该点左、右固定尺寸窗口内的数据点，以及各点序号值(负、零、正序号)，以多项式方法拟合这些数据，最后基于解析后的方程计算平滑目标数据点。
	惩罚最小二乘平滑		同时基于最小二乘和粗糙惩罚项构造目标函数以平滑数据，增加粗糙惩罚项在于粗糙量测数据中的噪音成分很大，在平滑中亦应受到更大的惩罚。
11	**求导**		
	Gap 法		Gap-Seg 法求导的特殊情形，其中分割尺寸为 1。该方法要求被处理数据无缺失值，单样本含有 5 个以上变量，且均为数值。
	Gap-Seg 法		导数计算时以数据二侧一定窗口尺寸内数据均值之差的替换原始数据值，且二数据中间由一段

			数据分隔。
	Savitzky-Golay 法		基于局部分割窗口，而非邻近点计算某特定数据点下的 *N* 阶导数，实际上使用了平滑后的数据，以克服噪声的影响。
	直接差分法		直接采用相邻数据点获得导数，计算简单便捷，但噪声对求导的影响较大。
12	**背景扣除**		
	airPLS 法		迭代加权拟合基线与原始信号，每次迭代计算中以自适应加权惩罚的方式对拟合基线与信号间误差平方和重加权，直至达到迭代中止条件。本法快速，使用灵活，结果较好。
	手动法		手动选择背景，基于线性模型构造并扣除背景。
	airPLS 与手动联合法		同时基于上述二种方法联动扣除背景，即在 airPLS 方法扣除背景的基础上，程序继续使用手动法扣除背景。
	线性补偿法		原始数据扣除一个固定的已知背景。本软件中整体减去每点变量最小值。
13	**漂移校正**		
	手动法		手动选择校正的样本，并选择该样本中的校正目标点，其他样本则依次对照这些点进行局部线性

			校正。
	COW 法		基于数据分割的信号漂移校正方法，事先无需做任何预处理或参数估计，使用非常广泛。缺点是其计算过程比较费时，且不能处理含有缺失值和非数值的数据。
14	多元散射校正		最初用于补偿光谱数据的加和与乘积效应，亦可用于漂移补偿和部分消除干扰信号等。
15	正交信号校正		用于消除自变量 **X** 中与响应变量 **Y** 无关的信息，即去除数据 **X** 中的无关方差，以构建更稳健可靠的回归模型。
16	去趋势化		用于去除光谱数据中的非线性趋势。

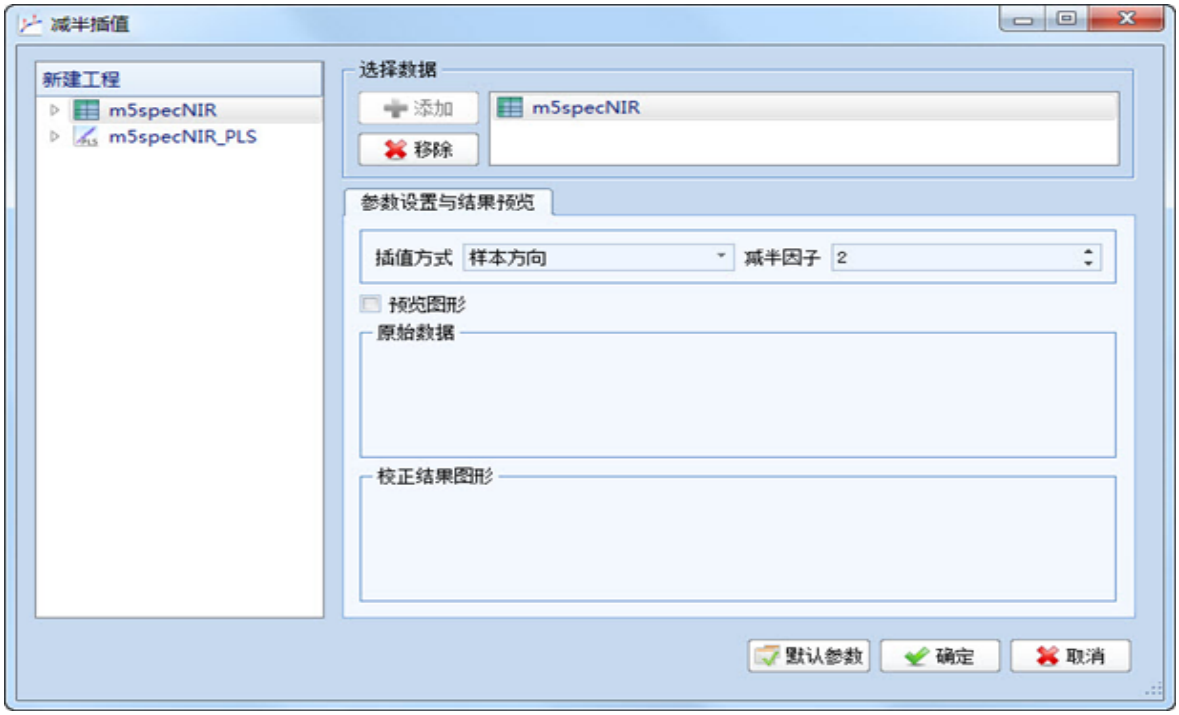
10.1.2. 通用步骤(归纳)

数据预处理部分有一些相同或相似的操作步骤，可先归纳出来做统一介绍。在介绍具体方法时，这些雷同的内容将不再赘述。

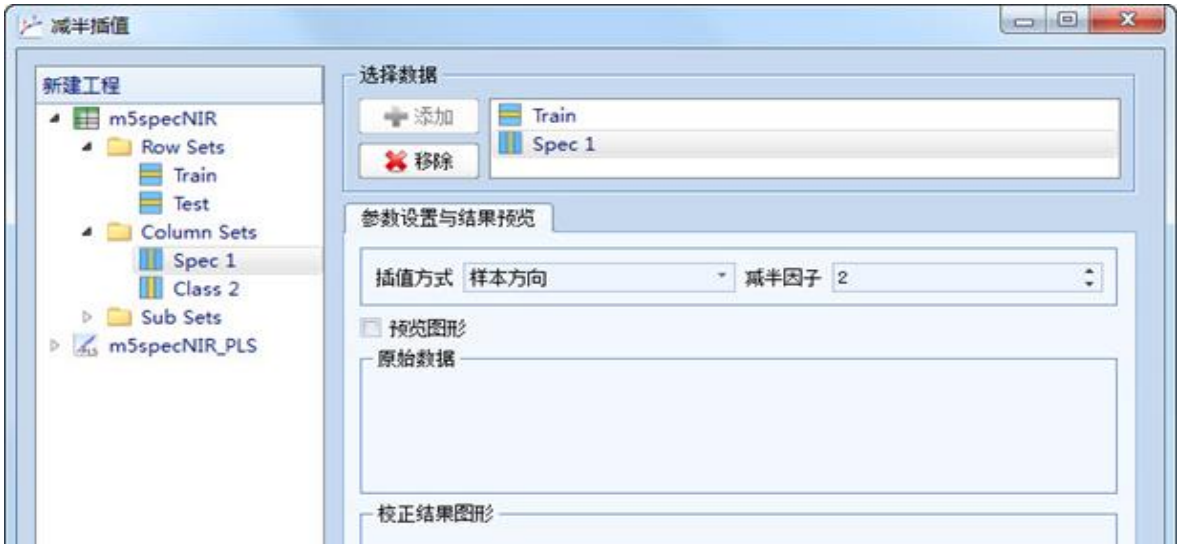
概括起来，这些共同的内容可归纳为选择数据，设置参数，预览图形(可跳过)，确定几个步骤，下面一一进行介绍。

1) 选择数据：数据选择是数据预处理的第一步，程序运算即是对被选数据的分析处理，被选数据可分为如下三种不同情形：

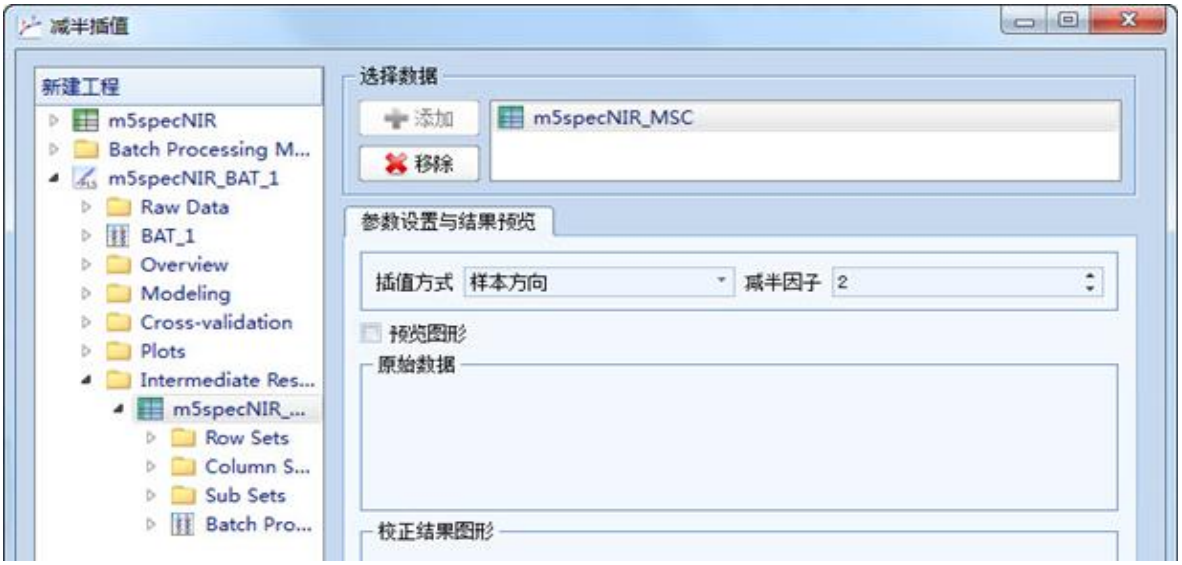
- 单个数据：指可直接被分析的数据，包括基本数据表，行或列划分数据及子数据。如下图所示，以减半差值为例，直接添加 m5specNIR 数据进行分析。



- 同源数据组：指一个行划分和一个列划分数据组，并该二数据来源相同，即来自于同一数据矩阵的划分。系统实际使用的数据则是同源数据组的公共交叉数据。如下图所示，行划分数据 Train 和列划分数据 Spec 1 均来自于基本数据 m5specNIR。在分析时，则使用 Train 和 Spec 1 的交叉部分数据。



- 结果矩阵：指建模运算后所产生的结果节点数据，如下图中使用算法流 BAT_1 获得模型中间结果 m5specNIR_MSC。



2) 默认参数：选择数据后，即可修改参数以获得最佳计算结果。用户亦可点击 默认参数 按钮，自动恢复方法参数为系统默认值。

3) 预览图形：勾选 预览图形 复选框，可即时查看原始数据，以及基于当前参数获得的计算结果，方便用户比较，并在需要的时候再次修改参数，计算新的结果，如下图所示。

用户可使用框选图形的功能，同步缩放原始数据及处理后的结果图形，即上下二图可同步缩放。

此外，用户亦可将鼠标置于任一图形上，通过滑动鼠标以缩放当前图形，而不是同时缩放二个图形。双击图形则可复原图形，恢复至初始状态。

4) 确定：点击 

按钮，若预处理计算成功，则数据结果将在工程导航栏中作为新的数据节点显示；若失败，则提示用户。若点击 

按钮，则取消操作，关闭对话框。

接下来依次介绍各预处理方法的使用与操作。

10.2. 减半插值

减半插值功能可同时用于处理样本和变量。若从样本方向处理数据，可直接去除重复的样本；若从变量方向处理数据，则可减少变量的数目。在数据处理过程中，实时使用该功能，可达到如下目的：

- 可能提高结果的准确性(如感官评价)。
- 获得更可靠的结果。
- 提高数据信噪比(如均值化处理同一样本的重复实验数据)。
- 结果更容易解释。
- 减少程序的运行时间(如含大量无信息变量的数据)。

操作步骤:

步骤 1: 点击预处理标签 -> 减半插值，弹出如下对话框:

接下来的操作步骤参照预处理之通用步骤。参数说明见下表:

参数	范围	说明
插值方式	样本方向或变量方向。	插值样本或插值变量。
减半因子	[2 num / 2]，其中 num 表示所选数据的长度。	插值后留下数据点的长度因子，比如设定为 2，则表示插值后的数据为原数据 1/2。

示例数据及结果：如下三图分别为原始数据，以及从样本和变量方向对数据进行减半插值的结果，其中减半因子设定为 2。系统默认从第一行或列数据开始插值，因此，若数据长度为基数时，则剩余数据的长度为 num / 2 + 1。

EJiao_FOSSNIR_2Class.mat | SSets | RSets | CSets | SSets_1 | SSets_2 | SSets_3

V	Var_1	Var_2	Var_3	Var_4	Var_5	Var_6	Var_7	Var_8	Var_9	Var_10
	1	2	3	4	5	6	7	8	9	10
1	0.62817	0.63456	0.64089	0.64714	0.65328	0.65926	0.66507	0.6707	0.67612	0.68133
2	0.59689	0.60243	0.60793	0.61333	0.61862	0.62377	0.62875	0.63355	0.63815	0.64255
3	0.50702	0.51137	0.51568	0.51993	0.52408	0.52812	0.53203	0.53579	0.5394	0.54284
4	0.68611	0.693	0.69984	0.7066	0.71324	0.71974	0.72607	0.73222	0.73817	0.74391
5	0.59479	0.60089	0.60694	0.6129	0.61874	0.62444	0.62997	0.63531	0.64044	0.64537
6	0.72971	0.73545	0.74115	0.74678	0.7523	0.7577	0.76295	0.76804	0.77295	0.77767
7	0.70737	0.713	0.71859	0.7241	0.72951	0.73478	0.7399	0.74486	0.74963	0.75421
8	0.67677	0.68271	0.68861	0.69442	0.70012	0.70568	0.71107	0.71629	0.72132	0.72615
9	0.68377	0.68929	0.69477	0.70017	0.70547	0.71062	0.71563	0.72046	0.72511	0.72957
10	0.60782	0.6137	0.61952	0.62526	0.63089	0.63637	0.64168	0.64681	0.65174	0.65646

EJiao_FOSSNIR_2Class.mat | SSets | RSets | CSets | SSets_1 | SSets_2 | SSets_3

V	Var_1	Var_2	Var_3	Var_4	Var_5	Var_6	Var_7	Var_8	Var_9	Var_10
	1	2	3	4	5	6	7	8	9	10
1	0.61253	0.618495	0.62441	0.630235	0.63595	0.641515	0.64691	0.652125	0.657135	0.66194
2	0.596565	0.602185	0.60776	0.613265	0.61866	0.62393	0.62905	0.634005	0.638785	0.643375
3	0.66225	0.66817	0.674045	0.67984	0.68552	0.69107	0.69646	0.701675	0.706695	0.71152
4	0.69207	0.697855	0.7036	0.70926	0.714815	0.72023	0.725485	0.730575	0.735475	0.74018
5	0.645795	0.651495	0.657145	0.662715	0.66818	0.673495	0.678655	0.683635	0.688425	0.693015

SSets | RSets | CSets | SSets_1 | SSet

V	Var_1	Var_3	Var_5	Var_7	Var_9
	1	2	3	4	5
1	0.631365	0.644015	0.65627	0.667885	0.678725
2	0.59966	0.61063	0.621195	0.63115	0.64035
3	0.509195	0.517805	0.5261	0.53391	0.54112
4	0.689555	0.70322	0.71649	0.729145	0.74104
5	0.59784	0.60992	0.62159	0.63264	0.642905
6	0.73258	0.743965	0.755	0.765495	0.77531
7	0.710185	0.721345	0.732145	0.74238	0.75192
8	0.67974	0.691515	0.7029	0.71368	0.723735
9	0.68653	0.69747	0.708045	0.718045	0.72734
10	0.61076	0.62239	0.63363	0.644245	0.6541

10.3. 通用插值

以更一般的方式对数据进行插值操作，即用户可设定任意的插值间距，在原始数据的起点到终点的范围内进行插值。通过此种插值方式，可很灵活地转换数据坐标，比如将二个化学坐标不完全一致的数据(采样频率和分辨率等实验条件的差异)，转换成相同的情形以方便处理，这在色谱、质谱)和光谱数据的分析中均可能涉及到，以转换保留时间、m/z 和波长等坐标。

数据插值的方法很多，包括线性、多项式或样条函数插值等。本软件则采用线性插值的方式，即假定二个被插值点中间存在线性关系，先构造二点的线性拟合函数，再在该线条上添加需要插入的数据点。对被选数据进行插值运算，可增加或减少原始数据的数据点数。

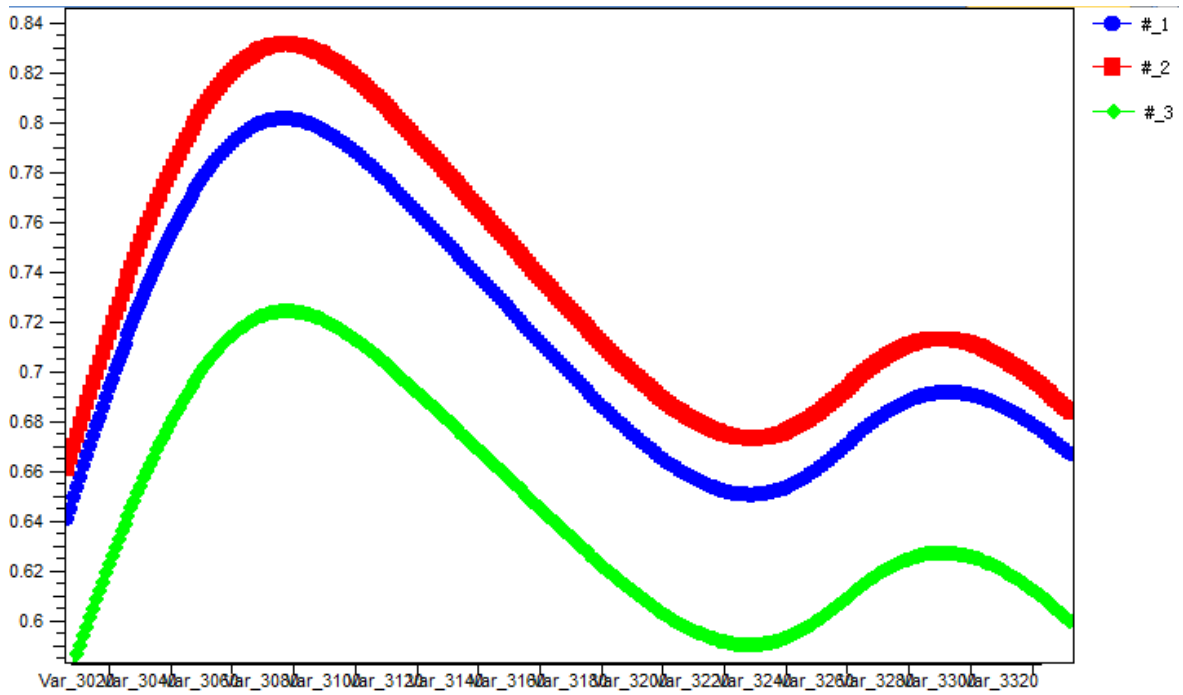
操作步骤:

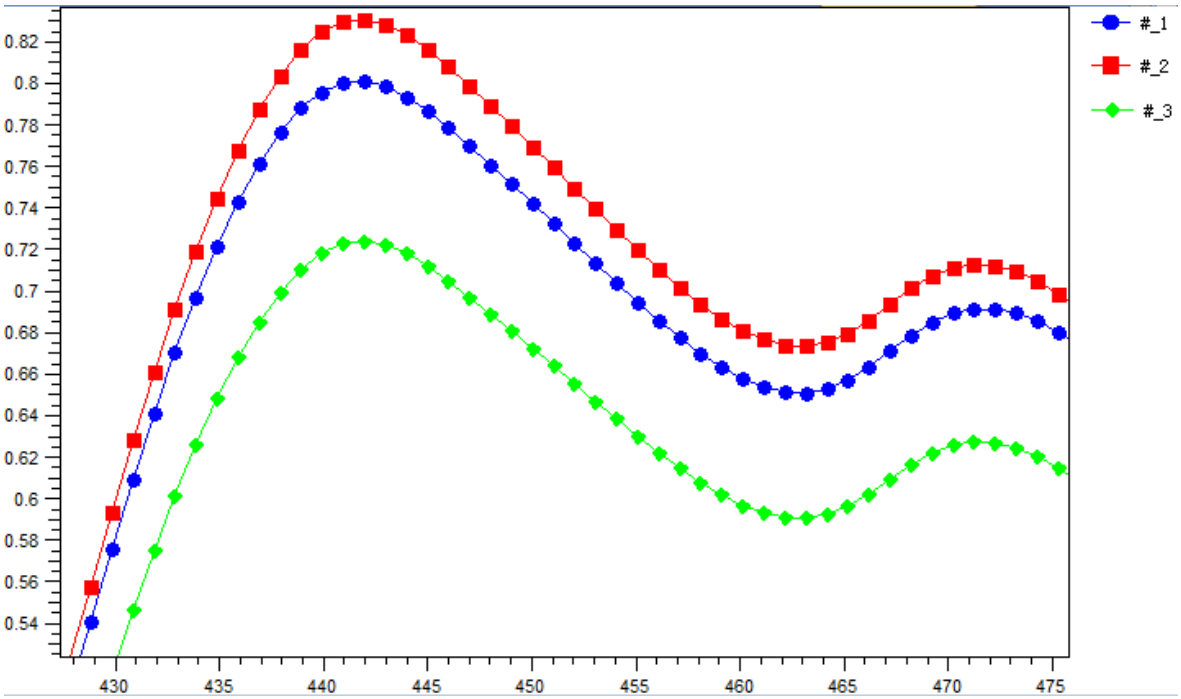
步骤 1: 点击**预处理标签** -> **通用插值**，弹出如下对话框:

接下来的操作步骤参照预处理之通用步骤。参数说明见下表:

参数	范围	说明
标签名称	下拉列表，选择 X 坐标轴名称。	用户可能定义多个 X 坐标(如化学与数学坐标)，则此处可选择。
起始点坐标	[min(x) max(x))，其中 X 表示 X 坐标。	可选择从最小值到最大值。
插值间距	(0 ∞]，其中∞表示无穷大。	指从起始点坐标到结束点坐标中间每二个点间的距离。
结束点坐标	(min(x) max(x)]，其中 x 表示 x 坐标。	可选择从最小值到最大值。

示例数据的一段图形及其结果如下图所示，原始数据共有 4200 个数据点，插值间距为 7，插值后的数据为 600 个数据点。





从图中可以看出，尽管数据点减少了 6/7，但原始图形中的信息还是得到了完整保留。

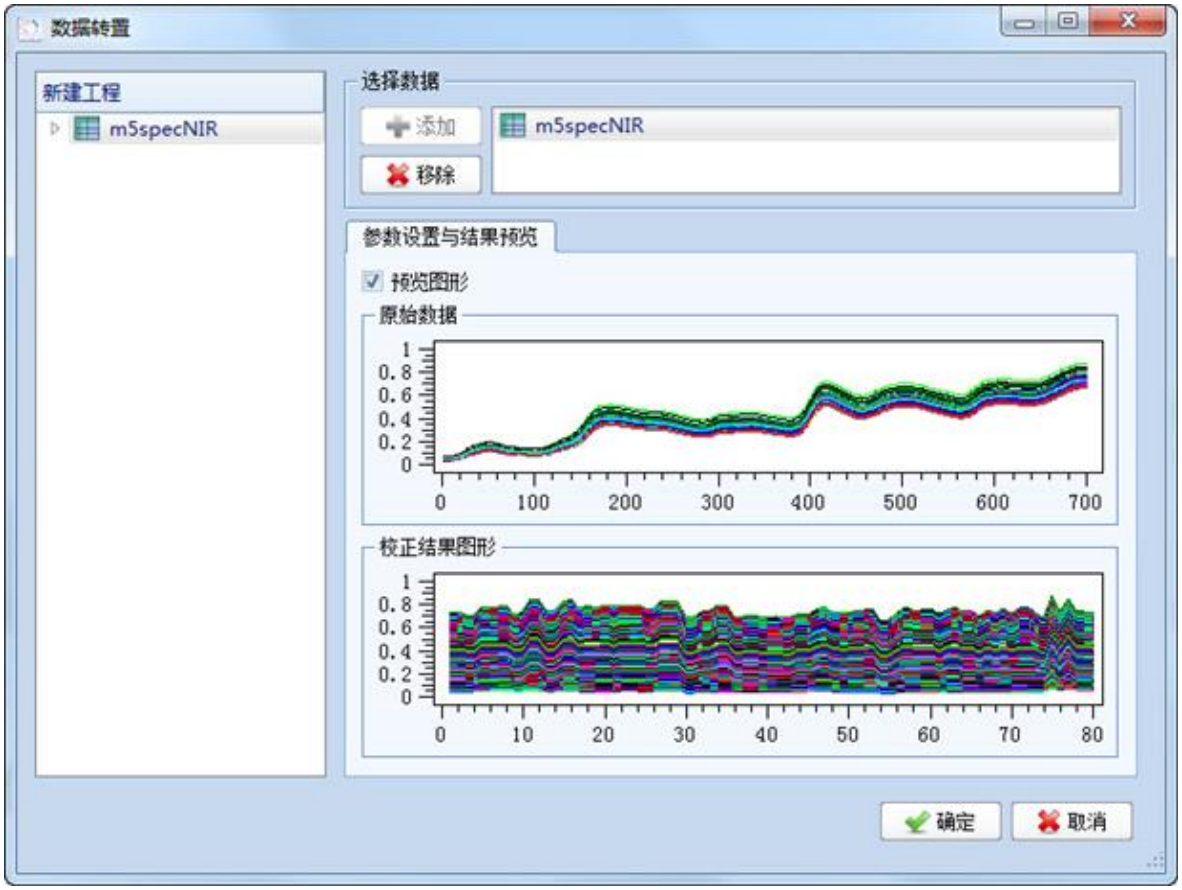
10.4. 数据转置

如前所述，本软件在进行数据分析时，约定数据行方向为样本，而列方向则为变量。若原始数据不符合该约定，则用户可数据载入时交换(转置)其行与列的方向，亦可在数据载入完成后，使用本部分所述功能完成数据的转置操作。

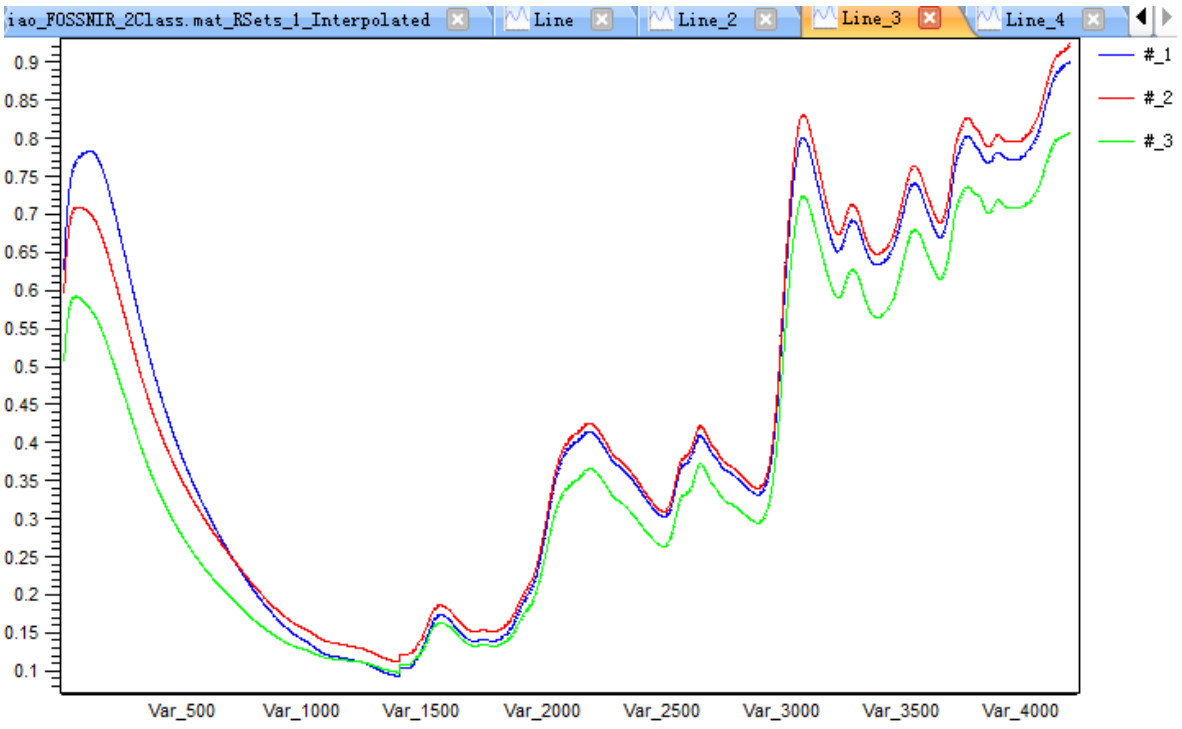
本软件在管理和分析数据的同时，亦包含丰富对数据样本(行)和变量(列)进行说明描述的功能，在进行数据转置时，将同时转置数据对应的说明信息。

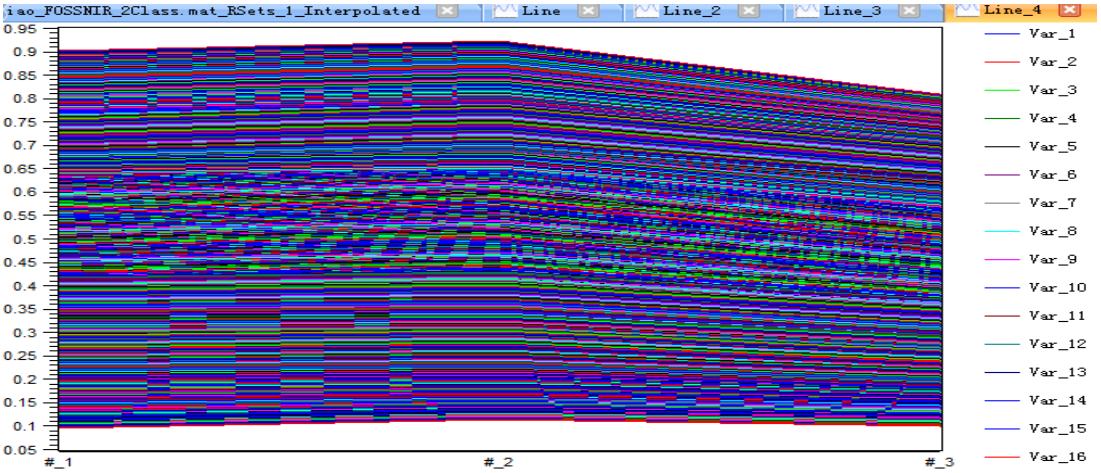
操作步骤:

步骤 1: 点击**预处理标签** -> **数据转置**，弹出如下对话框:



接下来的操作步骤参照预处理之通用步骤。转置前后的数据，所表达的意义完全不同，图形结果亦完全不同，如下二图所示。





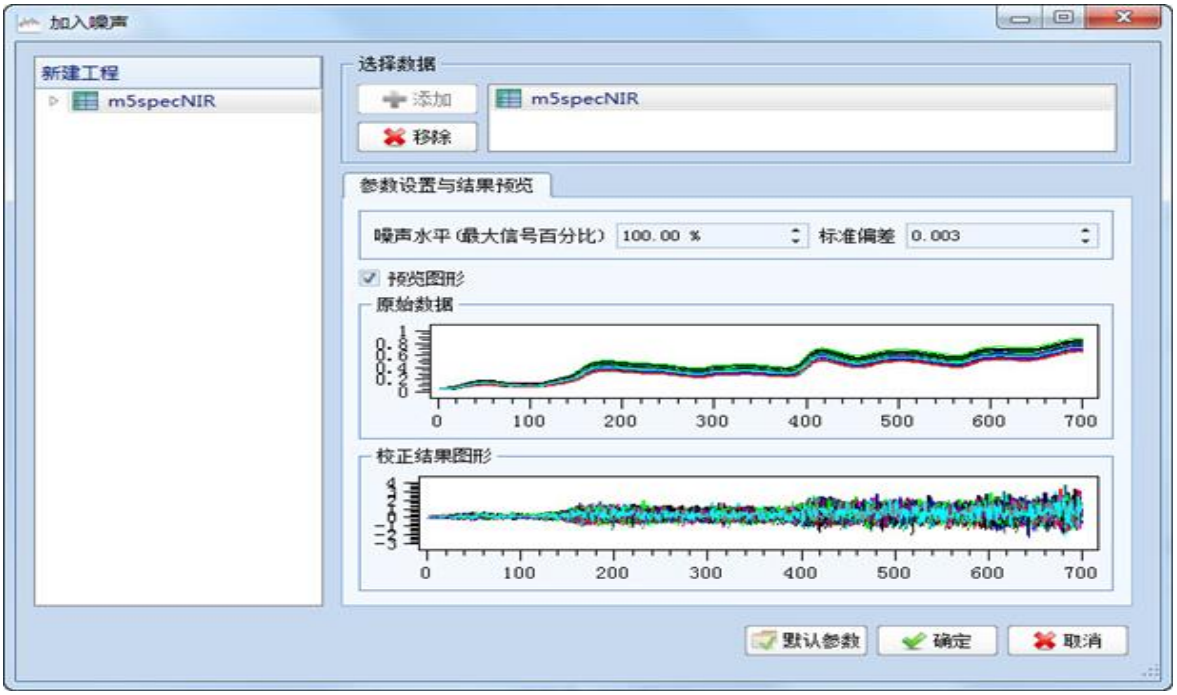
10.5. 加入噪声

对被选数据加入一定程度的噪声。

向数据中加入噪声，其目的在于考察其对模型结果稳健性和可靠性的影响。很显然，其他数据预处理与转换方法，在于提高数据质量或结果的准确性，而增加噪声则将降低分析结果的准确性。

操作步骤:

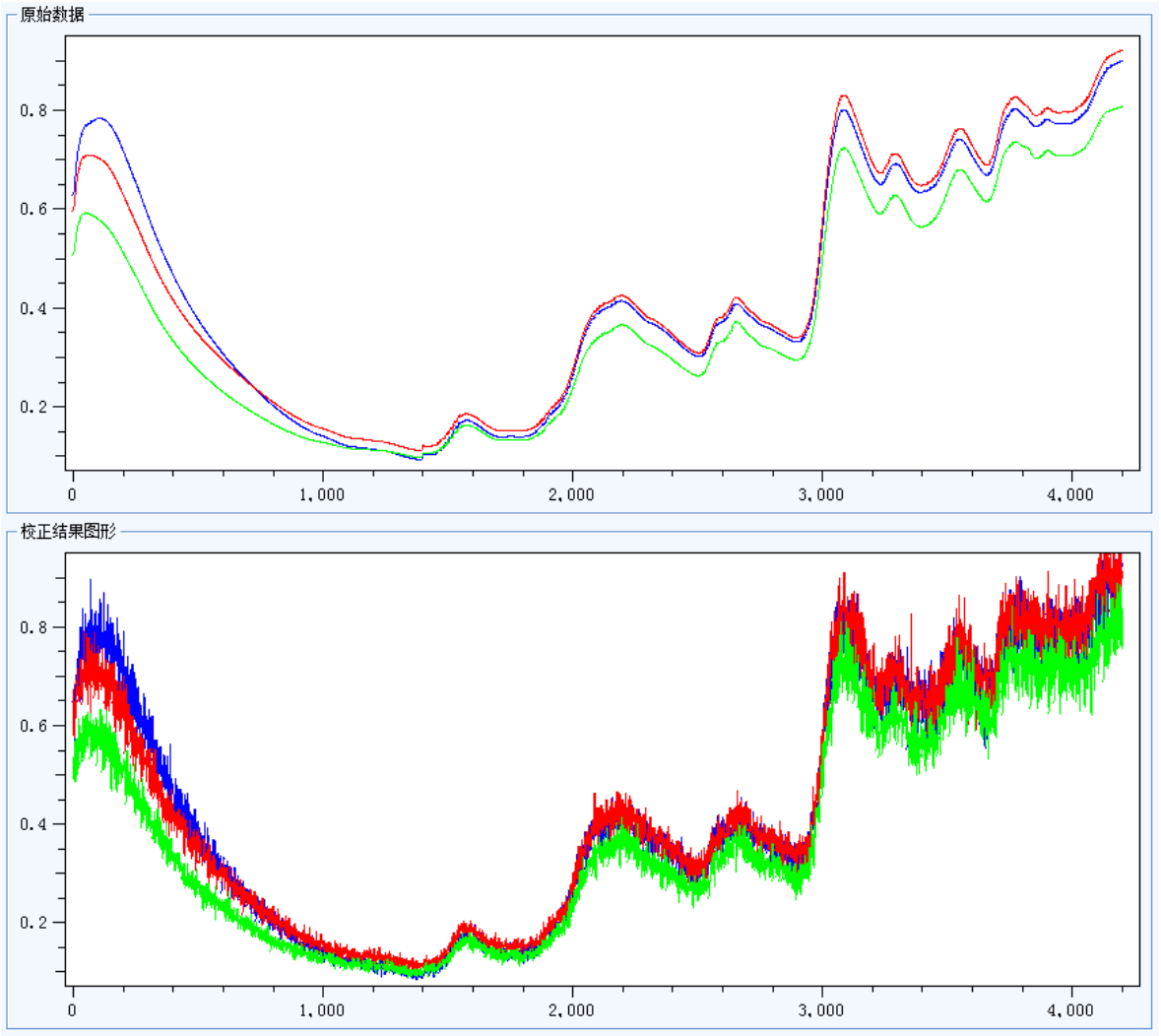
步骤 1: 点击**预处理标签** -> **加入噪声**，弹出如下对话框:



接下来的操作步骤参照预处理之通用步骤。参数说明见下表:

参数	范围	说明
噪声水平(最大信号百分比)	[0% 100%]	具体请参考数据处理方法。
标准偏差	[0 ∞]	具体请参考数据处理方法。

如下二图为数据加入噪声前后的对比图。



图中所加入噪声水平(最大信号百分比)和标准偏差分别设定为 5%和 0.001。

10.6. 样本归一化

归一化样本响应，以使其规范在相同或相近的标度范围内。归一化方法用于样本量测中检测器信号与样本质量呈某一函数关系等时的情形。

如下表则概括性描述本软件所涉及的归一化方法。

序号	方法名称	函数表达	说明
1	面积归一化	$\widehat{\mathbf{X}_i} = \mathbf{X}_i / \sum_j x_{i,j}$	计算量测信号曲线下面积，并用于样本归一化处理(如色谱分析)。
2	最大值归一化	$\widehat{\mathbf{X}_i} = \mathbf{X}_i / \mathbf{max}(\boldsymbol{X}_i)$	与平均值归一化方法对应，以最大值去除原始向量。在向量所有数据值符号不一致时应小心使用。
3	单位向量归一化	$\widehat{\mathbf{X}_i} = \mathbf{X}_i / \sqrt{\sum_j x_{i,j}^2}$	样本归一化处理后，其模为 1。
4	范围归一化	$\widehat{\mathbf{X}_i} = \mathbf{X}_i / (\mathbf{max}(\mathbf{X}_i) - \mathbf{min}(\mathbf{X}_i))$	与平均值归一化方法对应，以数据跨度范围(最大值减最小值)去除原始向量，使归一化后的数据跨度等于 1。
5	平均值归一化	$\widehat{\mathbf{X}_i} = \mathbf{X}_i / \overline{\mathbf{X}_i}$	最经典的方法之一，以向量均值去除原始向量得到的新值，可去除隐含未知因素对结果的影响，即使得转换后的值与分析中无法考虑到的因素成一比例关系变化，例如在色谱分析中，可校正进样量对色谱响应的影响。
6	峰值归一化	$\widehat{\mathbf{X}_i} = \mathbf{X}_i / x_{i,k}$	以某一个固定的数据点去除原始数据向量，对训练集和未知样本预测集数据，均使用同一数

			据点。本法可校正光程变化对数据结果的影响。但使用本方法时，需要特别注意其对基线偏移，边坡效应和波长漂移等非常敏感。

对被选数据进行样本方向的标准化操作。

操作步骤:

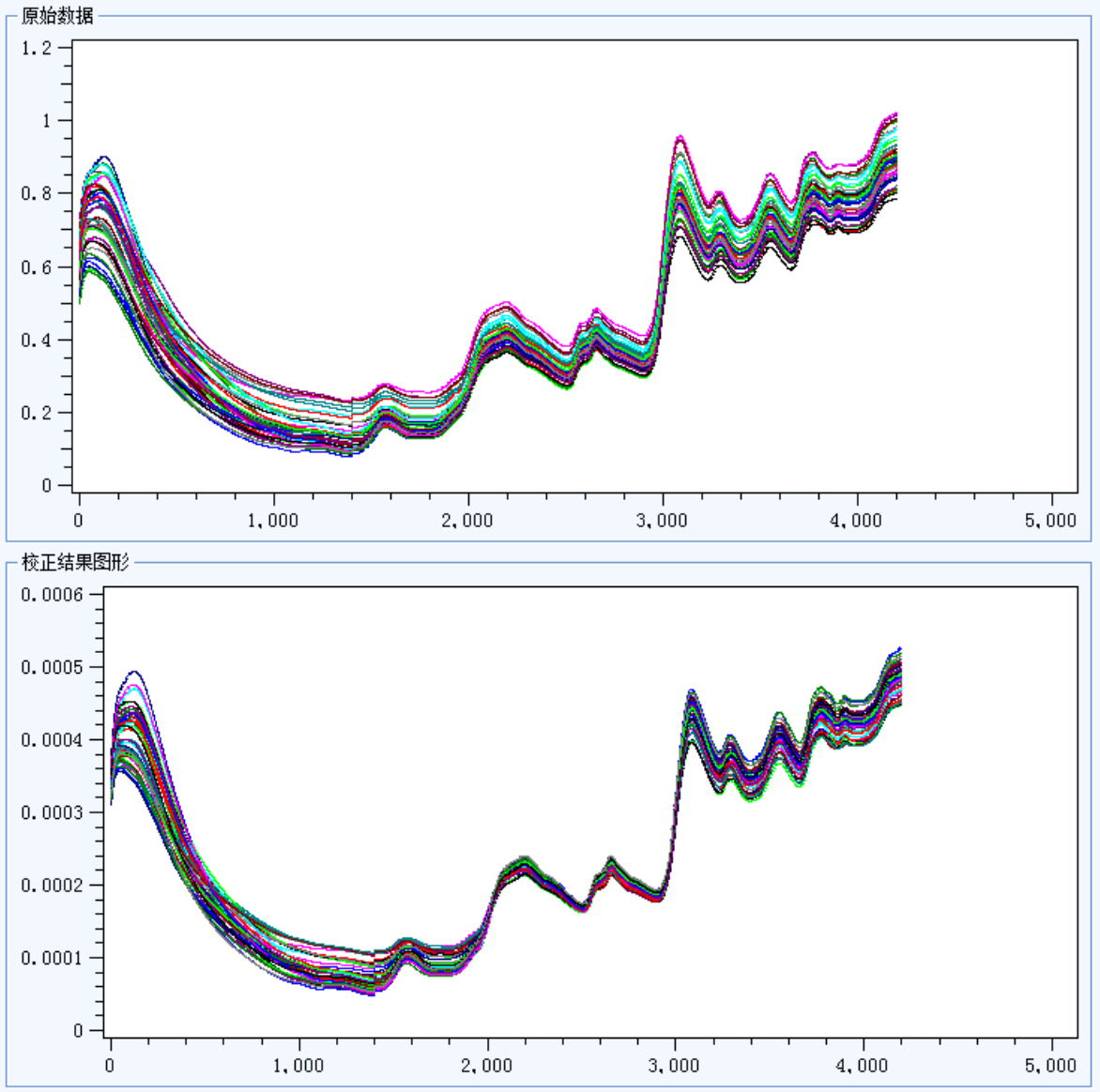
步骤 1: 点击**预处理** -> **样本归一化**，弹出如下对话框:

接下来的操作步骤参照预处理之通用步骤。

参数说明: 本部分仅峰值归一化方法涉及参数设置，该参数表示所选用于数据归一化

的峰位置。

以面积归一化方法为例，针对示例数据，可得到如下图所示的结果。



10.7. 变量标度化

变量标度化是对数据从变量方向的转换处理，包括二个方面，其一是中心化，其二是标度化。

数据中心化通常是多变量数据建模的第一步，是指数据中的各个变量减去其均值所

得的结果，以研究数据在其均值附近的变化，而不是数据的绝对值。根据实际问题的不同，有时亦使用其他的数值，而不一定是均值。

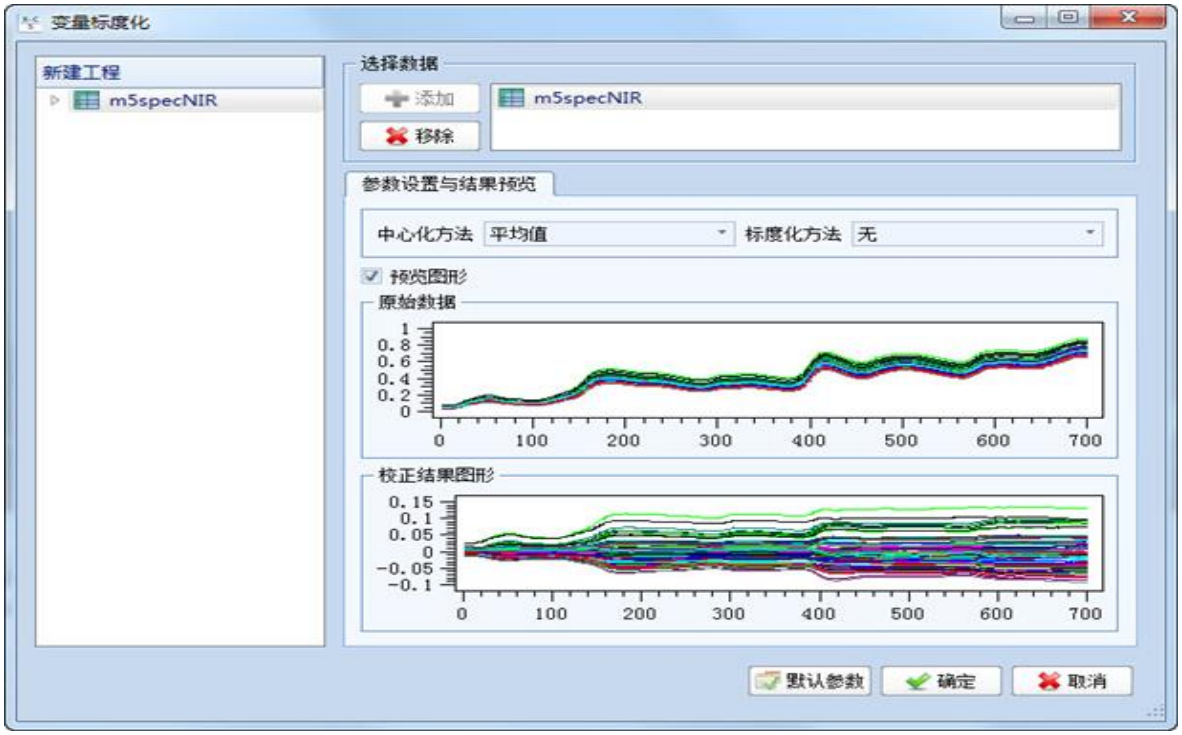
数据标度化则是指数据除以其估计范围，比如标准偏差。当不同变量的相对数值范围相差很大时，标度化则尤为重要，其原因在于具有更大方差的变量，其在回归分析时影响亦越大。

如下表则详述本软件所涉及的中心化和标度化方法，本节所说的标度化，则是而方法的叠加，即$\widehat{X_i} = (X_i - \mathbf{a})/b$，其中 a 为中心化方法所得的值，而 b 则是标度化方法所得的值。

方法	方法名称	说明
中心化方法	无	不做处理。
	平均值	使用广泛，为首选方法。着力观察值间的差异，而非其绝对值。
	中位数	平均值方法文件替代方法。当部分变量非对称分布时，本法可使得变量更趋近于数据中心(重心)。
	最小值	保证数据非负性，如色谱分析。
标度化方法	无	不做处理。
	标准偏差	通常与均值中心化方法搭配使用；数据变换后各变量方差均为 1。
	标准偏差开方	通常与中位数中心化方法搭配。
	四分位距(IQR)	通常与中位数中心化方法搭配。

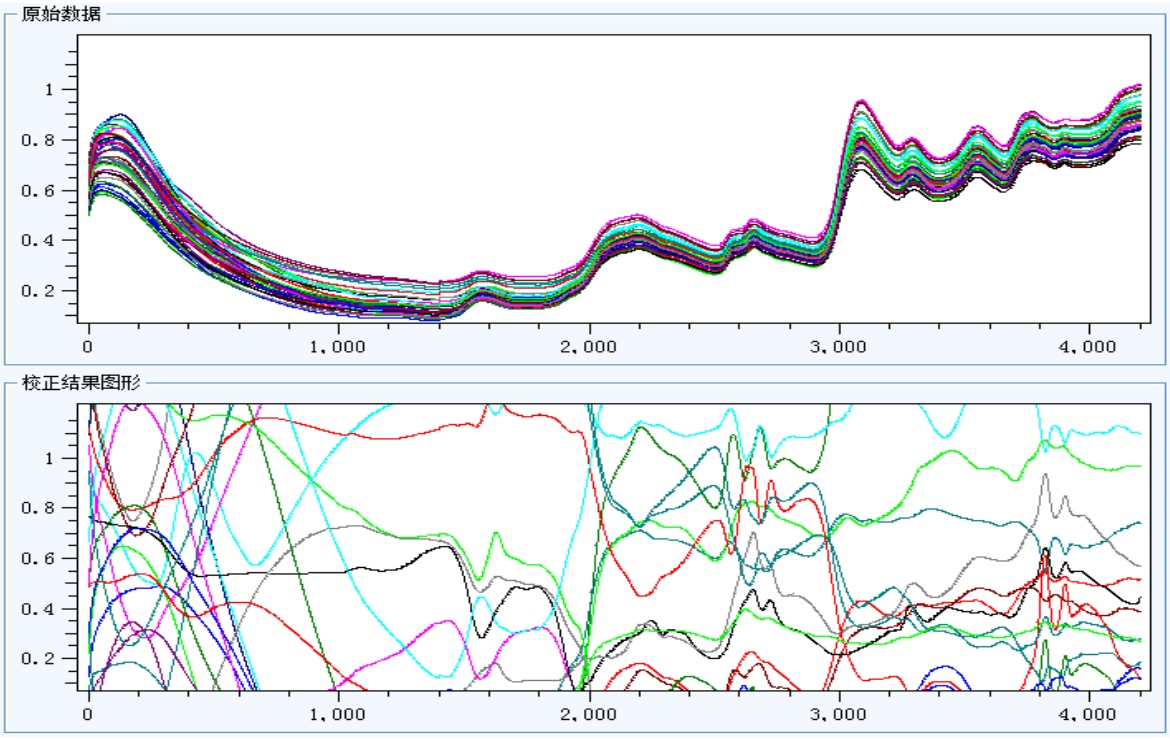
操作步骤:

步骤 1: 点击**预处理标签** -> **变量标度化**，弹出如下对话框:



接下来的操作步骤参照预处理之通用步骤。

如下二图则分别表示原始数据以及经过变量标度化后得到的结果，中心化和标度化方法分别为平均值和标准偏差。



10.8. SNV 变换(标准正态变量变换)

标准正态变量变换，从样本方向对数据进行中心化和标度化处理。广泛用于光谱数据处理，可去除光谱散射效应。该方法有时与去趋势化方法联合使用，以减少多重共线性与基线漂移等。该方法同样不能用于非数值数据的分析。其计算公式如下，即原始光谱先减去其平均值，再除以其标准偏差：

$$\widehat{\mathbf{x}_k} = \frac{\mathbf{x}_k - \mathrm{Mean}(\mathbf{X})}{\mathrm{SDev}(\mathbf{X})}$$

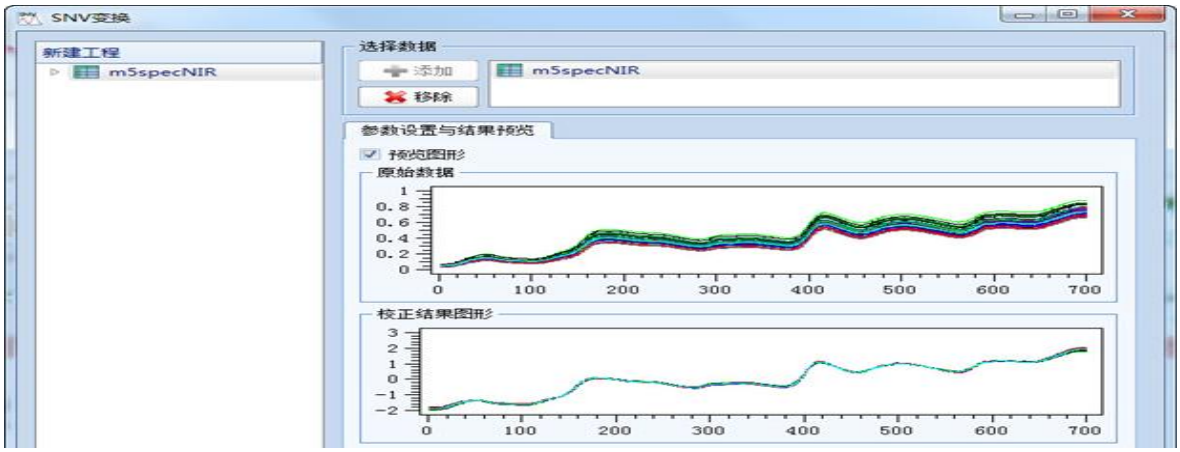
与多元散射校正(MSC)方法类似，SNV 法所得到的实际结果同样可去除光谱数据的散射，粒子尺寸效应，以及光程变化的影响。

经过 SNV 变换后，数据样本方向每行元素的均值为 0，方差和标准偏差均为 1。变换后数据强度得到调整，从而达到散射校正的目的。

至此，用户应已发现，SNV 算法公式与上述变量标度化时中心化使用均值，标度化使用标准偏差完全相同！然而，前者是从样本方向，针对光谱观察值的处理，而后者则是从变量方向，对同一变量在不同样本中变化的处理。且后者需要同时对数据矩阵 **X** 和响应因变量 **y** 进行处理(标度化)。

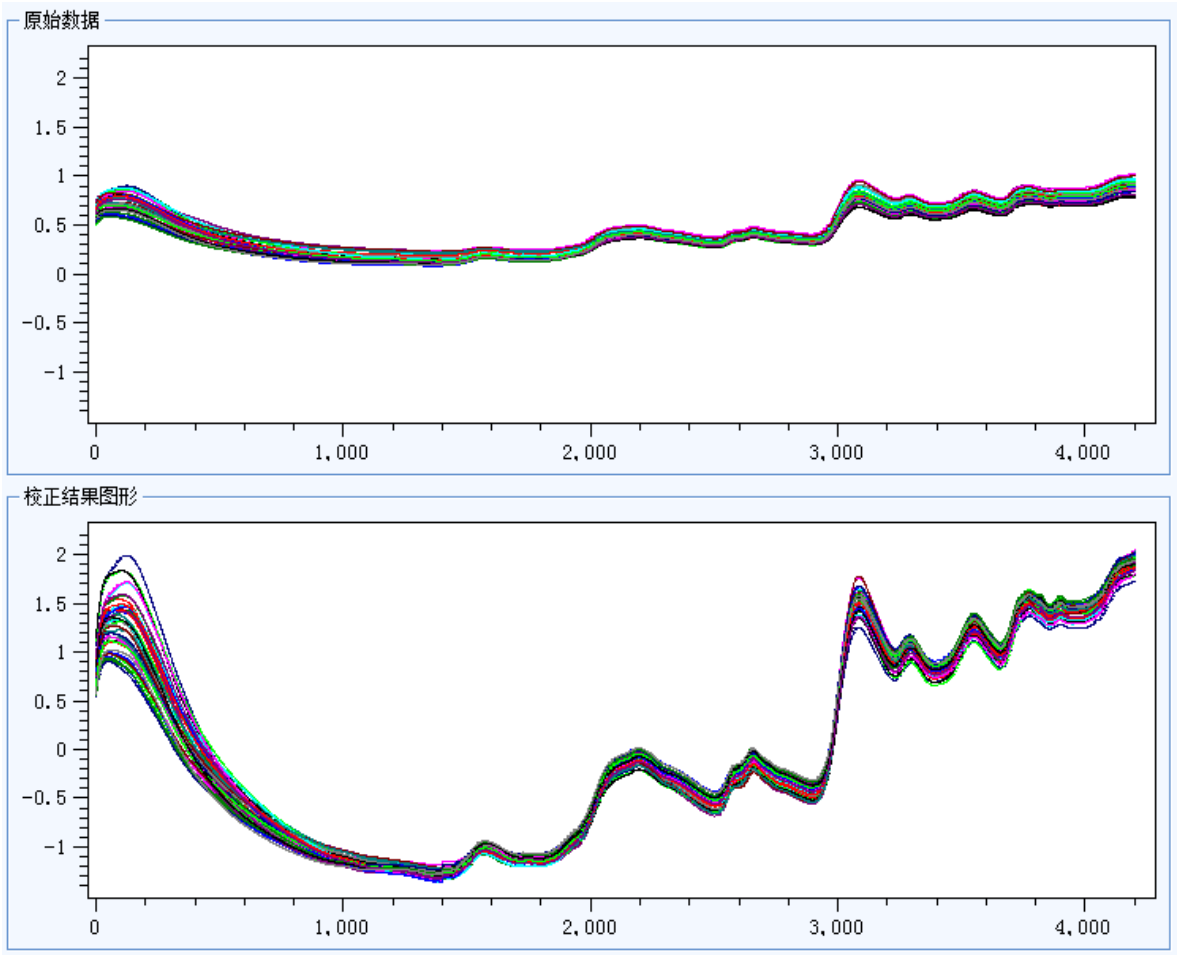
操作步骤:

步骤 1: 点击**预处理标签** -> **SNV 变换**，弹出如下对话框:



接下来的操作步骤参照预处理之通用步骤。

如下二图为示例数据 SNV 变换前后的结果对比。



10.9. Quantile 标准化

分位数标准化，最初用于去除基因芯片数据多次平行实验间的数据差异，目前亦较多地用于基因组学和代谢组学的数据分析中。经过 Quantile 标准化后，数据矩阵中的所有样本将具有相同经验分布，显然这样的假设难以存在于光谱数据中。若该假设存在，则 Quantile 标准化方法可较好地去除样本间的背景差异，如在基因组学或代谢组学的研究中，不同实验条件下得到主要变量间(基因或代谢物小分子)没有统计差异性。

通常，Quantile 标准化方法包括平均值，中位数和参考向量三种计算方式，本软件提供前二者方法，每个样本数据先按照从小到大的顺序排列，再选择各均值或中位数作为参

考分布，然后每个样本的最小值以参考分布的最小值替换，最后每个样本经过数据变换后，其所包含的数据与参考分布完全相同。

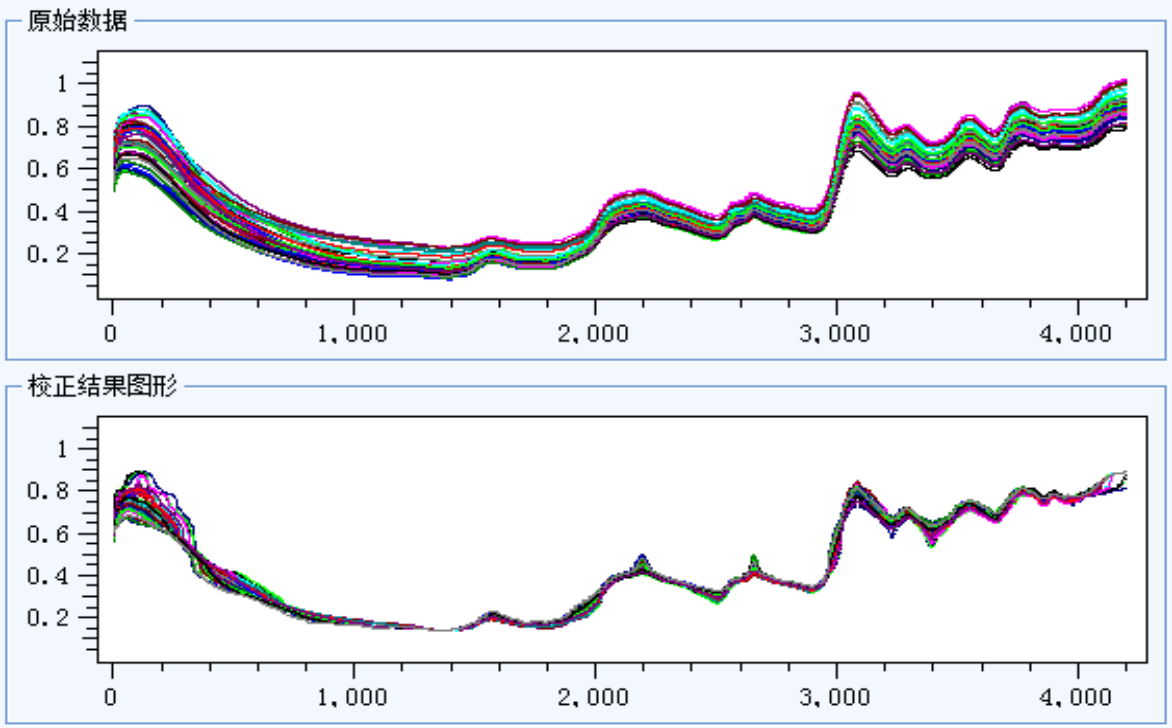
操作步骤:

步骤 1: 点击**预处理** -> **Quantile 标准化**，弹出如下对话框:

接下来的操作步骤参照预处理之通用步骤。

参数说明: 如上所述。

如下二图为示例数据 Quantile 变换前后的结果对比，方法选择为平均值。



10.10. 数据运算

从数据样本或变量方向对数据进行新计算，以产生新的样本或变量，或替换原有的样本或变量。

具体所能提供的计算方法请参见如下图形等。

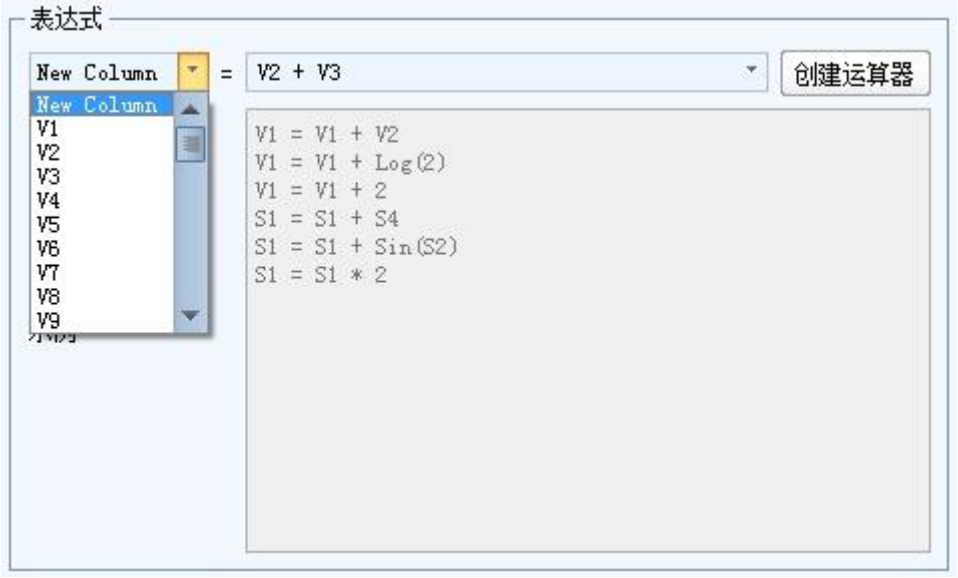
操作步骤:

步骤 1: 点击**预处理** -> **数据运算**，弹出如下对话框:

步骤 2: 选择数据，参照预处理的通用操作之选择数据。

步骤 3: 选择是对样本或变量进行计算。

步骤 4: 从表达式左侧选择新产生样本或变量名，或覆盖原有样本或变量，如下图所示。



当用户选择对样本进行计算时，若左侧选择新行，则表示产生新的样本，否则为使用新产生的样本覆盖已经存在样本。

当用户选择对变量进行计算时，若左侧选择新列，则表示产生新的变量，否则为使用新产生的变量覆盖已经存在变量，

在步骤 3 中，用户亦可通过直接输入表达式的方式，完成样本/变量计算，或者通过点击 创建运算器 按钮，弹出如下对话框以构建计算表达式，用户可通过选择任意样本或变量，以及丰富的运算方式，产生计算结果。

- 点击按钮 撤销，可撤销用户之前的操作，每点击一次，便撤销一步。
- 点击按钮 清空，则清空表达式显示框中的所有内容。
- 该运算器支持的计算方式及其含义如下:

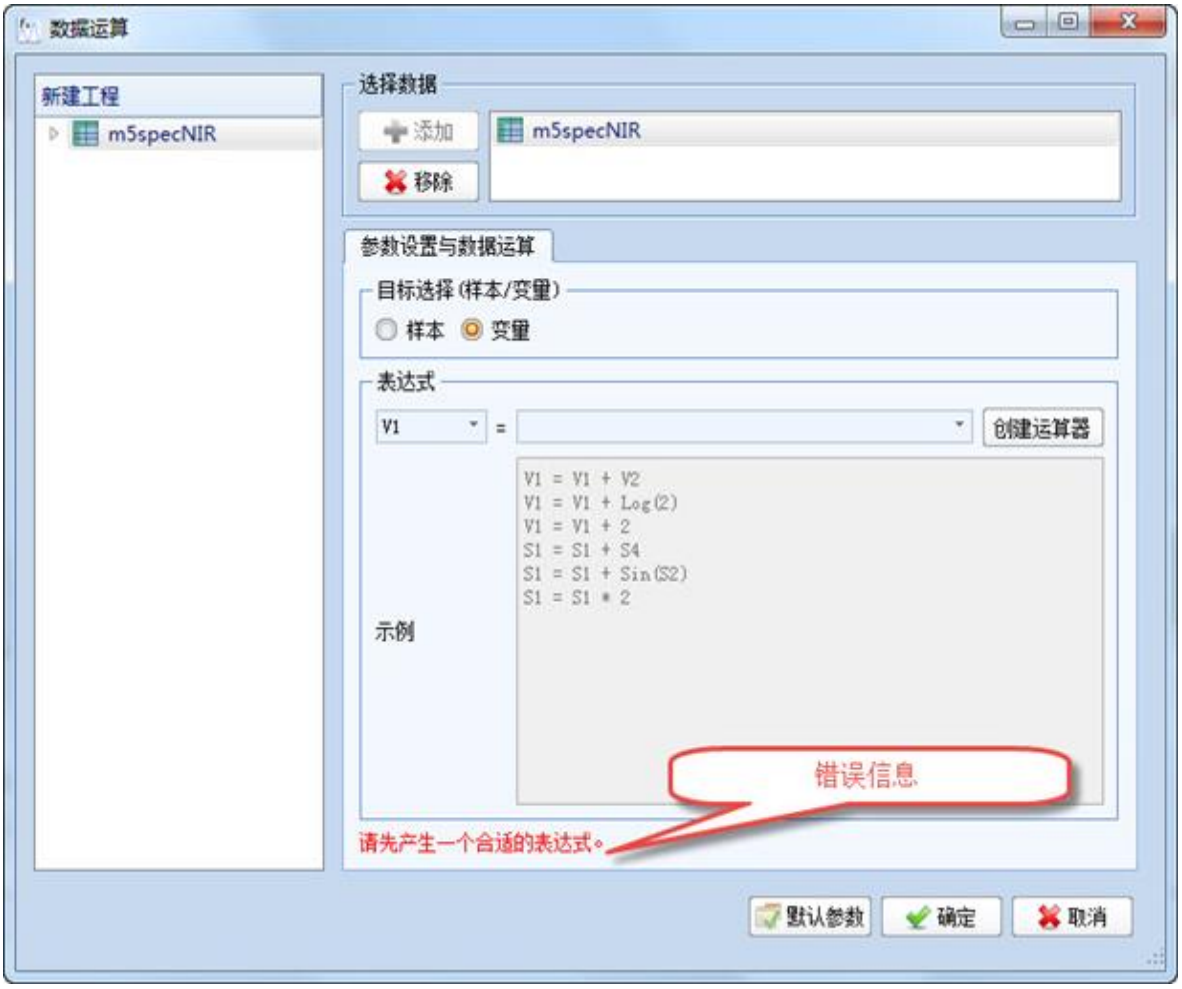
运算类别	运算方法	说明
常规运算	+	加法运算。
	-	减法运算。
	*	乘法运算。
	/	除法运算。

特殊运算	LOG10	以 10 为底的对数运算。
	LOG2	以 2 为底的对数运算。
	LN	自然对数运算。
	ABS	绝对值运算。
	EXP	指数运算。
	SQRT	开方运算。
	POW	乘方运算。
	SQUARE	平方运算。
	FLOOR	下取整运算。
	CEIL	上取整运算。
	ROUND	四舍五入运算。
	SIGN	符号函数运算。
	SIN	正弦运算。
	COS	余弦运算。
	TAN	正切运算。
	SINH	反正弦运算。
	COSH	反余弦运算。

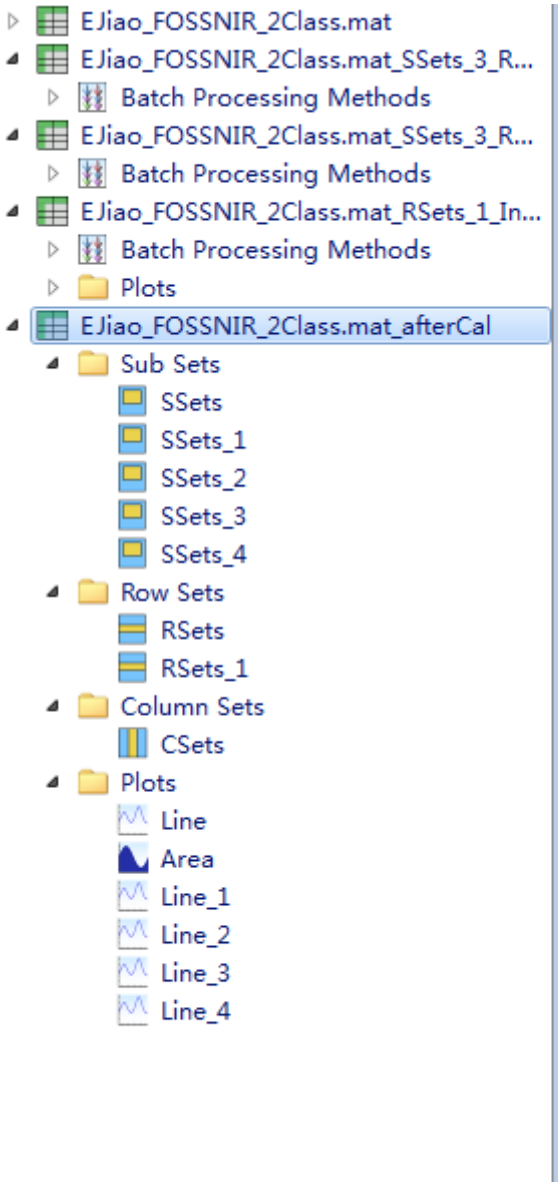
	TANH	反正切运算。

除此之外，亦包括括号和数字运算等，功非常能强大。

步骤 6: 点击数据运算对话框的确定按钮后便开始计算。计算成功后，将加入一个结果矩阵节点到工程导航栏中；若计算失败，则显示错误提示信息，如下图所示。



若选择产生新的样本或变量，则完成后，将自动排列在样本或变量的最末尾。以一个含有 43 个样本的数据为例，经过从样本方向的数据运算，增加了一个模拟样本，总样本数达到 44，如下图所示。

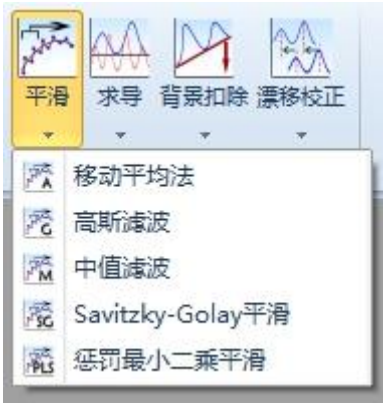


#	x	1	2
		1	
#_22	23	0.57134	0.5
#_23	24	0.64068	0.6
#_24	25	0.6015	0.6
#_25	26	0.53284	0.5
#_26	27	0.68762	0.6
#_27	28	0.69709	0.7
#_28	29	0.6509	0.6
#_29	30	0.66647	0.6
#_30	31	0.60559	0.6
#_31	32	0.61774	0.6
#_32	33	0.70375	0.7
#_33	34	0.65021	0.6
#_34	35	0.70645	0.7
#_35	36	0.67737	0.6
#_36	37	0.52041	0.5
#_37	38	0.63874	0.6
#_38	39	0.64952	0.6
#_39	40	0.59799	0.6
#_40	41	0.49752	0.5
#_41	42	0.56134	0.5
#_42	43	0.6205	0.6
#_43	44	0.54313	0.5

10.11. 平滑

数据平滑可帮助在不减少变量数的情况下，消除随机误差的影响，并提高数据信噪比，以减少噪声对数据处理及其结果的影响。从另一个方面来说，平滑可消除小方差信号，保留大方差信号，但不合适的平滑处理可能将微弱信号当成噪声处理。因此，用户在使用这些方法时，需要特别小心参数的选择。

本软件提供如下图所示的平滑方法。



10.11.1 移动平均法平滑

本法是非常经典的平滑方法之一，使用一定窗口尺寸内各数据点的算术平均值替代目标数据点(包括目标点本身)。在选取数据窗口时，分为左、右窗口二部分，分别从目标数据点的左、右二侧选择一定量的数据点。

操作步骤:

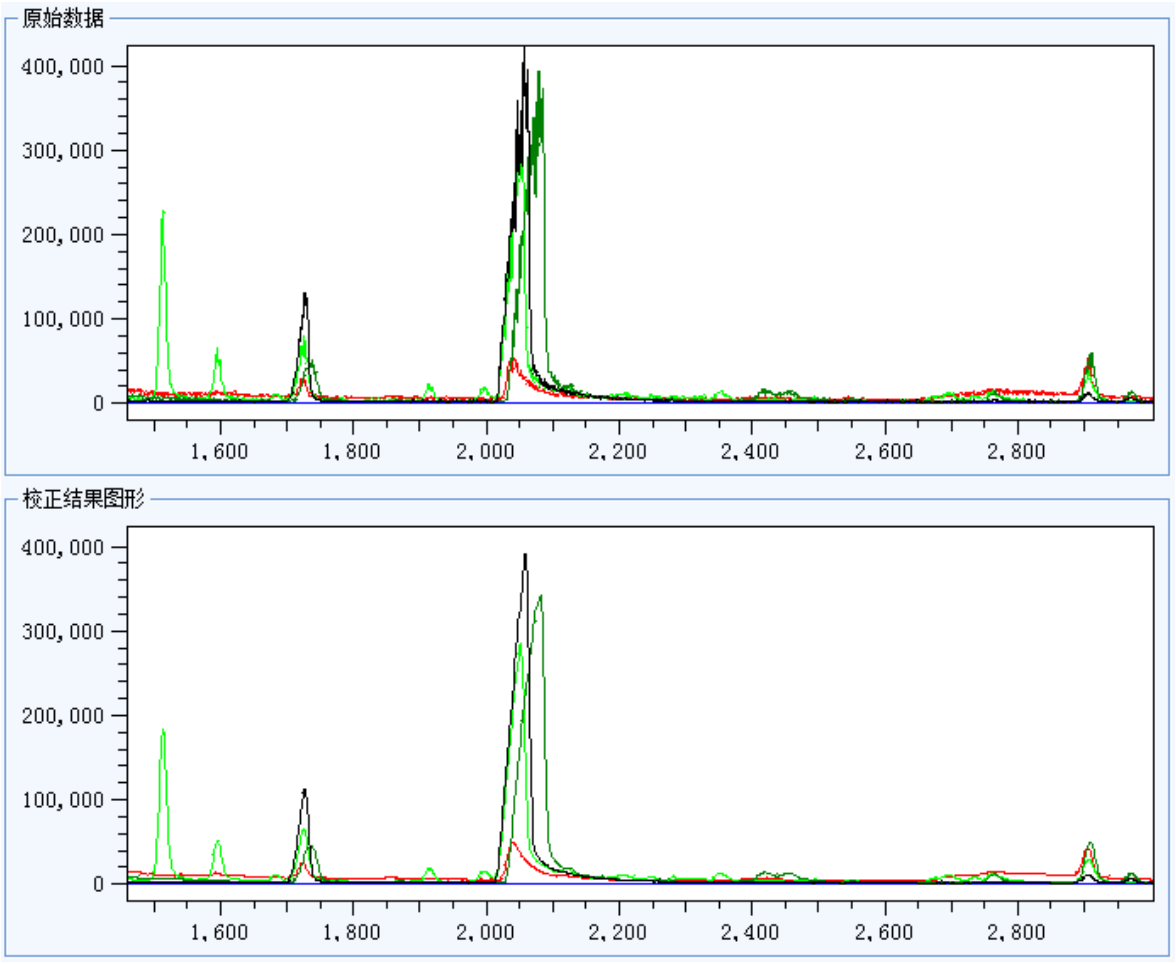
步骤 1: 点击**预处理标签** -> **平滑** -> **移动平均法**，弹出如下对话框:

接下来的操作步骤参照预处理之通用步骤。

参数说明见下表:

参数	范围	说明
左窗口尺寸	[0 num]，其中 num 表示所选数据的长度。	如前所述，所选目标数据点左侧窗口的尺寸大小。
右窗口尺寸	[0 num]，其中 num 表示所选数据的长度。	如前所述，所选目标数据点右侧窗口的尺寸大小。

下图是含有较大噪声的色谱示例数据与平滑后的结果比较(左、右窗口大小均为 5)。

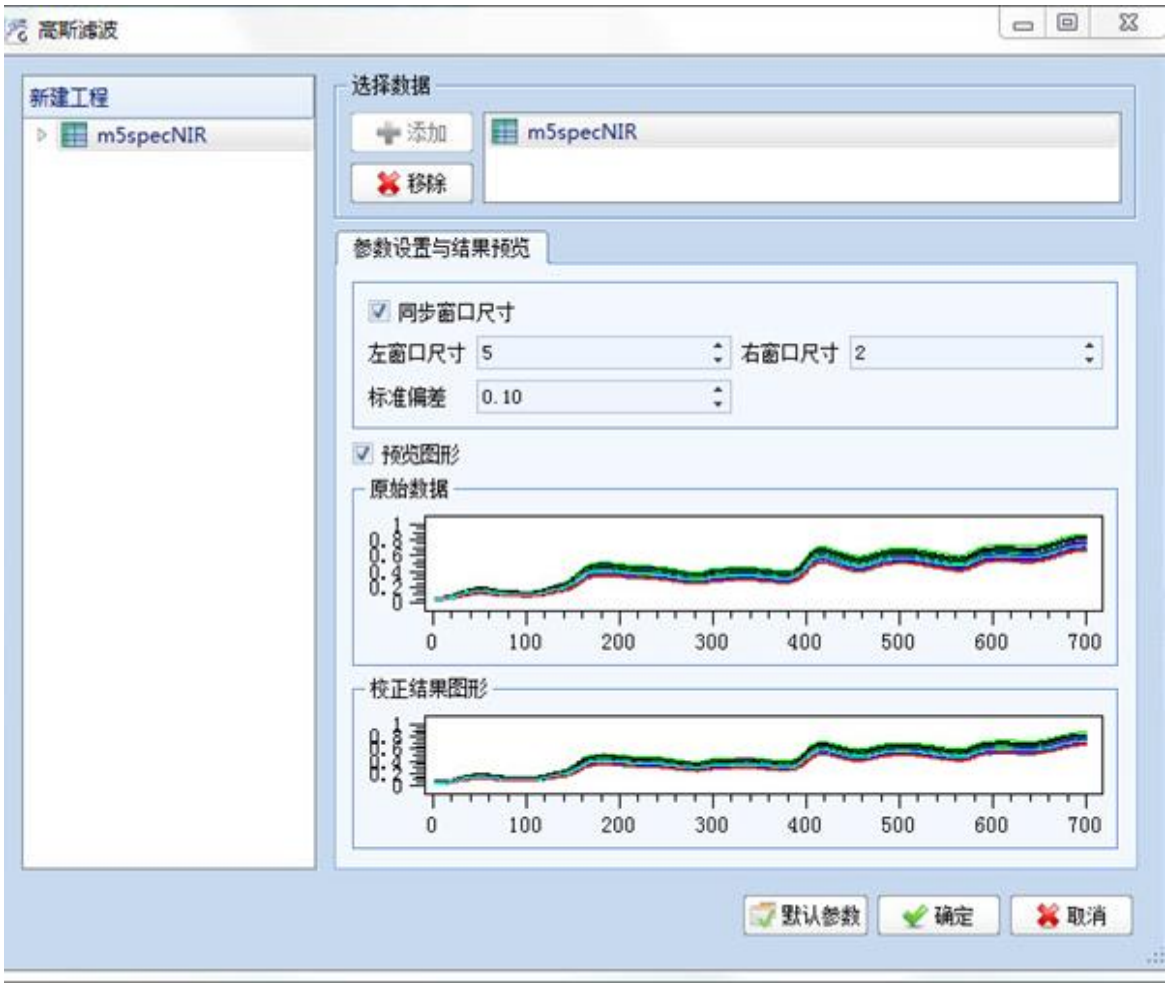


10.11.2. 高斯滤波平滑

适合于消除服从正态分布的噪声。本法原理上与移动平均法类似，差异在于被替换值不是基于某一窗口内各点的平均值，而是其基于用户自定义参数高斯函数的加权平均值计算得到。

操作步骤:

步骤 1: 点击**预处理标签** -> **平滑** -> **高斯滤波平滑**，弹出如下对话框:

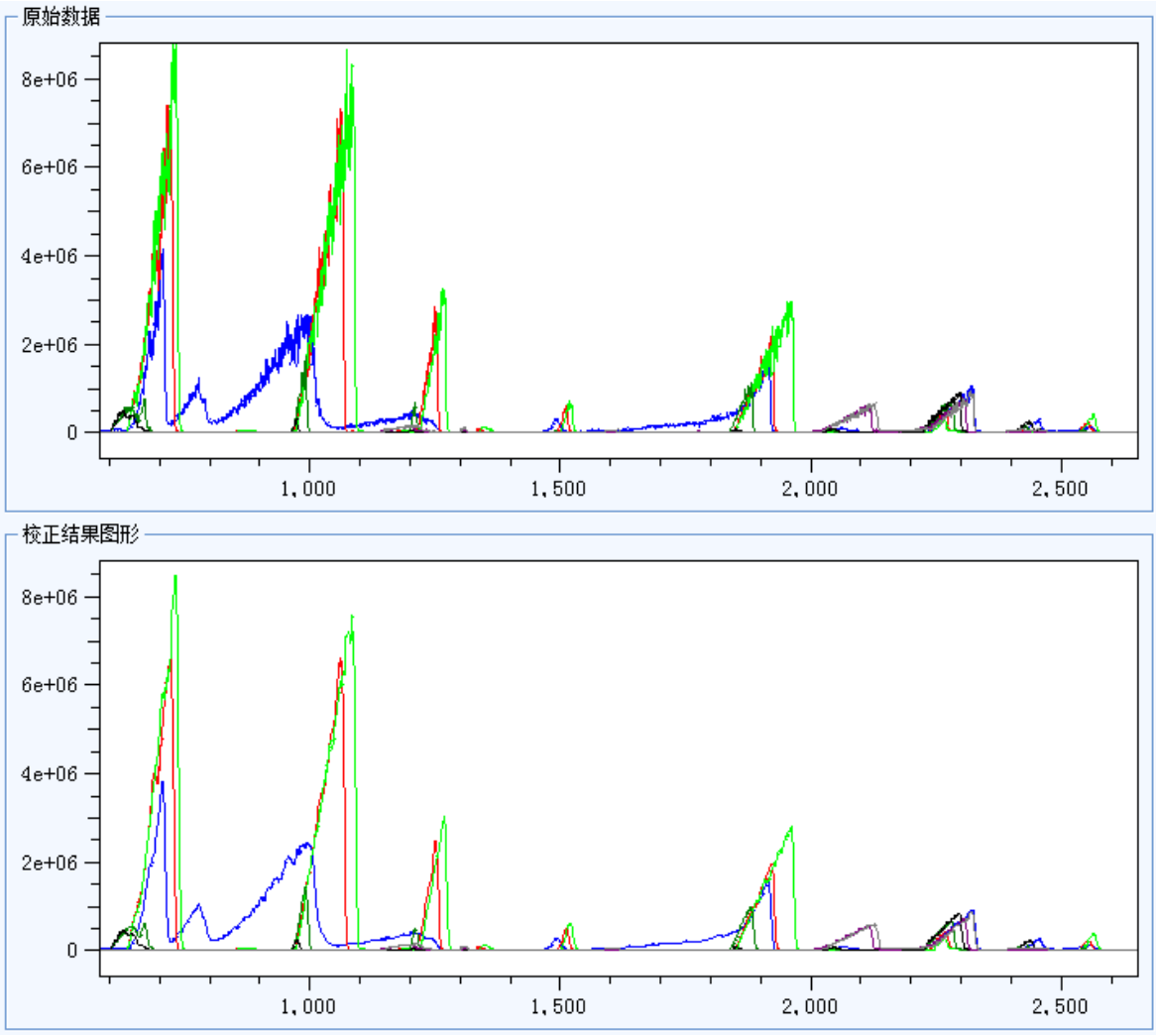


接下来的操作步骤参照预处理之通用步骤。

参数说明见下表:

参数	范围	说明
左窗口尺寸	[0 num]，其中 num 表示所选数据的长度。	如前所述，所选目标数据点左侧窗口的尺寸大小。
右窗口尺寸	[0 num]，其中 num 表示所选数据的长度。	如前所述，所选目标数据点右侧窗口的尺寸大小。
标准偏差	[0.1 ∞]，其中∞表示无穷大。	加权高斯函数参数值。

如下二图即为示例数据的结果比较。

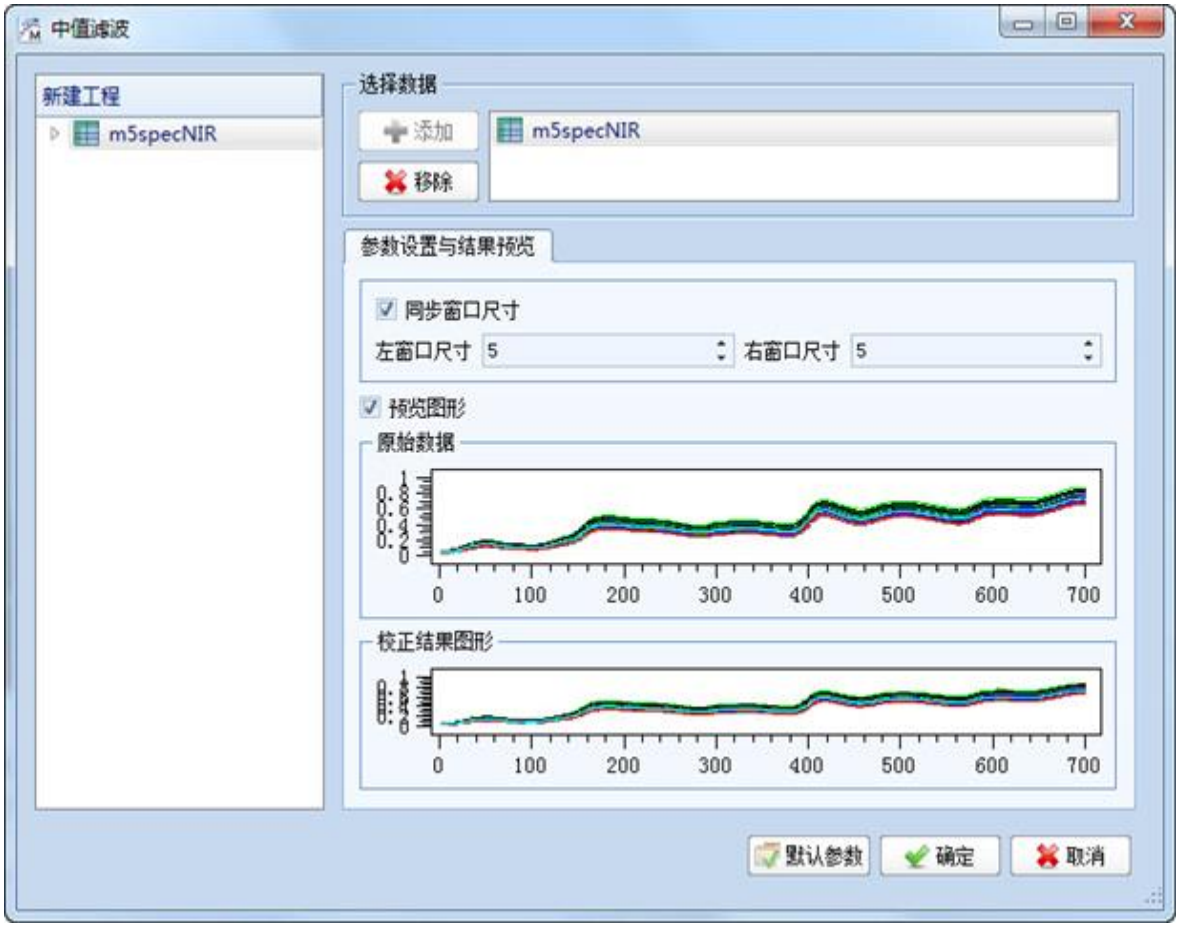


10.11.3. 中值滤波平滑

与移动窗口类似，本法以目标数据预设窗口内邻近点的中位数替换原始数据。预设窗口尺寸应为奇数。

操作步骤:

步骤 1: 点击**预处理标签** -> **平滑** -> **中值滤波**，弹出如下对话框:



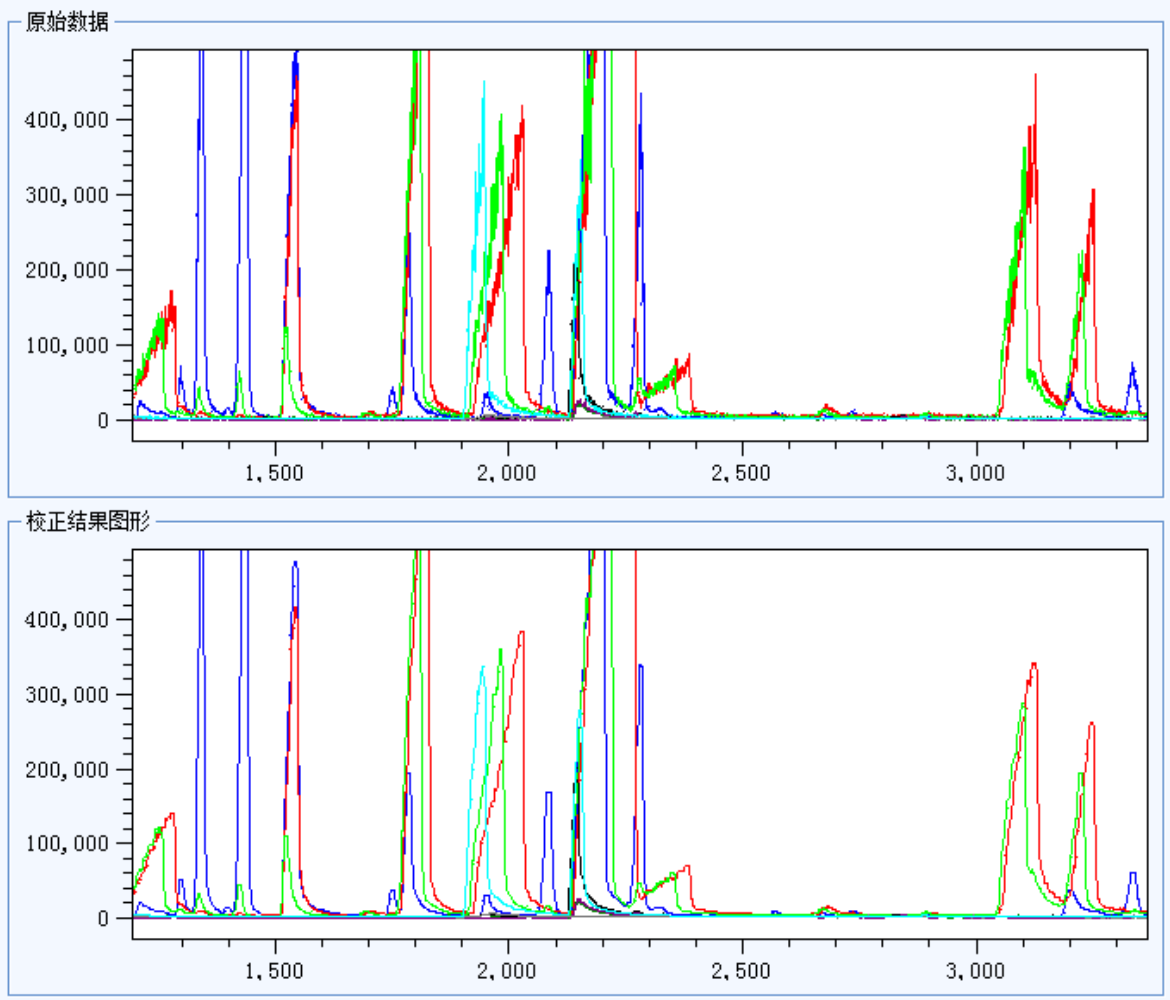
接下来的操作步骤参照预处理之通用步骤。

参数说明见下表:

参数	范围	说明
左窗口尺寸	[0 num]，其中 num 表示所选数据的长度。	如前所述，所选目标数据点

		左侧窗口的尺寸大小。
右窗口尺寸	[0 num]，其中 num 表示所选数据的长度。	如前所述，所选目标数据点右侧窗口的尺寸大小。

如下二图即为示例数据的结果比较。



10.11.4. Savitzky-Golay 平滑

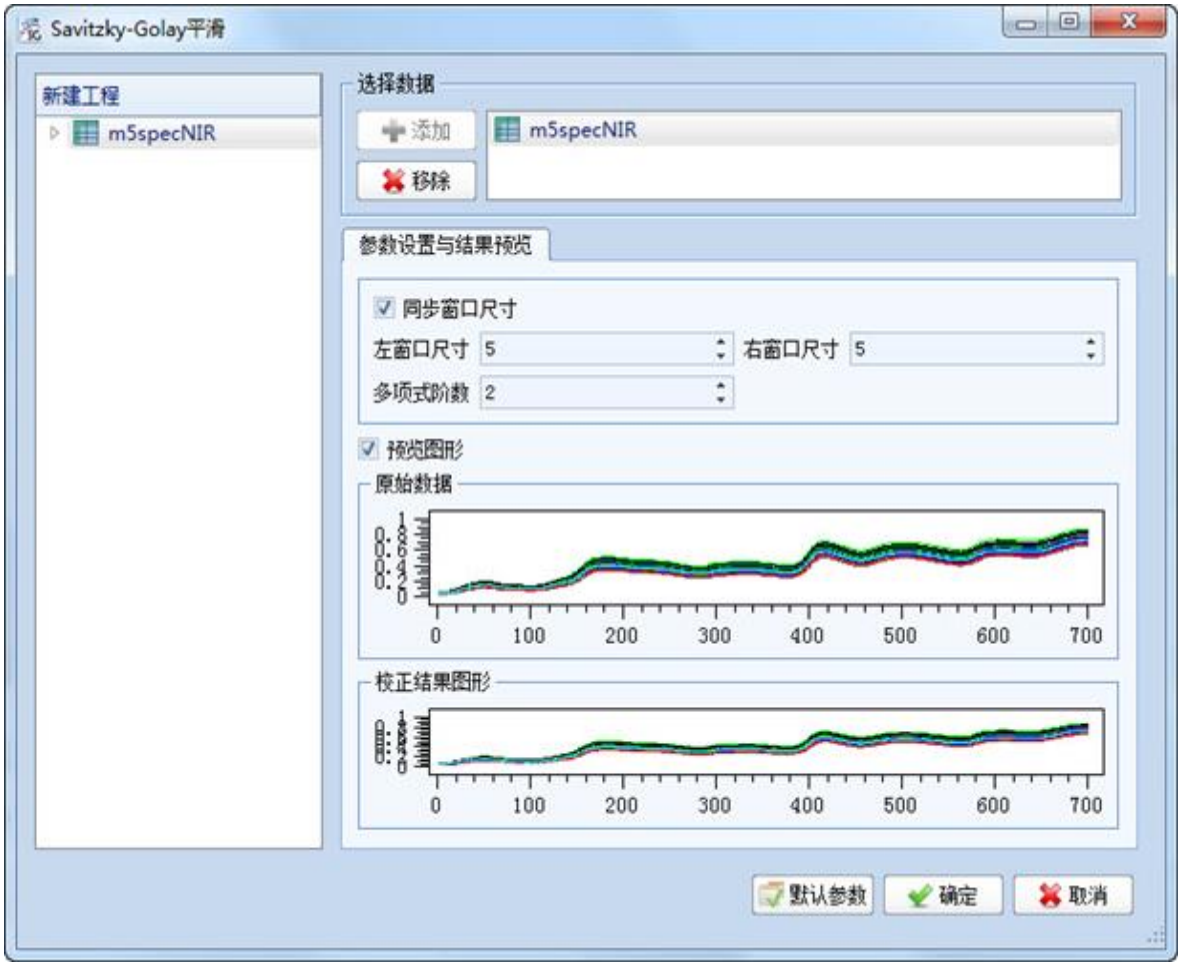
本法通过对移动窗口尺寸内的数据进行多项式最最小二乘拟合，实现对目标数据点的平滑。

特别需要注意的是窗口大小的选择，当窗口宽度选择太小时，平滑的效果不见得很理想，

而当窗口尺寸太大时，尽管平滑效果更好些，但往往会丢失较多信息，导致信号失真。

操作步骤:

步骤 1: 点击**预处理** -> **平滑** -> **Savitzky-Golay 平滑**，弹出如下对话框:



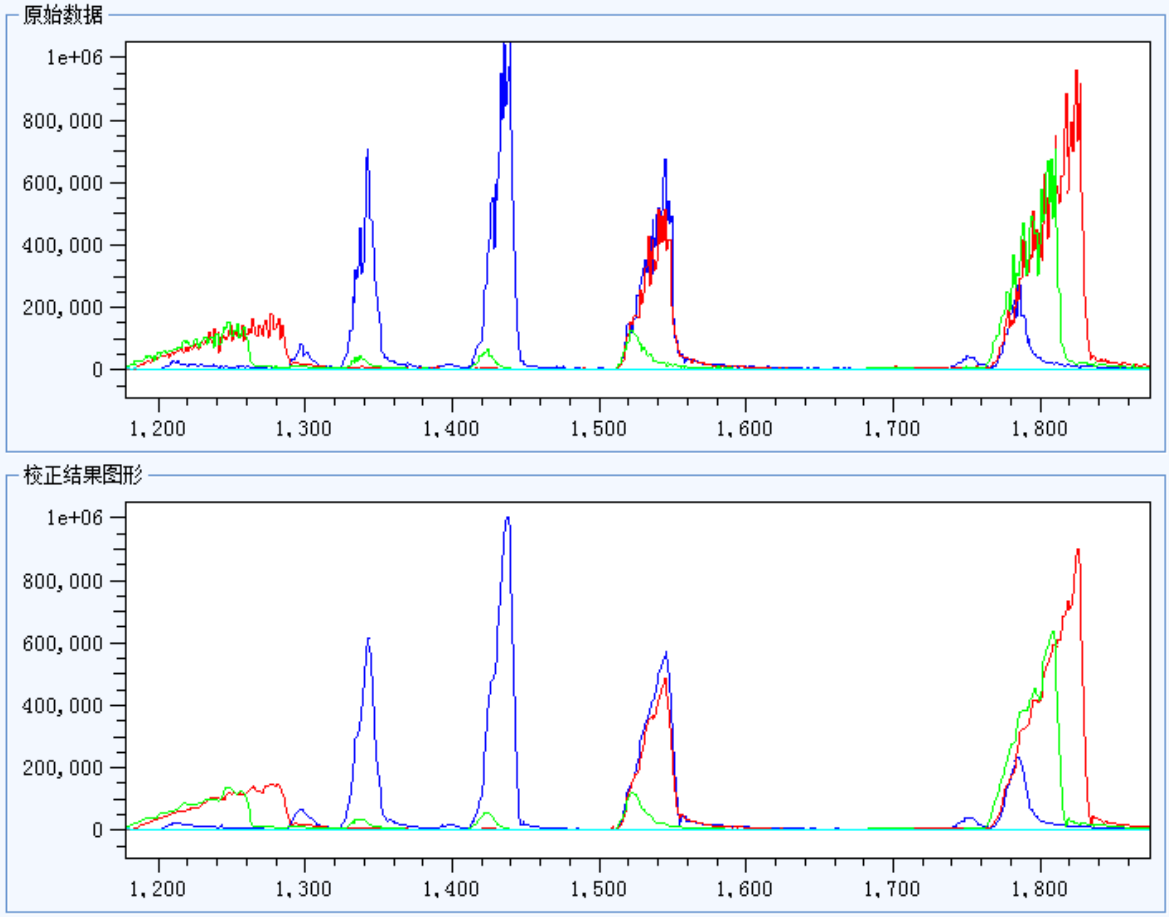
接下来的操作步骤参照预处理之通用步骤。

参数说明见下表:

参数	范围	说明
左窗口尺寸	[0 num]，其中 num 表示所选数据的长度。	如前所述，所选目标数据点左侧窗口的尺寸大小。

右窗口尺寸	[0 num]，其中 num 表示所选数据的长度。	如前所述，所选目标数据点右侧窗口的尺寸大小。
标准偏差	[1 左窗口尺寸 + 右窗口尺寸 - 1]。	加权高斯函数参数值。

如下二图即为示例数据的结果比较。

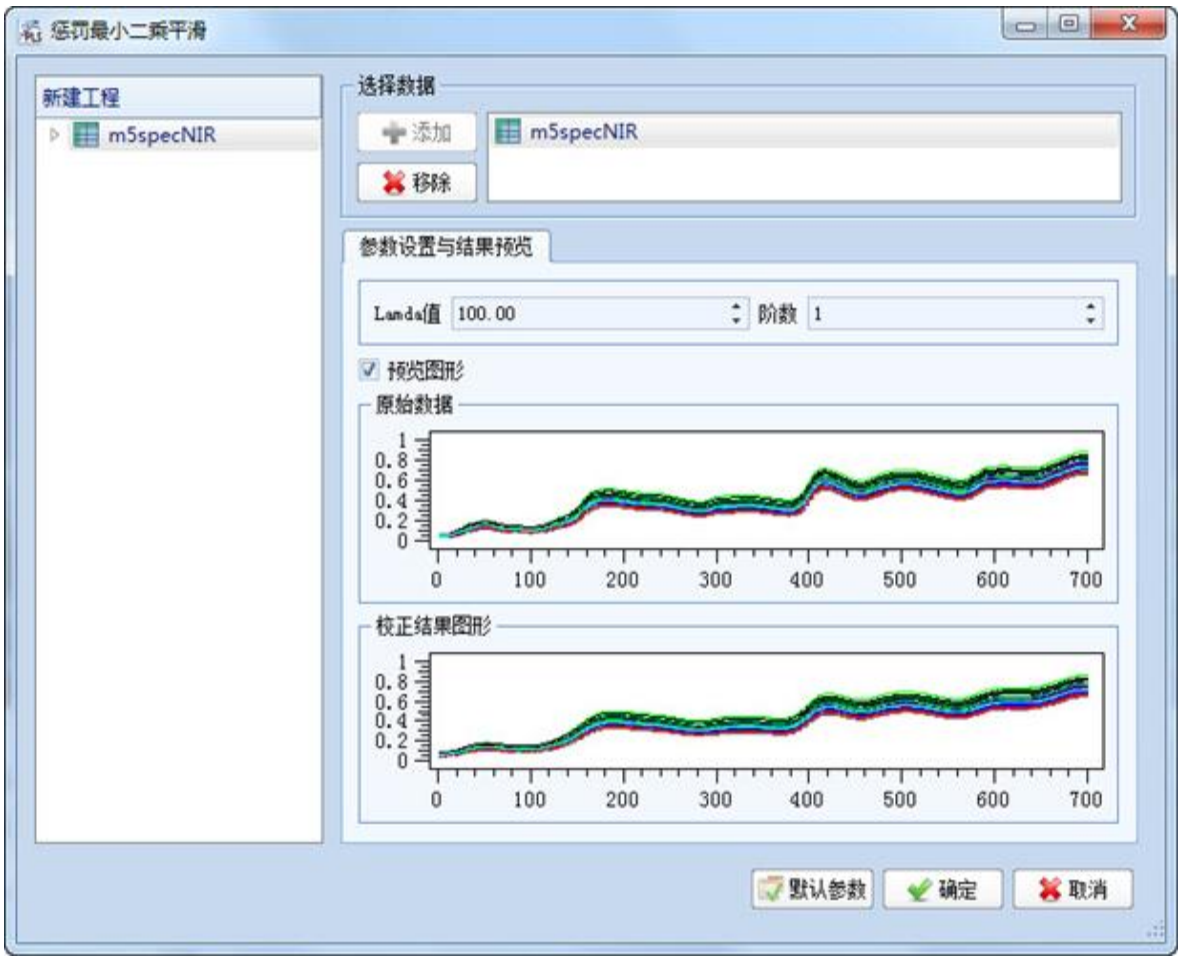


10.11.5. 惩罚最小二乘平滑

该方法滤除噪声的效果非常好。其核心在于平滑的优化目标函数同时包含原始光谱与除噪后光谱的最小二乘项，以及最小二乘的惩罚项(除噪后光谱的一阶导数)，前者用于控制拟合误差，而后者则用于限定除噪后光谱的平滑程度。

操作步骤:

步骤 1: 点击**预处理标签** -> **平滑** -> **惩罚最小二乘平滑**，弹出如下对话框:

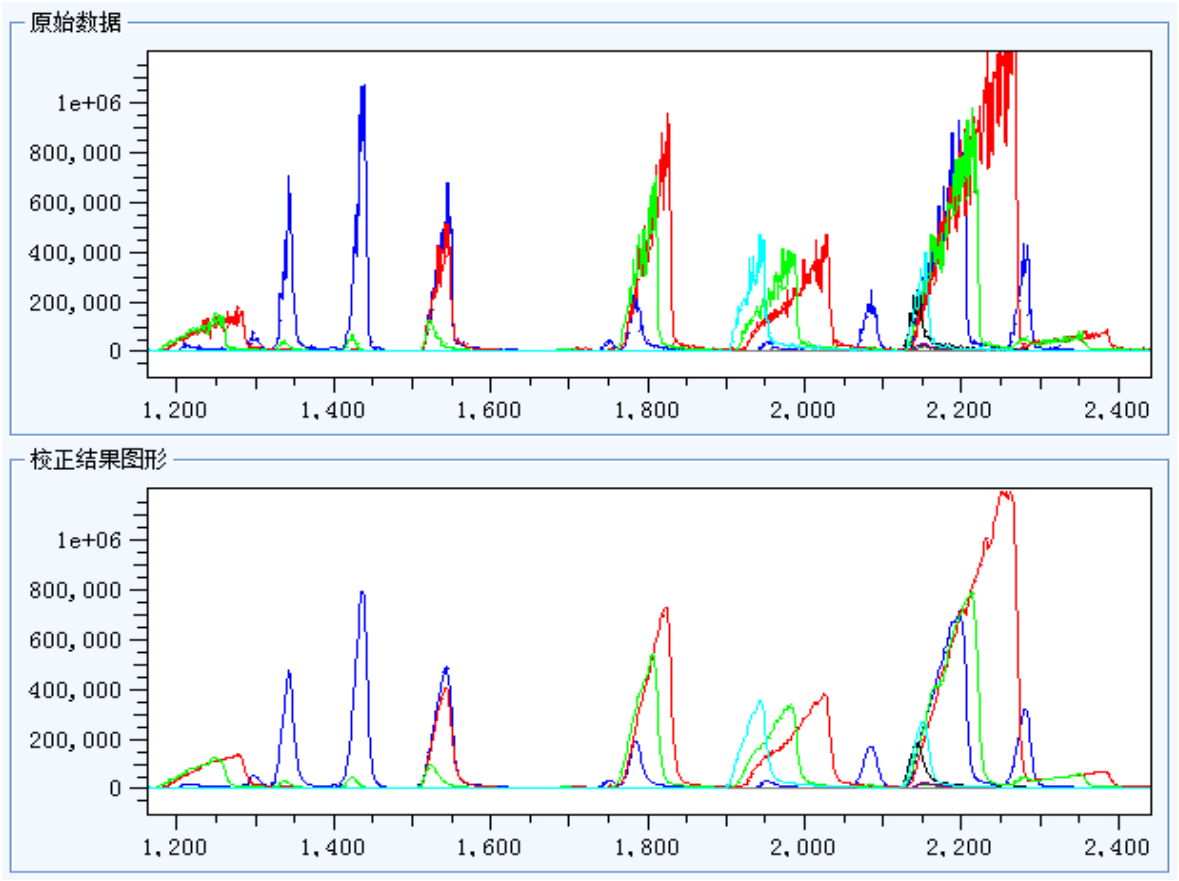


接下来的操作步骤参照预处理之通用步骤。

参数说明见下表:

参数	范围	说明
Lamda 值	[1 ∞]，其中∞表示无穷大。	惩罚系数。
阶数	[1 num-1]，其中 num 表示所选数据的长度。	多项式拟合阶数。

如下二图即为示例数据的结果比较。



从结果上述数据平滑结果可以清晰看到，本法对原始数据保持很好的保真度，平滑效果亦较佳。

10.12. 求导

求导作为光谱数据分析的重要方法，其意义和作用是不言而喻的，可达到去除非化学效益，并构建更稳健校正模型的目的，亦可部分消除样本组份间相互干扰导致的响应重叠。一阶或二阶导数光谱可更容易展示隐藏在原始光谱中的信息；前者以相邻两点间的差值除以采样点间隔获得，其物理意义代表该点曲线斜率，而后者则在前者的基础上，继续求导获得，其物理意义代表该点曲线斜率的变化。

ⓘ 尽管更高阶导数的应用相对较少，但有时亦可获得光谱数据中低阶导数不具备的特性，但他们显然会极大地降低转换信号的强度。

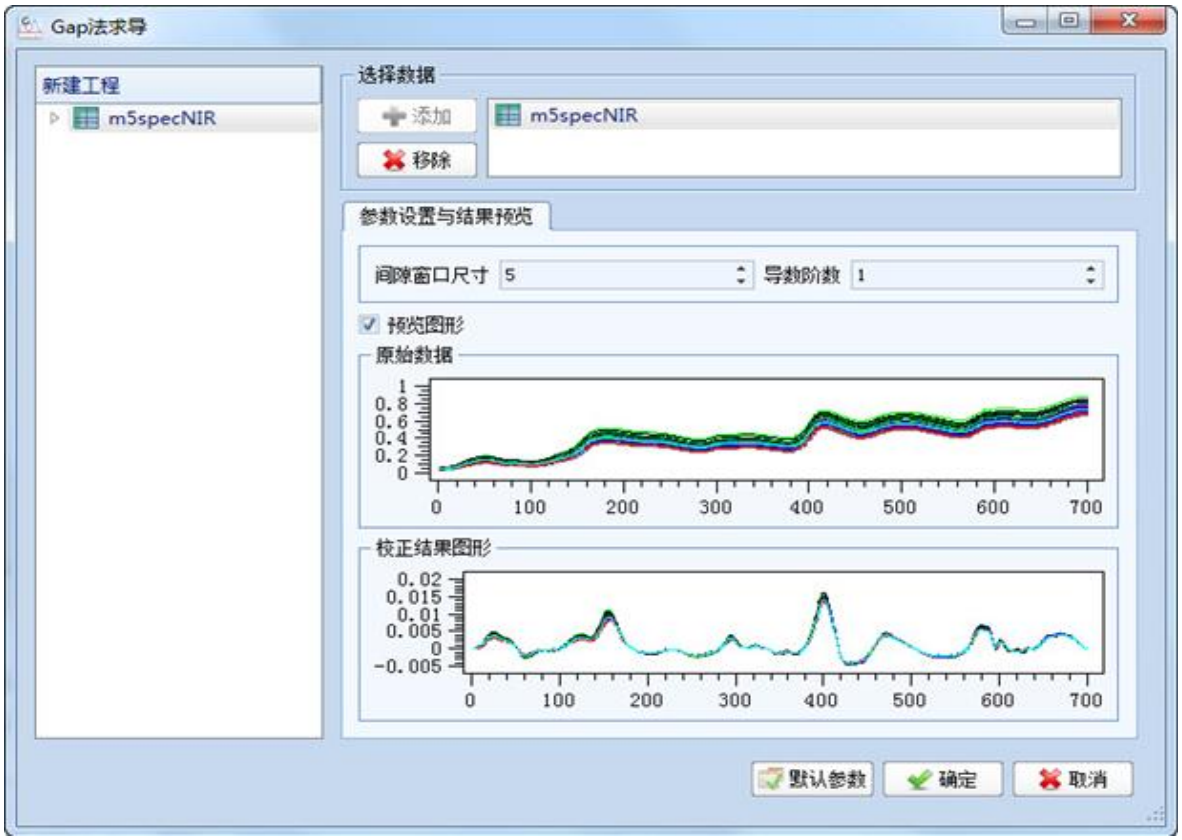
本软件所提供的求导方法如下图所示:

10.12.1. Gap 法

本法由美国著名科学家 Karl Norris 教授提出，他被认为是“现代近红外光谱之父”。除此方法以外，他亦提出 Norris 回归方法，均是近红外光谱数据预处理的重要手段。Gap 方法可提高排除干扰信号的能力，是 Gap-Seg 法求导的特殊情形，此时分割尺寸为 1。该方法要求被处理数据无缺失值，单样本含有 5 个以上变量，且均为数值。

操作步骤:

步骤 1: 点击**预处理标签** -> **求导** -> **Gap 法**，弹出如下对话框:

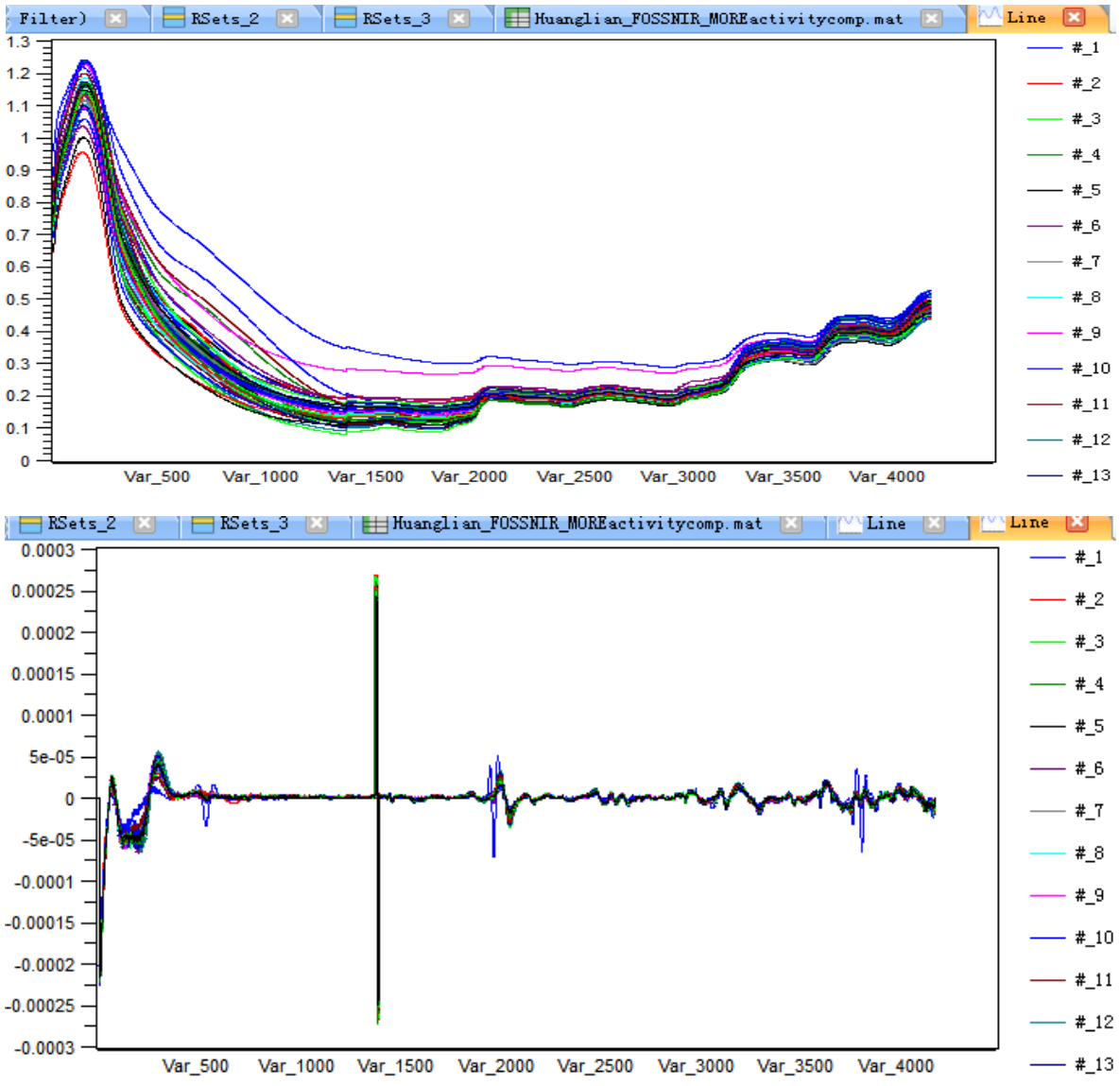


接下来的操作步骤参照预处理之通用步骤。

参数说明见下表:

参数	范围	说明
间隙窗口尺寸	[1 num-1]	当导数阶数等于 1 或者 3 的时候，间隙窗口尺寸必须为奇数；当导数阶数等于 2 或者 4 的时候，间隙窗口尺寸则无此限制。
导数阶数	[1 4]	求导阶数。

如下二图则是典型近红外数据及其二阶导数的求导结果。



10.12.2. Gap-Seg 法

该方法由 Karl Norris 教授提出。其特点在于求导时，其 X 值由该点二侧一定窗口尺寸内的数据均值决定，即该点二侧的数据段决定，且数据中间由一段数据分隔。求导点的原始值则以上述二数据均值之差来替换，从而获得该点导数的估计值。

分割尺寸是本法的重要参数，若该值太小，则结果不一定比简单差分好；若该值太大，则结果必定不代表光谱的局部行为。决定分割尺寸的参考方法是，可以是选取足够多的数据点，使其覆盖最大吸收谱峰的半峰宽，或者基于不同尺寸下模型的准确性和稳健性来选择。

操作步骤:

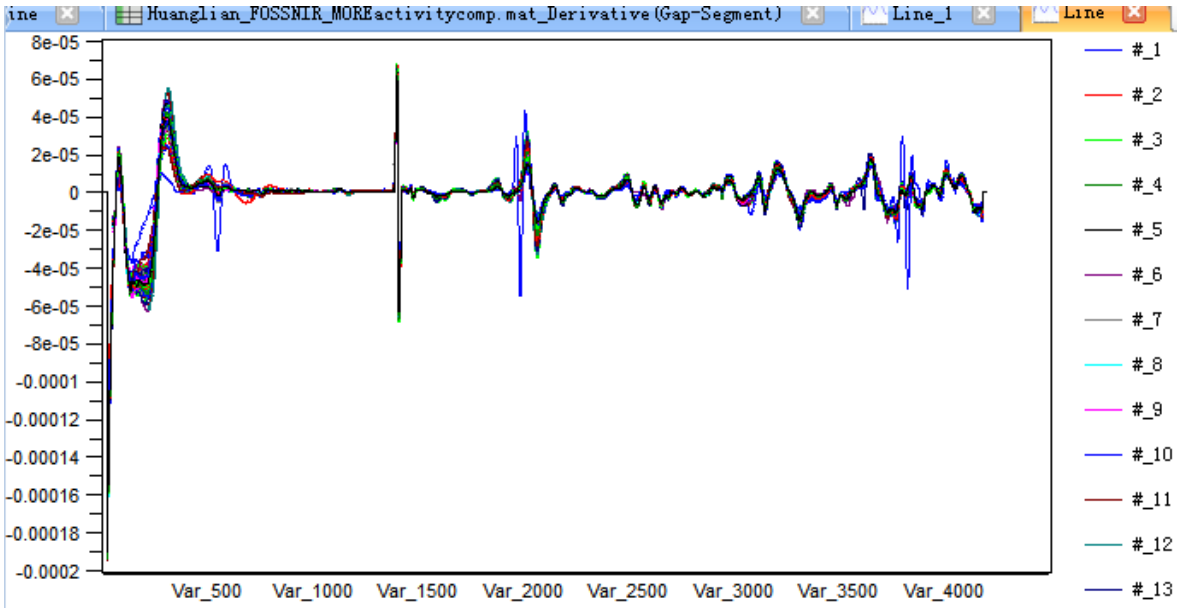
步骤 1: 点击**预处理标签** -> **求导** -> **Gap-Seg 法**，弹出如下对话框:

接下来的操作步骤参照预处理之通用步骤。

参数说明见下表:

参数	范围	说明
分割大小	[1 num-1]，其中 num 表示所选数据的长度。	X 值间距长度。当导数阶数等于 2 或者 4 的时候，分割大小必须为奇数。
间隔大小	[1 num-1]，其中 num 表示所选数据的长度。	隔断上述二分割的间距长度。当导数阶数等于 1 或者 3 的时候，间隔大小必须为奇数。
导数阶数	[1 4]	求导阶数。

如下图则是典型近红外数据的二阶导数结果，原始数据如图所示。



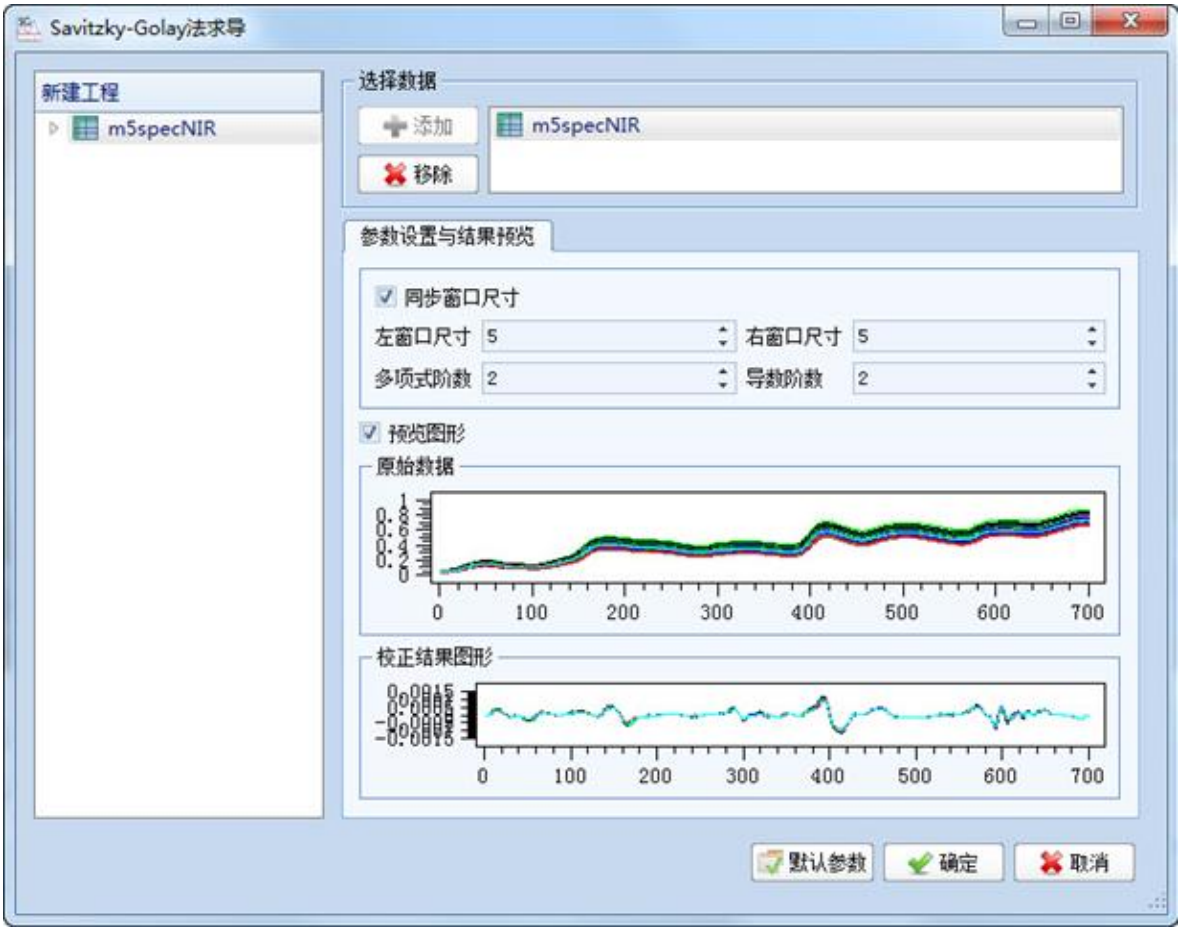
10.12.3. Savitzky-Golay 法

基于局部分割窗口，而非邻近点计算某特定数据点下的 N 阶导数，实际上使用了平滑后的，以克服噪声的影响。具体来说，先通过多项式最小二乘拟合的方式，平滑原始数据中的每个点，然后计算拟合多项式中各点的导数获得。

特别需要强调的是，窗口宽度是求导的重要参数，且求导后结果的波长点数将比原始数据少，其数目等于窗口宽度。

操作步骤:

步骤 1: 点击**预处理** -> **求导** -> **Savitzky-Golay 法**，弹出如下对话框:



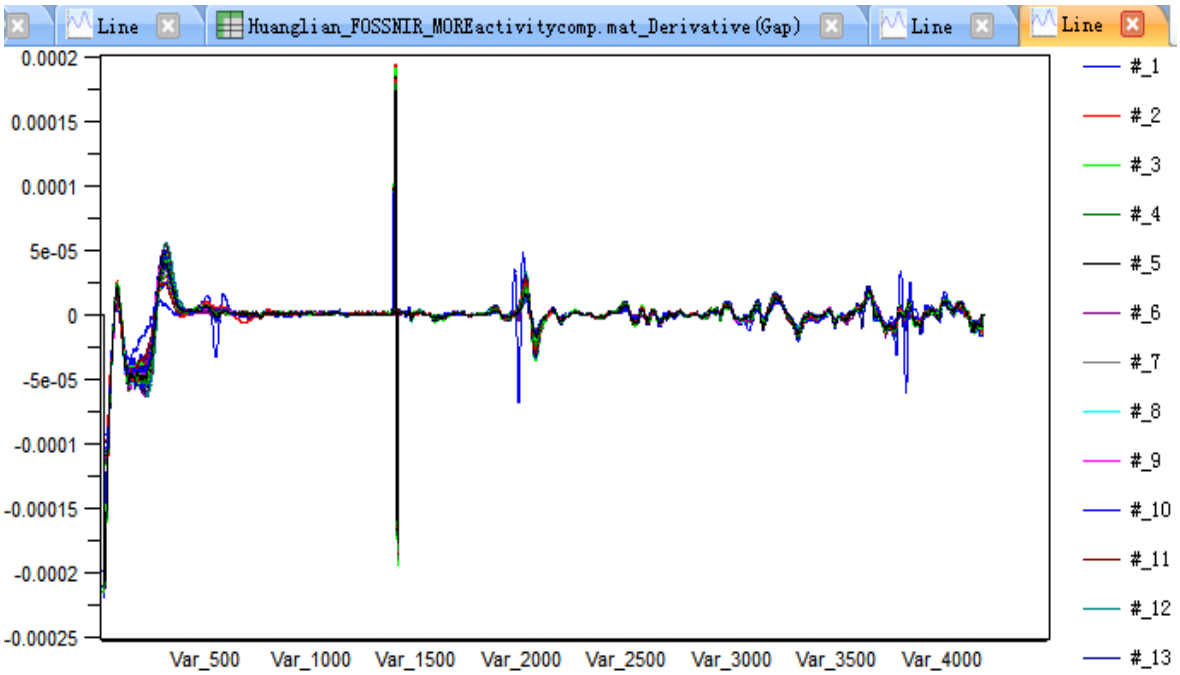
接下来的操作步骤参照预处理之通用步骤。

参数说明见下表:

参数	范围	说明
左窗口尺寸	[1 num]，其中 num 表示所选数据的长度。	如前所述，所选目标数据点左侧窗口的尺寸大小。

右窗口尺寸	[1 num]，其中 num 表示所选数据的长度。	如前所述，所选目标数据点右侧窗口的尺寸大小。
多项式阶数	[1 左窗口尺寸+ 右窗口尺寸- 1]	拟合多项式阶数。
导数阶数	[1 多项式阶数]	求导阶数。

如下图则是典型近红外数据的二阶导数结果，原始数据如图所示。



10.12.4. 直接差分法

直接采用相邻数据点获得导数，计算简单便捷，但噪声对求导的影响较大。尽管本法对光谱波长采样点较多的数据结果影响不大，但是对稀疏波长采样点光谱则可能带来较大的误差，使用时需特别注意。若出现此种情形，可换用 Savitzky-Golay 方法等。

操作步骤：

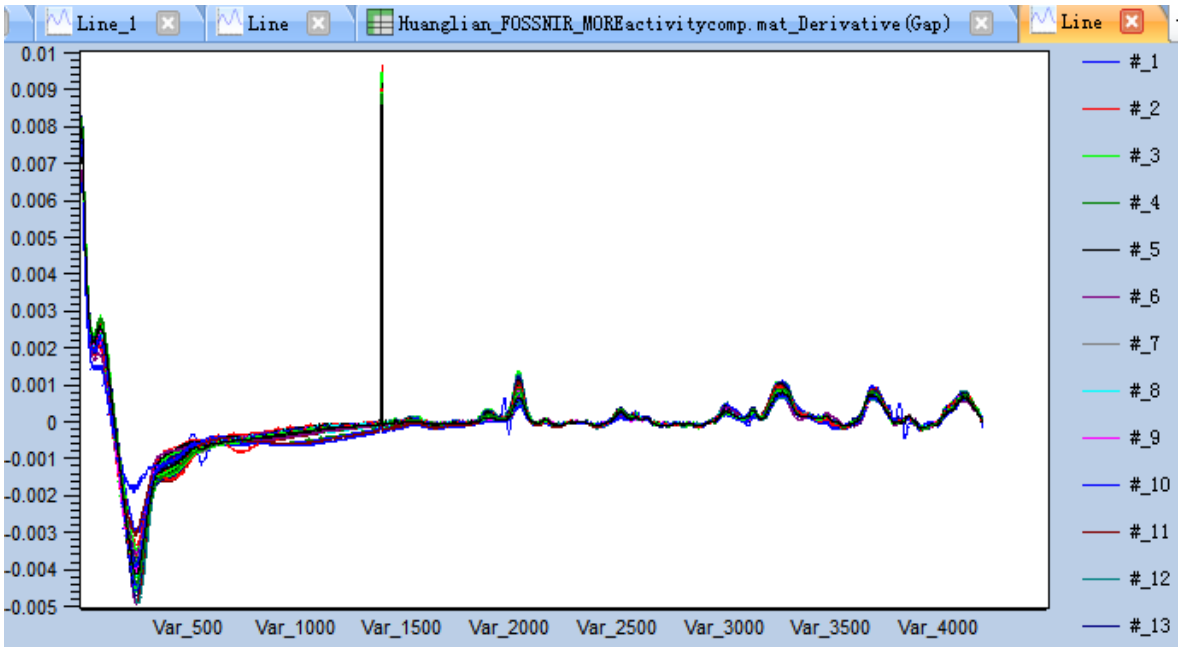
步骤 1: 点击**预处理标签** -> **求导** -> **直接差分法**，弹出如下对话框:

接下来的操作步骤参照预处理之通用步骤。

参数说明见下表：

参数	范围	说明
导数阶数	[1 num-1]，其中 num 表示所选数据的长度。	求导阶数。

如下图则是典型近红外数据的一阶导数结果，原始数据如图所示。



10.13. 背景扣除

在复杂分析仪器数据中，比如色谱和质谱等，除各组分或基团的特征峰外，通常还同时存在连续、缓慢或相对较快变化的背景，且数据背景在不同样本中亦具有差异。若在进行数据分析前不对背景进行有效校正，一方面导致难于确定目标峰的峰位置或强度，另一方面降低所建模型的稳健性和可解释性。

ⓘ 背景扣除是通过手动或自动选取背景数据点，以连续函数模型拟合的方式逼近数据点，再将其从原始数据中减去以提高数据质量。

本软件所提供的背景扣除方法如下图所示:

10.13.1. airPLS 法

本法快速，使用灵活，结果较好。其主要原理是通过迭代加权方式拟合基线与原始信号，每次迭代计算中以自适应加权惩罚的方式对拟合基线与原始信号间整体方差重加权，直至达到迭代中止条件。即在惩罚最小二乘信号平滑的基础上，在自适应迭代重加权中将其转变为背景估计的一种方法。

操作步骤:

步骤 1: 点击**预处理标签** -> **背景扣除** -> **airPLS 法**，弹出如下对话框:

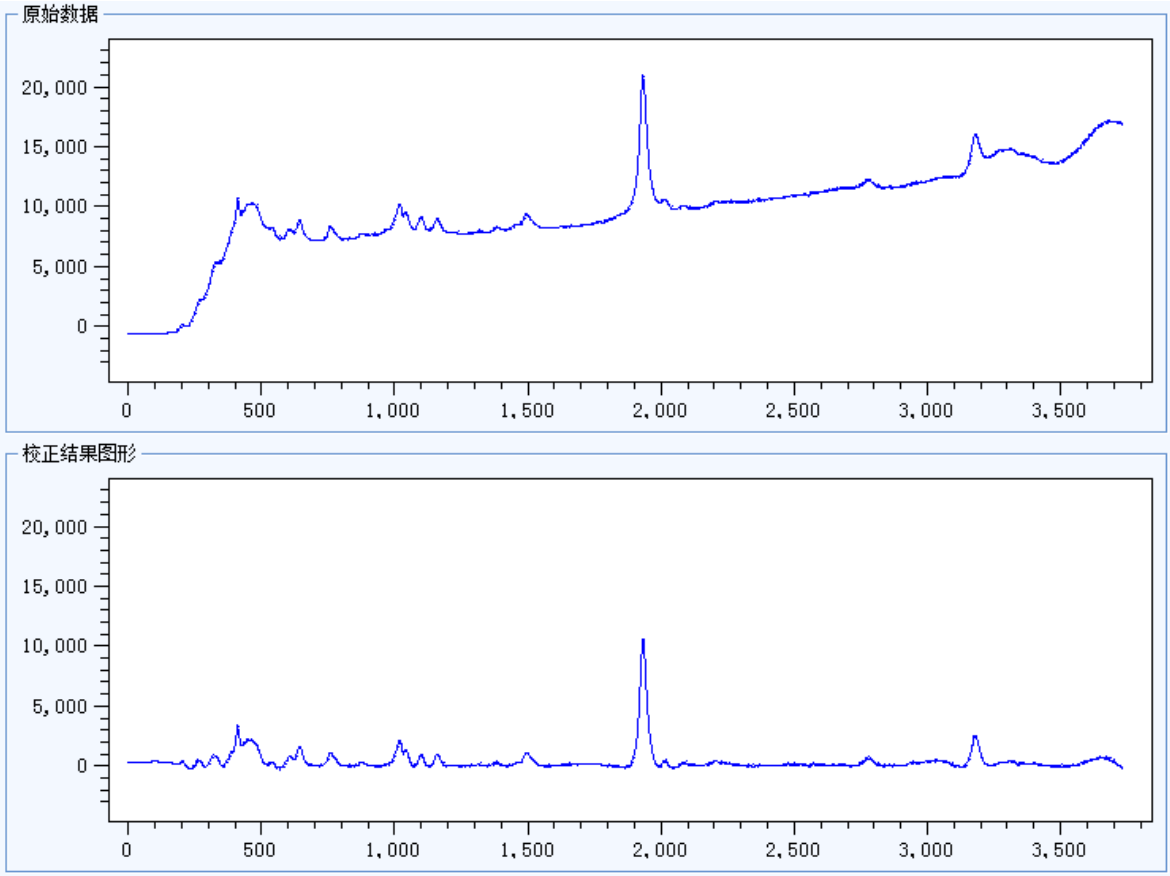
接下来的操作步骤参照预处理之通用步骤。

参数说明见下表:

参数	范围	说明
Lambda 值	[1 ∞]，其中∞表示无穷大。	粗糙度惩罚系数。
WEP 值	[0 0.5)	计算加权惩罚权重向量的参数，以决定该向量节点处的值。
DP 值	(0 1]	迭代中当前步骤的权重替换值。
阶数	[1 9]	拟合数据粗糙度的阶数。
噪声水平	[0 1]	噪声水平。
最大迭代次数	[1 ∞]，其中∞表示无穷大。	最大迭代运算的次数。

在上表所列的参数中，Lambda 值最为关键，若背景扣除结果不理想，通常可通过调节(增

大)该参数来达到。如下二图则是典型拉曼光谱数据及其背景扣除的结果。



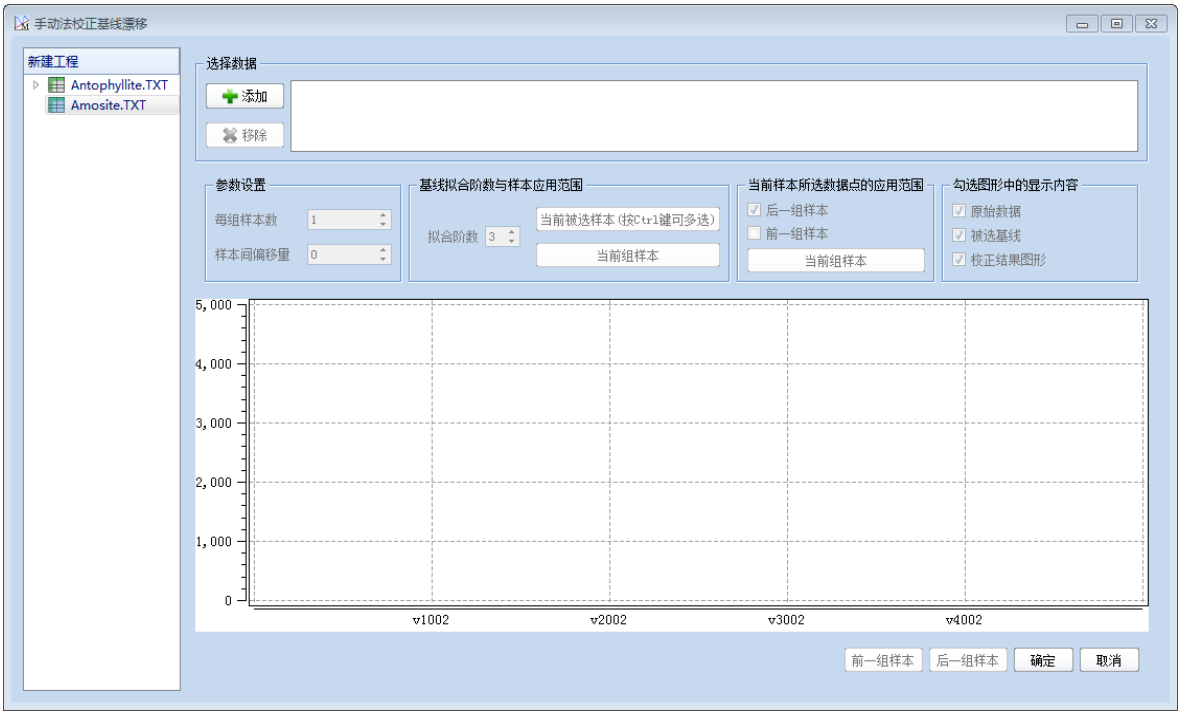
10.13.2. 手动法

手动法通过从原始数据(如光谱或色谱)中人工判断并选择可能的背景数据点，以自定义模型(如局部线性或多项式函数)拟合的形式获得整个数据背景，最后将其从从原始光谱中扣除以达到背景扣除的目的。其优点是可处理较为复杂的背景漂移情形，且结果通常不会太差，但强烈依赖于使用者的实际经验，且因数据点选择的差异，导致结果往往很难重复。

本软件以多项式函数拟合数据背景。

操作步骤:

步骤 1: 点击**预处理标签** -> **背景扣除** -> **手动法**，弹出如下对话框:



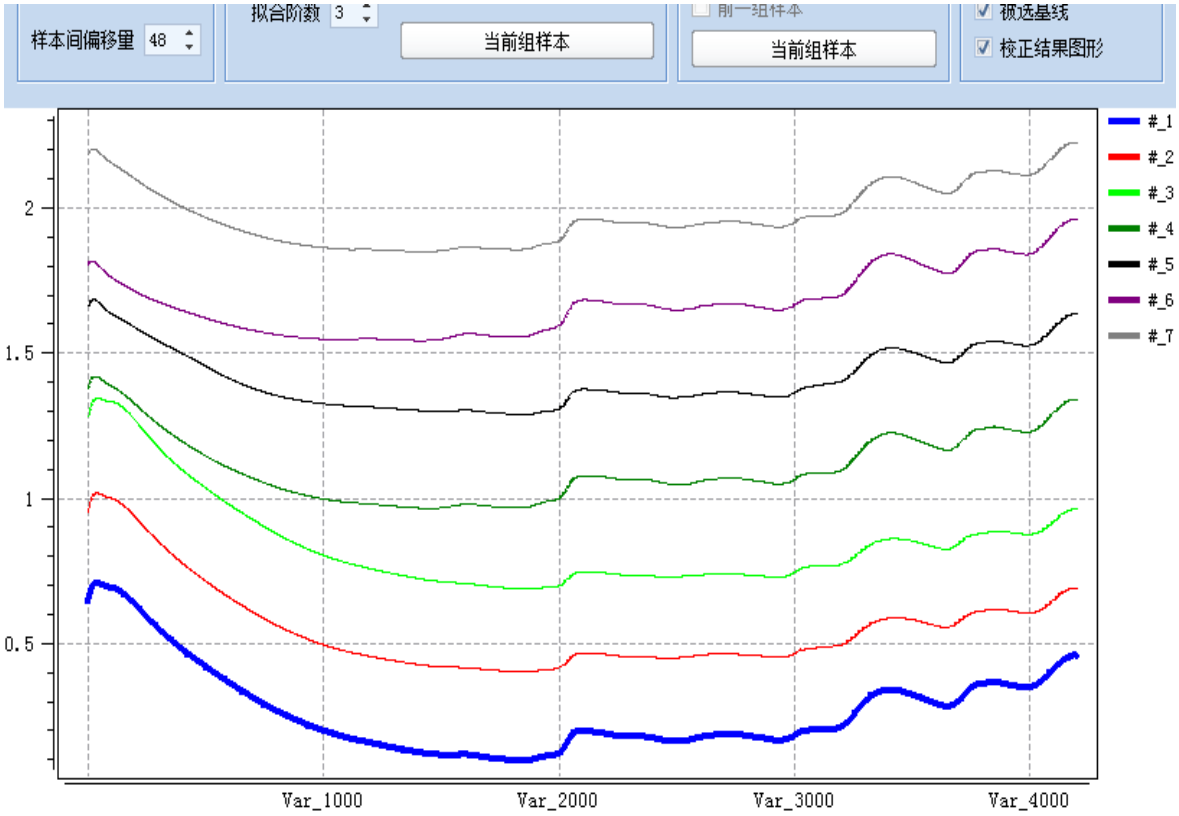
接下来的操作步骤参照预处理之通用步骤。

实因手动背景扣除涉及的问题较多，为了达到良好的用户体验，本软件提供相对于自动背景扣除复杂的使用界面，其操作一一介绍如下。

1) 参数设置：针对手动背景扣除中涉及的主要参数进行预先设置，参数的具体意义如下。

参数	范围	说明
每组样本数	[1 num]，其中 num 表示数据样本总数。	被处理的数据样本可能很多，每次在当前窗口中处理的数据总有限，该参数是指每次成批被处理的样本数目。
样本间偏移量	[0 ∞]，其中∞指无穷大。	该参数实现绘制当前组样本时，不同样本间一定的间距错开，避免重叠，以方便背景选择如下图所示。

修改样本间偏移量，可得到如下所示图形。

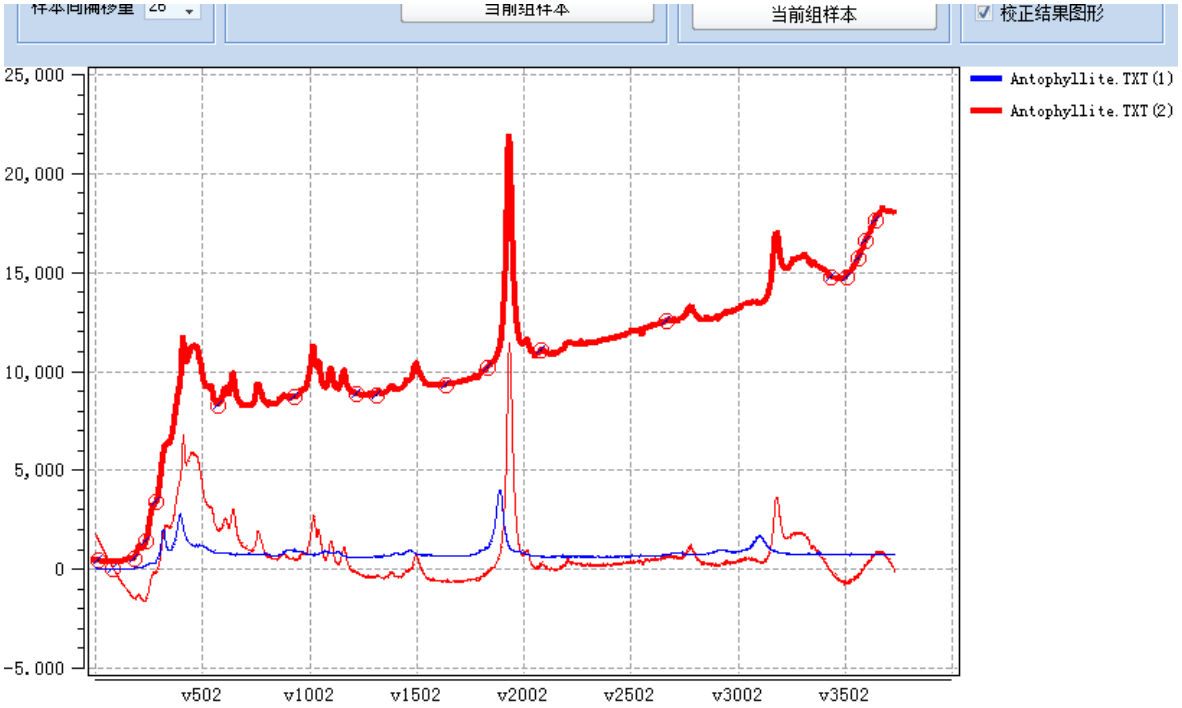


2) 基线拟合阶数与样本应用范围：设置多项式拟合阶数，并应用于目标样本，以便进行背景扣除。

用户先设置基线拟合阶数(或使用默认值)，该设置将仅作用当前组的当前样本。若用户希望对当前组的不同样本分别设置拟合阶数，则可分别选择这些样本，再改变阶数值。若希望将当前针对样本所设置的阶数值，作用于当前组的其他左右样本，则可通过点击当前样本做来完成。

特别需要注意的是，用户选择的数据点数目，满足多项式拟合的要求时，当前组样本的所选样本将直接拟合得到结果。

若用户选择绘出图形时，将直接在当前界面获得图形结果，如下图所示。



在上图中，最上面的红线为当前被处理的样本，加粗显示；红色上圆圈点所表示的位置，则是用户自定义的数据背景，而下面的红线是该线扣除背景后得到的结果；中间的蓝色则是另一未被处理的样本。

用户特别需要注意的是，当前被界面中正在被处理的样本，其线条将以加粗的形式表示，以方便用户识别。

此外，用户亦须了解：点击鼠标左键，选择并标记数据背景点，若同时按住 Ctrl 键，可实现同时多选数据样本，以实现同时参数设置；点击鼠标右键，则去除对背景数据点的选择。

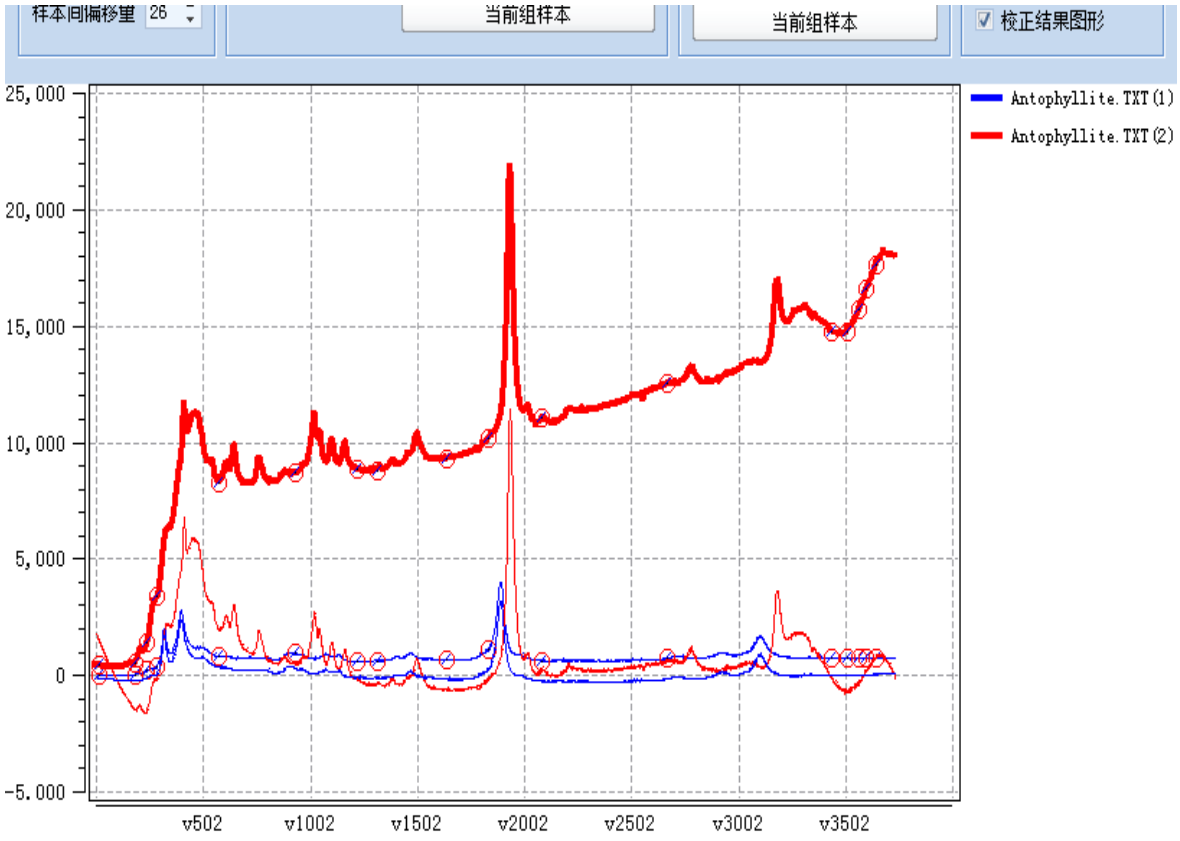
参数	范围	说明
拟合阶数	[1 ∞]，其中∞指无穷大。	多项式的拟合阶数。

3) 当前样本所选数据点的应用范围：指当前被选样本上所定义数据背景点的位置，将被应用于哪些样本，包括后一组样本，前一组样本和当前组样本三种可能。

将当前样本所选数据背景点应用于后一组样本或前一组样本，可通过勾选复选框来达到，而直接作用当前样本组中的其他样本，则可通过点击按钮 当前组样本 来达到，

此时若数据点数目满足背景的拟合需要，所有样本将直接得到背景扣除后的结果，如下图所示。

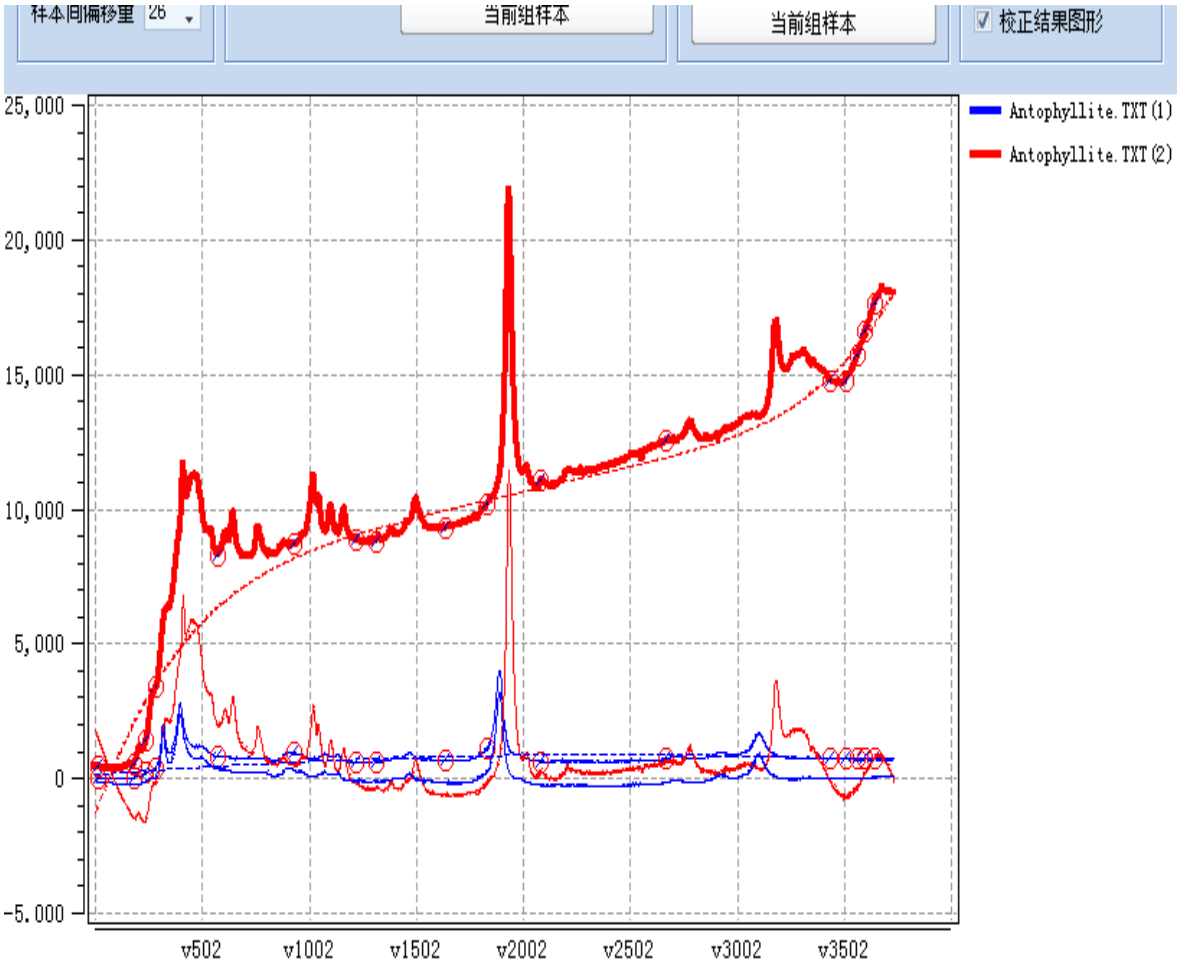
本功能对于大样本数据的处理是费用有用的，用户只需对其中一个样本选择数据背景点，即可将其作用其他样本，省去对每个样本进行背景标记的麻烦。但亦需考察从当前样本所选择的背景数据点，是否完全适合于所有其他样本，必要的时候可去除某些点，或者增加某些点。



上图为在图的基础上，点击当前组样本的结果，图中的蓝线亦同时自动添加上图中红色粗线所选的背景数据点(亦以红色圆圈表示)，并获得背景扣除后的结果。

4) 勾选图形中的显示内容：指背景扣除界面中所显示的图形内容。

可显示的图形内容包括：原始数据，被选基线以及校正结果。用户可以通过复选框决定是否需要显示这些内容。勾选，则表示显示；反之，则为显示，如下图所示。



与上图相比，本图多显示了红色和蓝色二条虚线，分别为红色和蓝色线条所代表样本的拟合背景。

5) 跳转到前一组样本或后一组样本：指通过切换当前被处理的样本组，实现对被选数据中所有样本的分析处理。

用户可通过点击按钮 前一组样本 ，跳转当前被处理的数据到前一组样本，或点击按钮 后一组样本 ，跳转数据到后一组样本。

ⓘ 无论当前样本是否已经完成背景扣除，用户均可通过该功能实现对其他样本组的处理，使用非常方便灵活。

10.13.3. airPLS 与手动联合法

该功能实现上述二种背景扣除方法的联合使用，以充分发挥二种方法的优点和长处，达到最佳的背景扣除效果和用户体验。

操作步骤:

步骤 1: 点击**预处理标签** -> **背景扣除** -> **airPLS 与手动联合法**，同样弹出如图所示的对话框:

步骤 2: 参照背景扣除 -> airPLS 法的操作步骤，产生经过 airPLS 结果扣除背景的结果，再点击按钮 

，即弹出图所示的对话框:

步骤 3: 参照背景扣除 -> 手动法的操作步骤，得到最终结果。

10.13.4. 线性补偿法

本法通过如下方式校正数据背景，即:

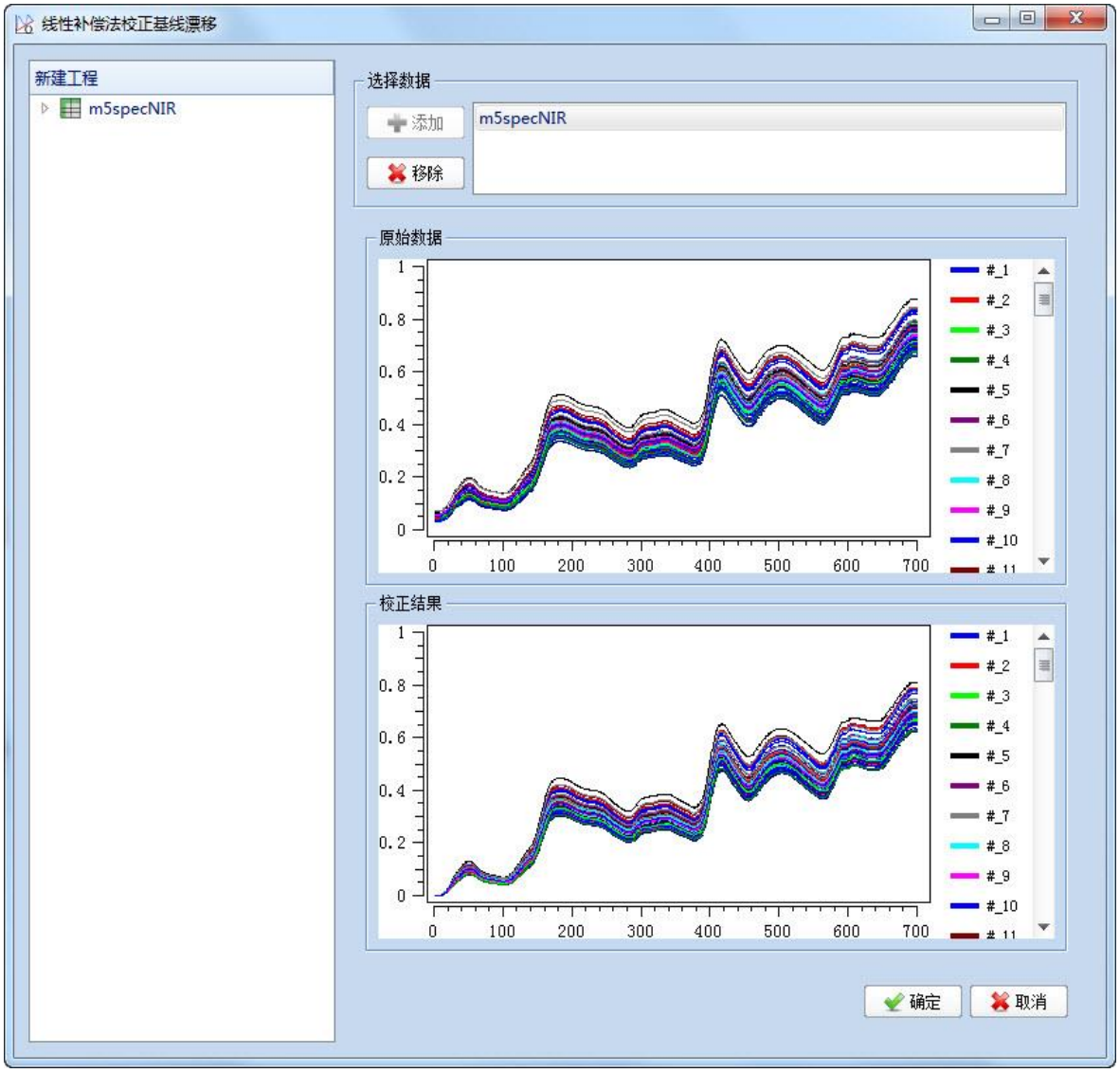
$$f(x_i) = x_i - \min(\mathbf{x})$$

其中，数据 x_i 和 **x** 分别表示目标数据点和样本向量。

大部分方法扣除背景，实因数据的的复杂性和方法的局限性，背景校正后的结果，某些数据点可能存在负数的情形，影响其后的数据处理或模型构建。而通过该方法扣除背景，可以保证所得到的数据结果值均非非负数。

操作步骤:

步骤 1: 点击**预处理标签** -> **背景扣除** -> **线性补偿法**，弹出如下对话框:

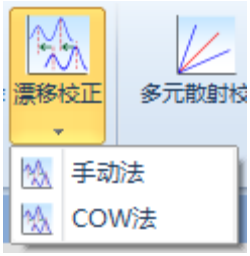


接下来的操作步骤参照预处理之通用步骤。

10.14. 漂移校正

本软件所述漂移是指不同样本或同一样本，在不同实验条件下所得到的数据，它们的相同组份或信号间存在沿着化学坐标方向(如色谱保留时间或光谱波长)的偏移。若不经过任何处理，直接对这样的数据进行处理，显然将极大地影响分析结果。比如同一复杂生物血液样本的二次色谱重复进样实验，即使实验条件完全相同，他们中相同代谢小分子的保留时间亦存在差别。使用原始数据进行聚类分析或模型构建等，其结果必然产生某些偏差，甚

至错误的结论。漂移校正是指通过数学模型的方法，校正上述差异。本软件所提供的漂移校正方法，如下图所示:

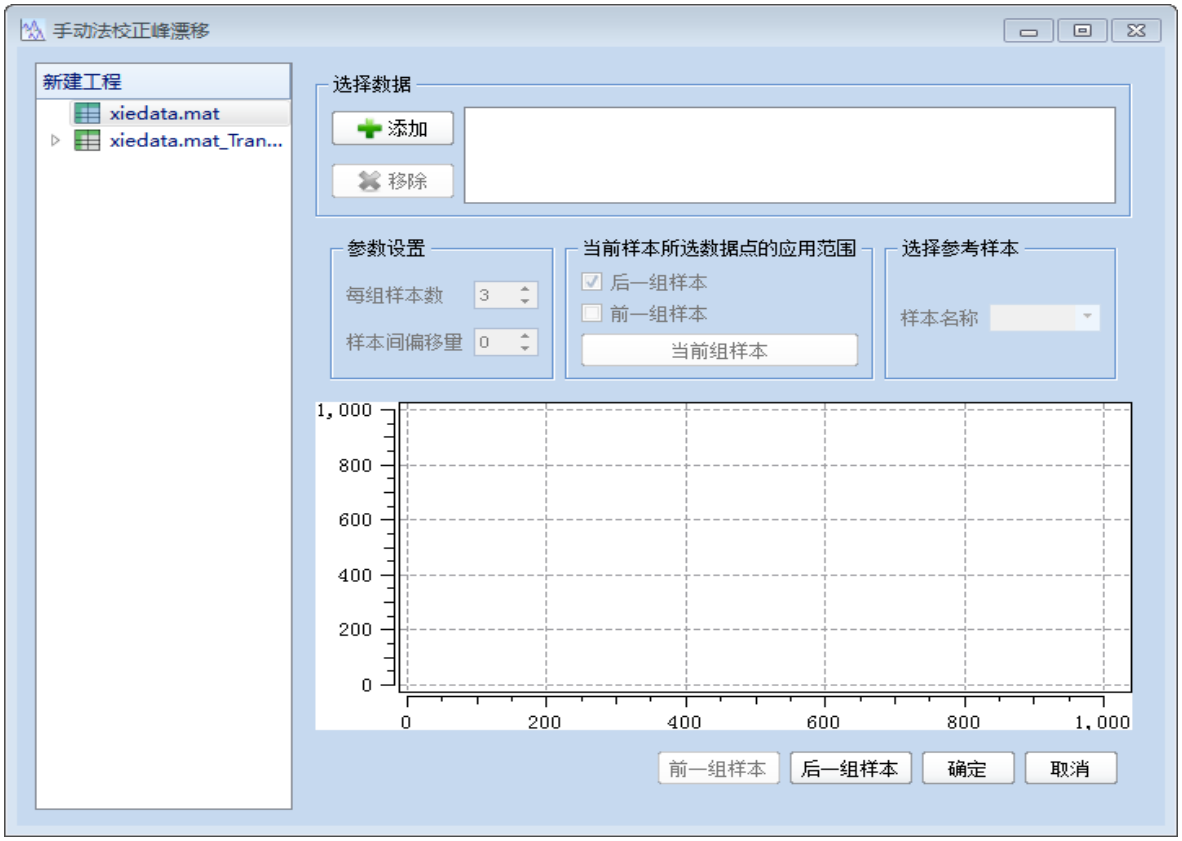


10.14.1. 手动法

手动法校正样本间漂移，与手动法扣除数据背景雷同。其区别在于背景扣除中所选择的数据点，用于单个样本数据中的背景扣除，而漂移校正中所选数据点，用于与参考样本中所选的目标数据点比较，校正二个样本间的漂移。

操作步骤:

步骤 1: 点击**预处理标签** -> **漂移校正** -> **手动法**，弹出如下对话框:



上图与图非常类似，但亦有较大差别。接下来一一介绍各个功能及其使用方法。

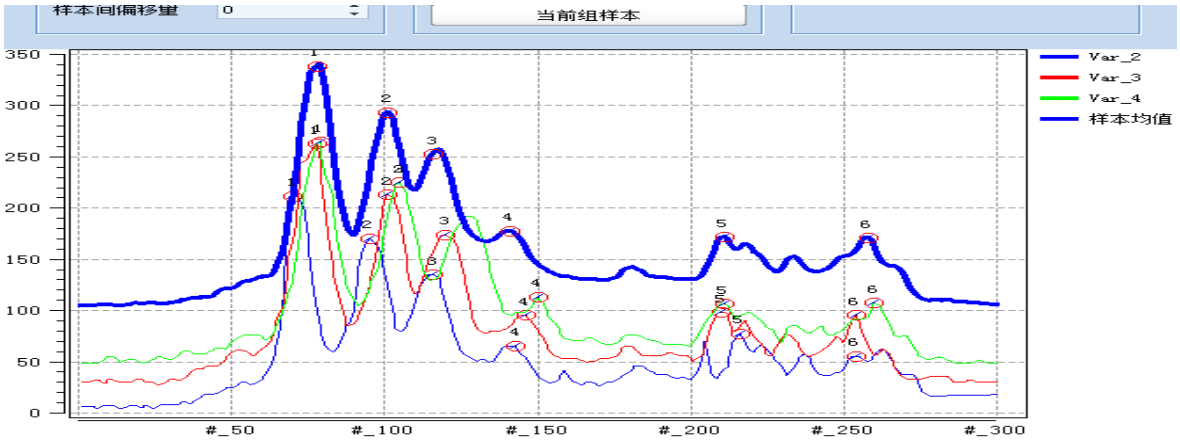
1) 参数设置：针对漂移校正中涉及的主要参数进行预先设置，参数的具体意义如下。

参数	范围	说明
每组样本数	[1 num]，其中 num 表示数据样本总数。	被处理的数据样本可能很多，每次在当前窗口中处理的数据总有限，该参数是指每次成批被处理的样本数目。
样本间偏移量	[0 ∞]，其中∞指无穷大。	该参数实现绘制当前组样本时，不同样本间一定的间距错开，避免重叠，以方便漂移校正，具体如图所示。

2) 当前样本所选数据点的应用范围：本部分内容与背景扣除雷同，可参考。其主要意义指当前窗口中参考样本上所选择的校正参考点位置，将被应用于哪些样本，包括后一组样本，前一组样本和当前组样本三种可能。

特别需要注意的是，用户在选择参考样本上的数据点后，亦可在实际样本点上任意调解或修改校正点位置。但需要注意的是个样本数据点将按照先后顺序依次使用，即在实际校正时，参考样本中标记的第一个数据点，将于实际样本的第一个数据点校正，依此类推。

若数据点标记时，由于样本所包含的组分太多，导致某些组份的标记困难，则可使用图形的局部放大功能，标记后的数据点如下图所示。

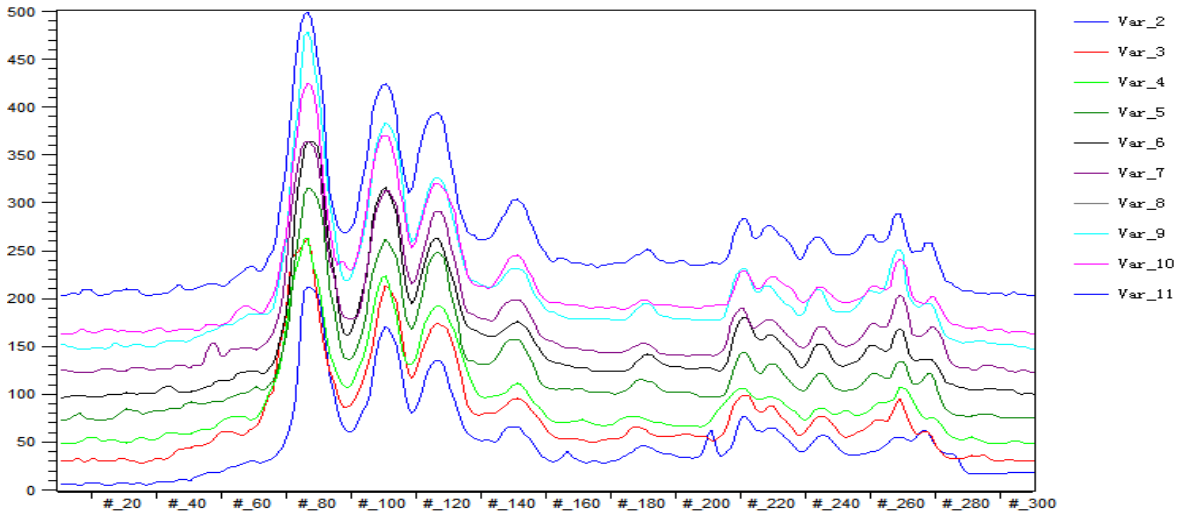


与此同时，本软件提供智能寻找峰顶点的功能，即在将参考样本中的标记点作用于其他样本时，将自动寻找对应组份的顶点，以便快速完成漂移校正。

3) 选择参考样本：指用户自定义用于漂移校正的参考样本。定义好校正参考样本后，其他样本将均以该样本为标准进行漂移校正，以保证校正后的结果，具有相互间的可比性。用户既可选择任一数据样本作为参考样本，亦可选择数据均值作为参考样本。

4) 跳转数据到前一组样本或后一组样本：此部分亦与背景扣除雷同，用户可参考。

上述数据通过手动漂移校正后，得到如下图形结果。



10.14.2. COW 法

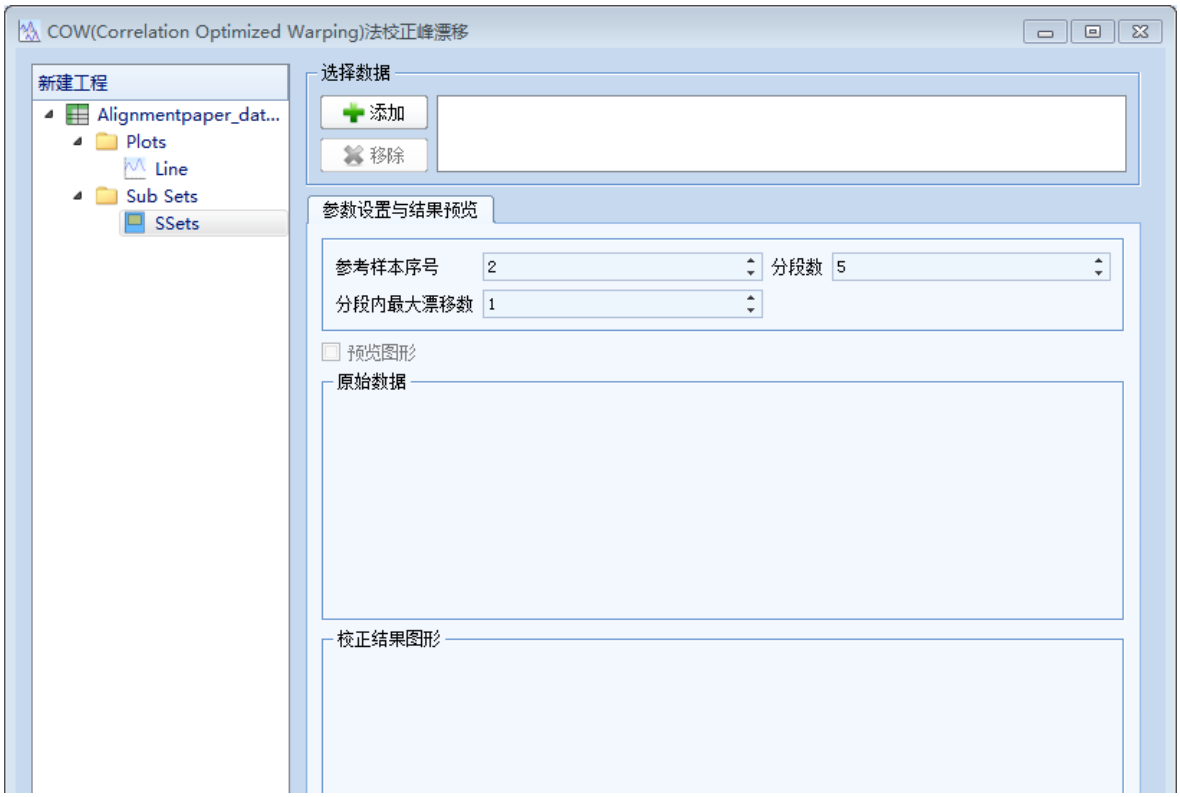
本法由 Nielsen N-PV 提出，使用非常广泛，已经成为色谱、拉曼和 NMR 峰漂移校正的最重要方法之一，无须任何数据预处理步骤，所需参数亦可由色谱峰宽估计得到。同样地，该方法亦不可用于非数值型数据或含有缺失值的数据，且消耗计算时间。

COW 法运算主要包括三个步骤，首先将参考和被校正的数据样本分割成 I 段，然后在每段数据内沿 X 轴线性伸缩目标数据长度 t，最后计算校正后样本与参考样本间的相关系数，以评价结果。

本软件中上述参数均由使用设定，实际上亦有研究人员发展了系列方法优化参数，以使程序自动达到较好的校正结果。

操作步骤:

步骤 1: 点击**预处理标签** -> **漂移校正** -> **COW 法**，弹出如下对话框:

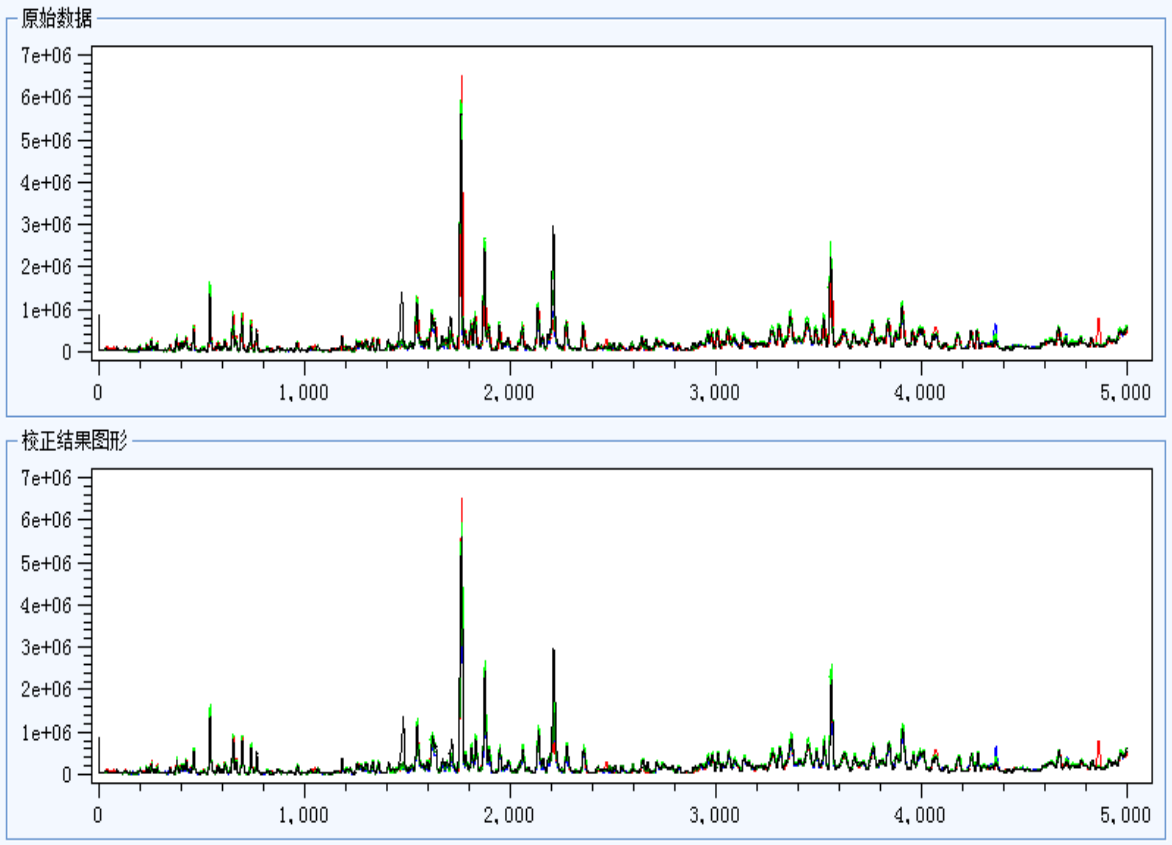


接下来的操作步骤参照预处理之通用步骤。

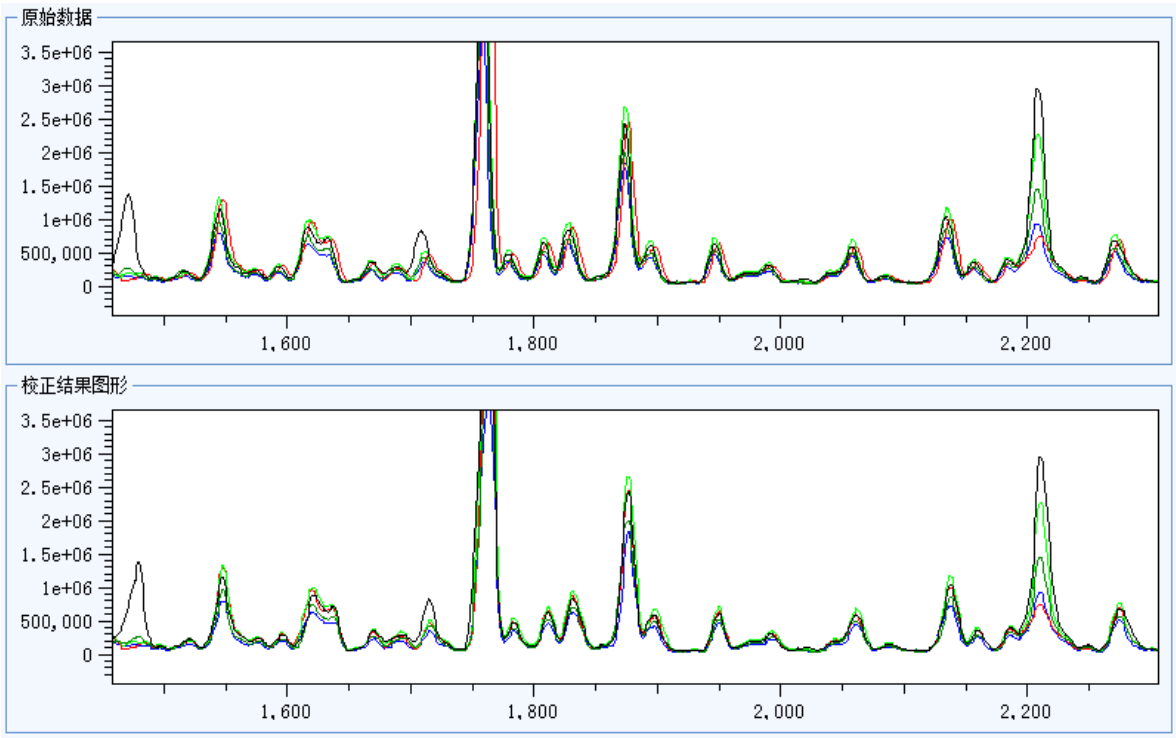
参数说明见下表:

参数	范围	说明
参考样本序号	[1 num]，其中 num 表示数据样本总数。	决定校正样本的参考样本。
分段数	[5 col-1]，其中 col 表示所选数据的长度。	将样本分割成 / 段。
分段内最大漂移数	[1 5]	每段内数据的伸缩长度。

如下二图即为典型数据的校正结果。



下图为上述图中一段典型数据的局部放大图，以更好地查看数据的校正结果。



从上图中可以看出，即时对较复杂数据，COW 法同样能得到较好的校正结果。

10.14.3. COW 与手动联合法

暂略。

10.15. 多元散射校正

多元散射校正由 Geladi 等人提出，至今已有 30 余年的历史，主要用于消除光谱数据中因样本颗粒分布的不均匀，以及颗粒大小差异而导致的光散射影响。本法是建立在各样本在各波长点量测下，其散射系数相同，且实际光谱与所谓的“理想”光谱呈线性关系这一基本假设的基础上。基于此，本法首先计算平均光谱替代实际上并不存在的“理想”光谱，然后将实际光谱与其构建并拟合线性回归模型，最后以如下形式校正实际光谱，获得多元散射校正后的光谱。

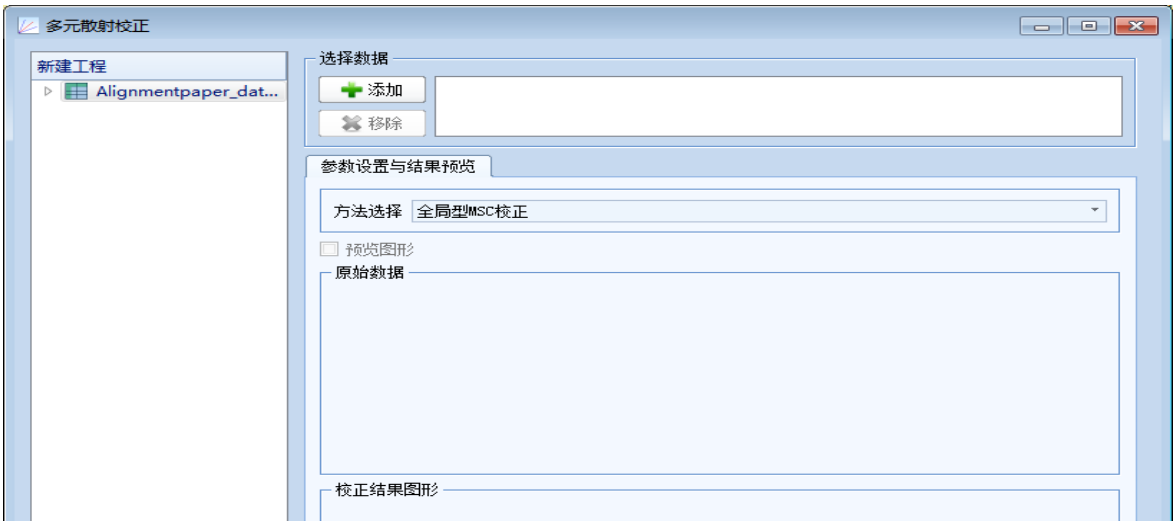
$$\mathbf{x}_{new} = (\mathbf{x} - b) / a$$

其中，a 和 b 分别表示所建线性模型的斜率和截距，x 则为实际量测光谱。

实因颗粒物的光散射与量测光谱并不存在严格的线性关系，在多元散射校正的基础上，有学者提出分段多元散射校正、循环多元散射校正以及扩展多元散射校正等。

操作步骤:

步骤 1: 点击**预处理标签 -> 多元散射校正**，弹出如下对话框:

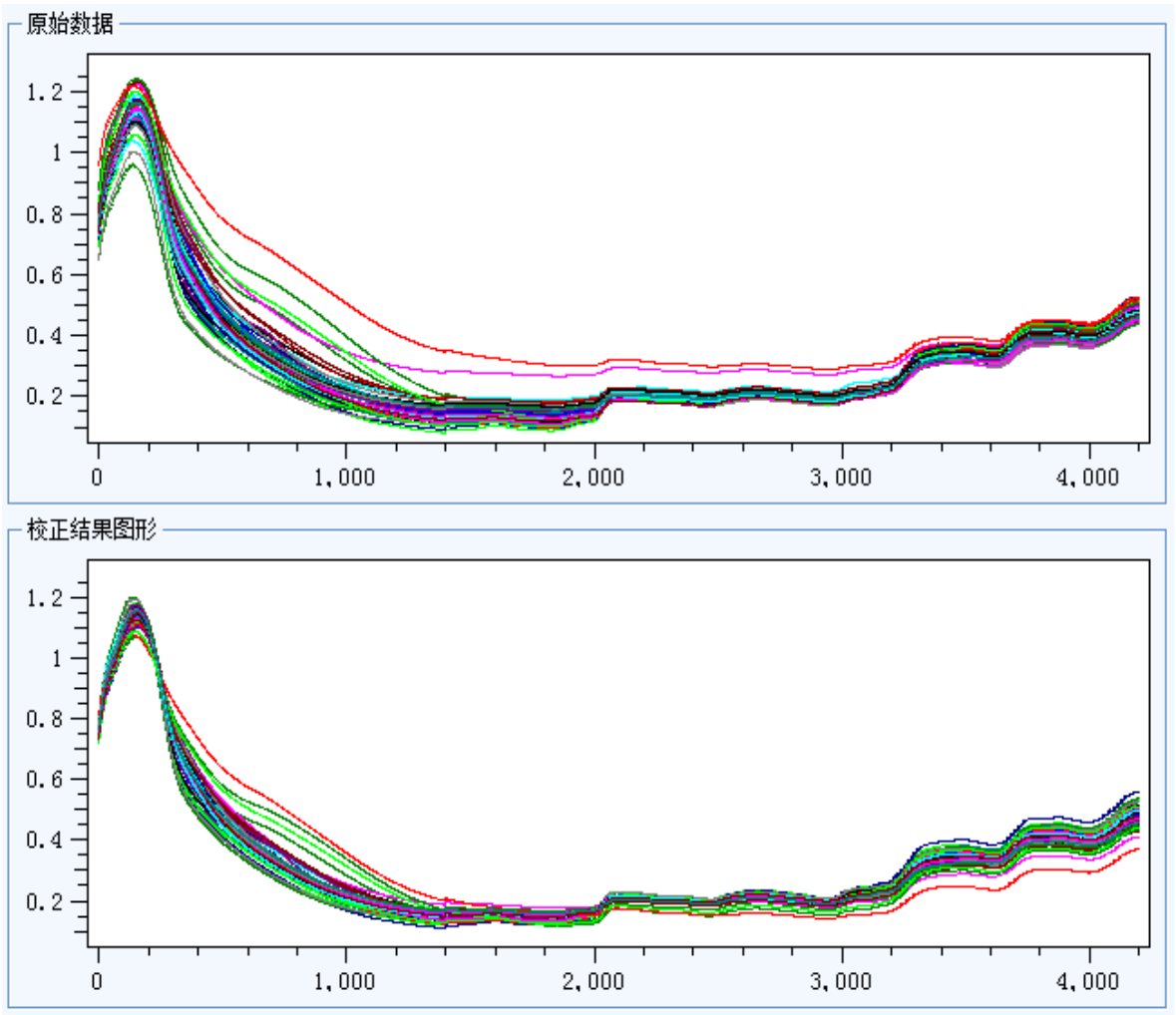


接下来的操作步骤参照预处理之通用步骤。

参数说明见下表:

MSC 方法序号	方法名称	说明
1	全局型 MSC 校正	使用模型 $\mathbf{x}_{new} = (\mathbf{x} - b) / a$ 校正光谱。
2	消除型 MSC 校正	使用模型 $\mathbf{x}_{new} = \mathbf{x} - b$ 校正光谱。
3	扩增型 MSC 校正	使用模型 $\mathbf{x}_{new} = \mathbf{x} / a$ 校正光谱。

如下二图则为典型近红外数据的多元散射校正结果。



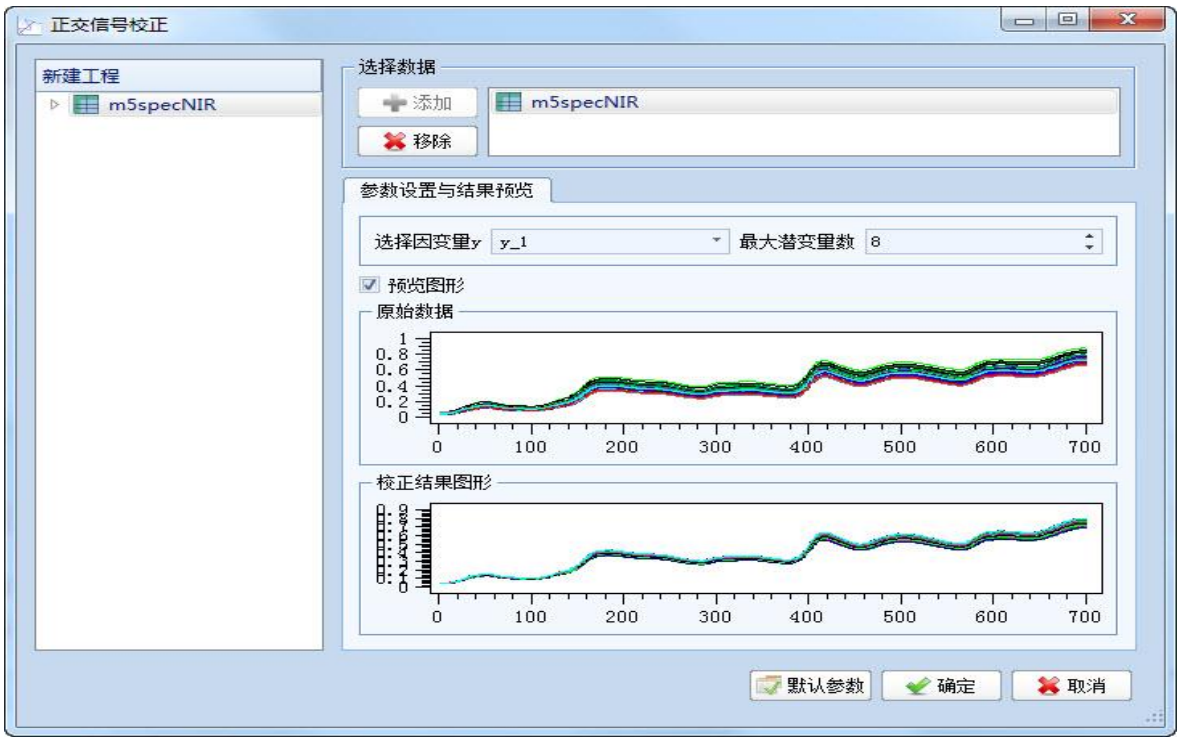
10.16. 正交信号校正

正交信号校正最早由国际著名化学计量学家 S Wold 提出，是为数不多的利用到浓度矩阵(本软件中所述因变量 **y**)信息的数据预处理方法，在去除所谓的光谱无用信息方面特别有效，以达到简化模型，并提高其稳健性与泛化能力，使用广泛。但特别需要指出的是，若将正交信号校正方法与偏最小二乘法联合使用，并不一定显著提高模型的预测能力，其原因在于偏最小二乘法本身具有一定的消除不相关变量的能力。若采用偏最小二乘方法构建模型后，第一潜变量可解释 >80%比例的 **X** 数据矩阵方差，却仅能解释 <15%比例的 **y** 响应方差，则使用正交信号校正方法对于提高模型的准确性是很有益的。

此外，由于使用本法的目的在于消除数据矩阵 **X** 中与向量变量 **y** 无关的信息，因此 y 量测的准确性对结果的影响很大，即 **y** 越准确，建模结果及其预测结果亦越准确，反之亦然。

操作步骤:

步骤 1: 点击**预处理标签** -> **正交信号校正**，弹出如下对话框:



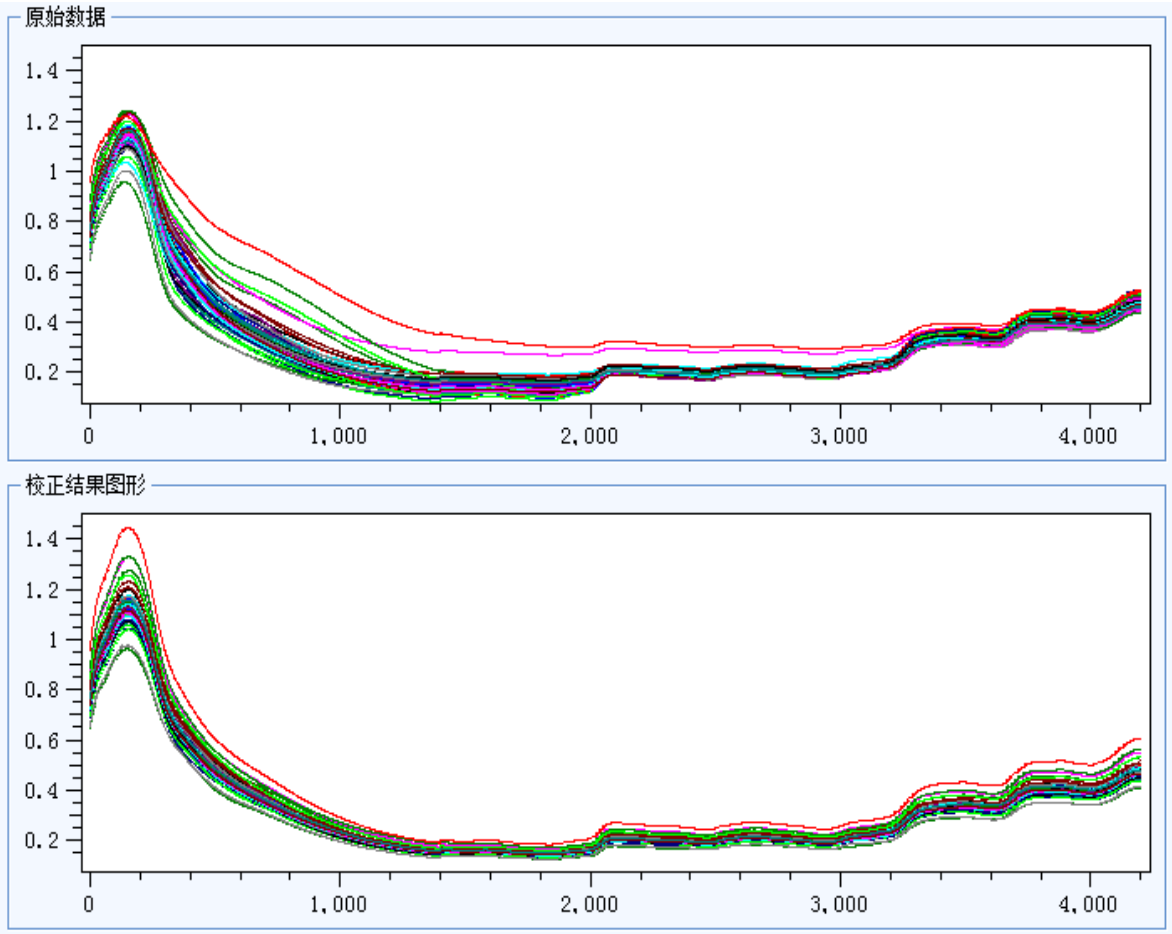
接下来的操作步骤参照预处理之通用步骤。

参数说明见下表：

参数	范围	说明
因变量 **y**	从因变量 **y** 数据集中选取。	取决于实际数据情况。
最大潜变量数	[1 col-1]，其中 col 表示所选数据的长度。	通常选取信号校正后的 1-3 个潜变量便足够。

若数据中不包含因变量 **y**，则系统将显示提示信息。

如下二图则为典型近红外数据的正交信号校正结果。与其他的数据与处理方法一样，接下来的数据处理则在校正后数据的基础上，更进一步进行分析。



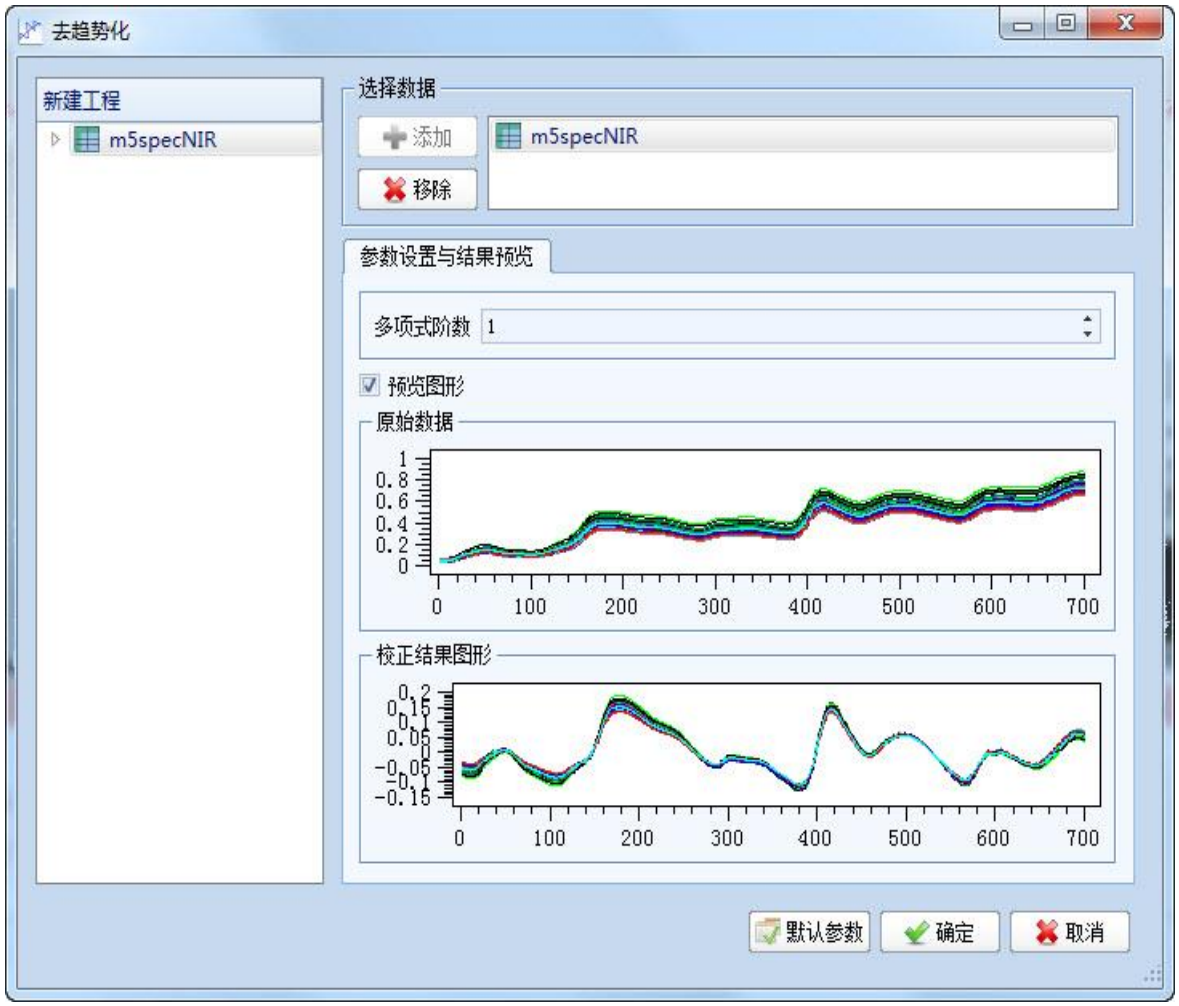
10.17. 去趋势化

顾名思义，去趋势化在于去除光谱数据中因散射现象导致的非线性数据趋势，1989 年由 Barnes R.J.等人，与 SNV 变换在同一篇论文中提出，这二种方法的组合使用，更可有效降低多重共线性和基线漂移等的影响。

本法基于多项式最小二乘拟合方法，计算原始光谱的多项式拟合函数，并将其从原始数据中扣除。与背景扣除类似，但此处并不仅仅单纯考虑数据背景，而是整个数据的趋势。随多项式使用阶数的变化，本法所扣除的基线效应亦不断变化。不难理解，零阶为简单偏移补偿，一阶为偏移加斜率补偿，二阶则可同时达到偏移、斜率和曲率补偿。

操作步骤:

步骤 1: 点击**预处理标签** -> **去趋势化**，弹出如下对话框:

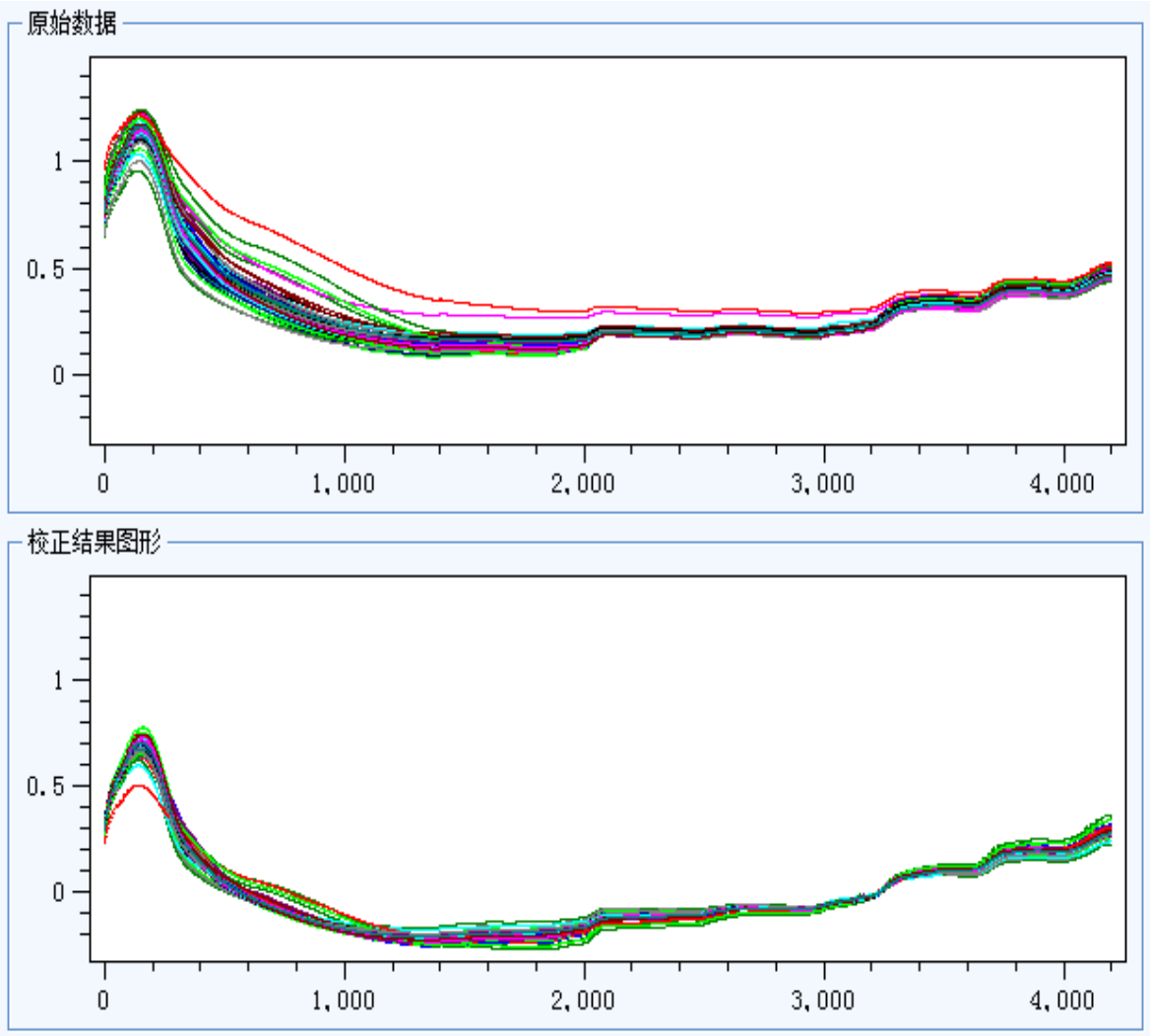


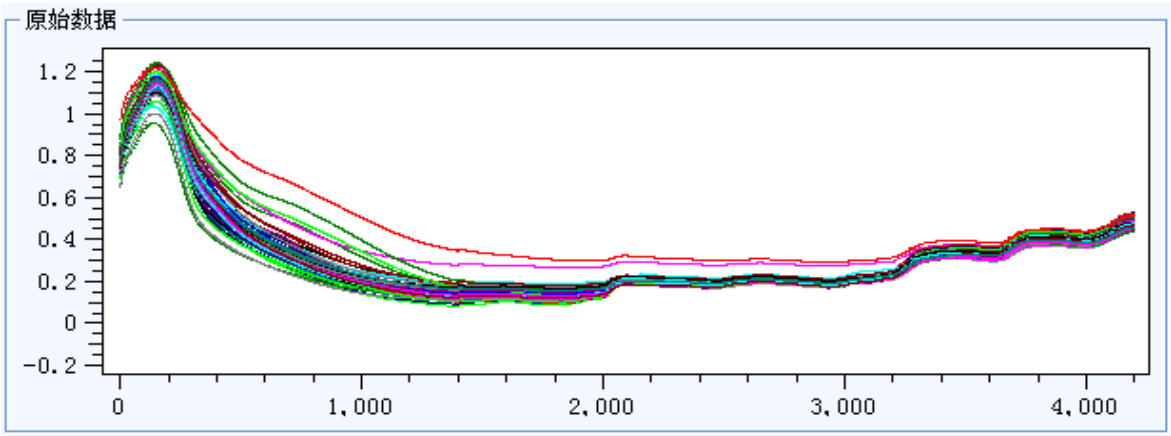
接下来的操作步骤参照预处理之通用步骤。

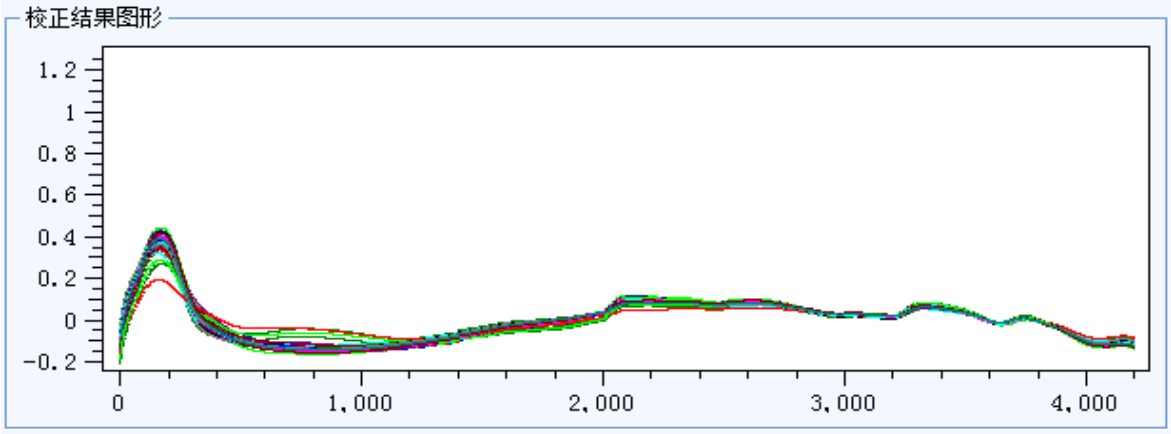
参数说明见下表:

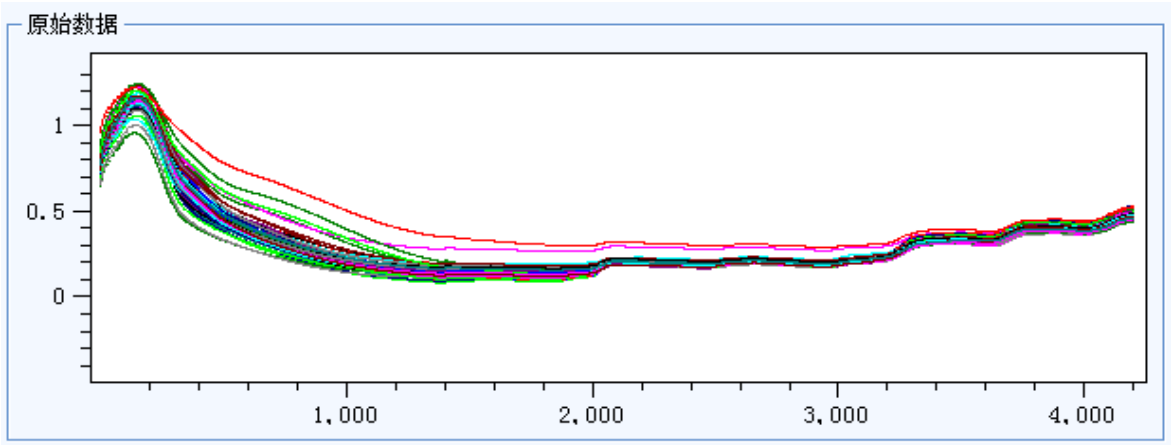
参数	范围	说明
多项式阶数	[0 ∞]，其中∞表示无穷大。	多项式拟合阶数。实际使用中以三阶或三阶以下居多。

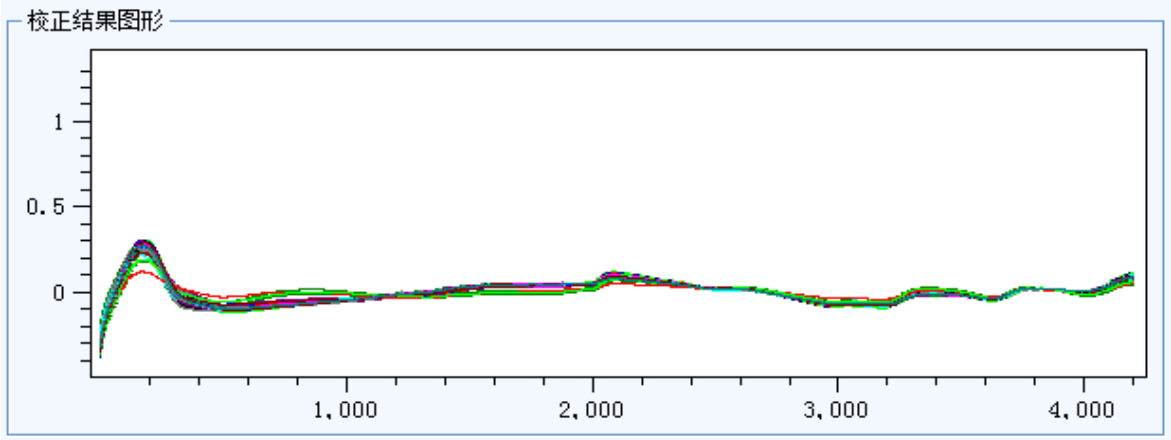
如下三组图形分别为设定多项式阶数从 1 到 3 得到的结果，各组图形中的上下二图则分别为典型近红外光谱原始数据及其去趋势化校正后的结果。




原始数据
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
1,000
2,000
3,000
4,000


校正结果图形
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
1,000
2,000
3,000
4,000


原始数据
1
0.5
0
1,000
2,000
3,000
4,000


校正结果图形
1
0.5
0
1,000
2,000
3,000
4,000

.第十一章　变量选择

变量选择是构建稳健可靠模型的关键步骤之一，变量选择的结果好坏，对模型预测的影响非常明显。

11.1. 整体介绍

变量选择是指从复杂数据中有效提取与分类或回归建模属性相关性高，预测能力强，可解释性好的数据特征，简单来说即是从无信息或干扰信息中挑选出对建模有用的信息(变量)。变量选择多变量数据分析的关键问题之一，也是目前受到广泛关注的研究内容。

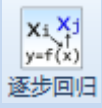
变量选择方法非常之多，而且不同针对数据类型的数据(色谱、质谱和光谱等)，或者解决不同的建模问题(如分类或回归)，所用的方法还不完全相同。以近红外数据的分析为例，使用较多的便包括如下表所列的系列方法。

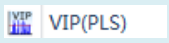
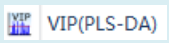
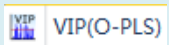
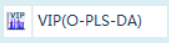
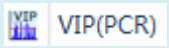
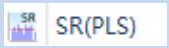
序号	英文简称	英文全称	说明
1	UVE	Uninformative Variable Elimination	见如下详情。
2	MWPLS	Moving Window Partial Least Squares Regression	见如下详情。
3	GA-PLS	Genetic Algorithms – Partial Least Squares	见如下详情。
4	VIP	Variable Importance In Projection	见如下详情。
5	SR	Selectivity Ratio	见如下详情。
6	MLR-step	Stepwise Multiple Linear Regression	见如下详情。
7	GO	Global Optimization	全局优化方法，如遗传算法、

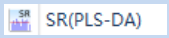
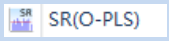
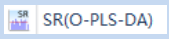
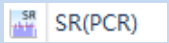
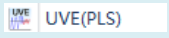
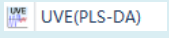
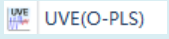
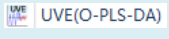
			粒子群优化、模拟退火和蚁群算法等。
8	MPA	Model Population Analysis	基于模型集群分析的一大类方法。

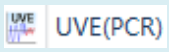
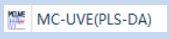
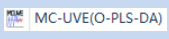
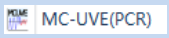
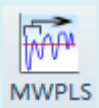
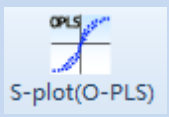
11.1.1. 概述

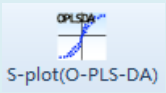
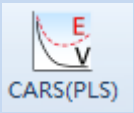
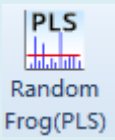
在本软件中，同时包括用于分类和回归的变量选择方法，并将其概括为三个大类，分别为经典方法，常用方法和基于模型集群的方法。通过使用这些方法，足够解决科学研究和实际应用中的变量选择问题。

序号	方法名	图标	说明
1	不加权		用于分类的变量选择方法，基于变量在不同组内与所有样本间的标准偏差选择重要变量，具体内容请参见对该方法的详细介绍。
2	加权		与不加权方法雷同，差异在于对标准偏差进行加权计算。
3	Fisher 比		用于分类的变量选择方法，通过计算变量在组内和组间的变化，构造 Fisher 比值选择重要变量，具体内容请参见对该方法的详细介绍。
4	逐步回归		主要用于回归的变量筛选方法。通过计算增加一个或多个变量后模型残差平方和(SSE)的变化，判断变量的重要性，且增加变量后，须对模型中所有变量重新审查加入的必要性，直到再增加变量对模型 SSE 的影响不再显著为止。显著减少；在前面步骤中增加的自变量在后面的步骤中有可能被剔除，而在前面步骤中剔除的自变量在

			后面的步骤中也可能重新进入到模型中。基于 F 检验判断增减变量对 SSE 变化的显著性。详情请参见对该方法的详细介绍。
5	VIP		
	VIP (PLS)		指数据矩阵中的变量对解释因变量的重要性，本法综合考虑各变量对获得模型得分的贡献，以及得分解释因变量 y 的能力，具体内容请参见对该方法的详细介绍。基于 PLS 法建模，用于回归中的变量选择。
	VIP (PLS-DA)		与上一方法雷同，基于 PLS-DA 法建模，用于分类中的变量选择。
	VIP(O-PLS)		与上一方法雷同，基于 O-PLS 法建模，用于回归中的变量选择。
	VIP(O-PLS-DA)		与上一方法雷同，基于 O-PLS-DA 法建模，用于分类中的变量选择。
	VIP(PCR)		与上一方法雷同，基于 PCR 法建模，用于回归中的变量选择。
6	SR		
	SR (PLS)		本法基于变量被模型解释与未被解释的方差比值构建，其核心思想是变量被解释的越多则越重要，具体内容请参见对该方法的详细介绍。基于 PLS 法建模，用于回归中的变量选择。

	SR(PLS-DA)		与上一方法雷同，基于 PLS-DA 法建模，用于分类中的变量选择。
	SR(O-PLS)		与上一方法雷同，基于 O-PLS 法建模，用于回归中的变量选择。
	SR(O-PLS-DA)		与上一方法雷同，基于 O-PLS-DA 法建模，用于分类中的变量选择。
	SR(PCR)		与上一方法雷同，基于 PCR 法建模，用于回归中的变量选择。
7	UVE		
	UVE (PLS)		加入一个与原始数据相同大小的噪声矩阵，基于交互验证构建模型并获得回归系数，将各变量回归系数与噪声比较，并以其稳定性作为评价指标。具体内容请参见对该方法的详细介绍。基于 PLS 法建模，用于回归中的变量选择。
	UVE(PLS-DA)		与上一方法雷同，基于 PLS-DA 法建模，用于分类中的变量选择。
	UVE(O-PLS)		与上一方法雷同，基于 O-PLS 法建模，用于回归中的变量选择。
	UVE(O-PLS-DA)		与上一方法雷同，基于 O-PLS-DA 法建模，用于分类中的变量选择。

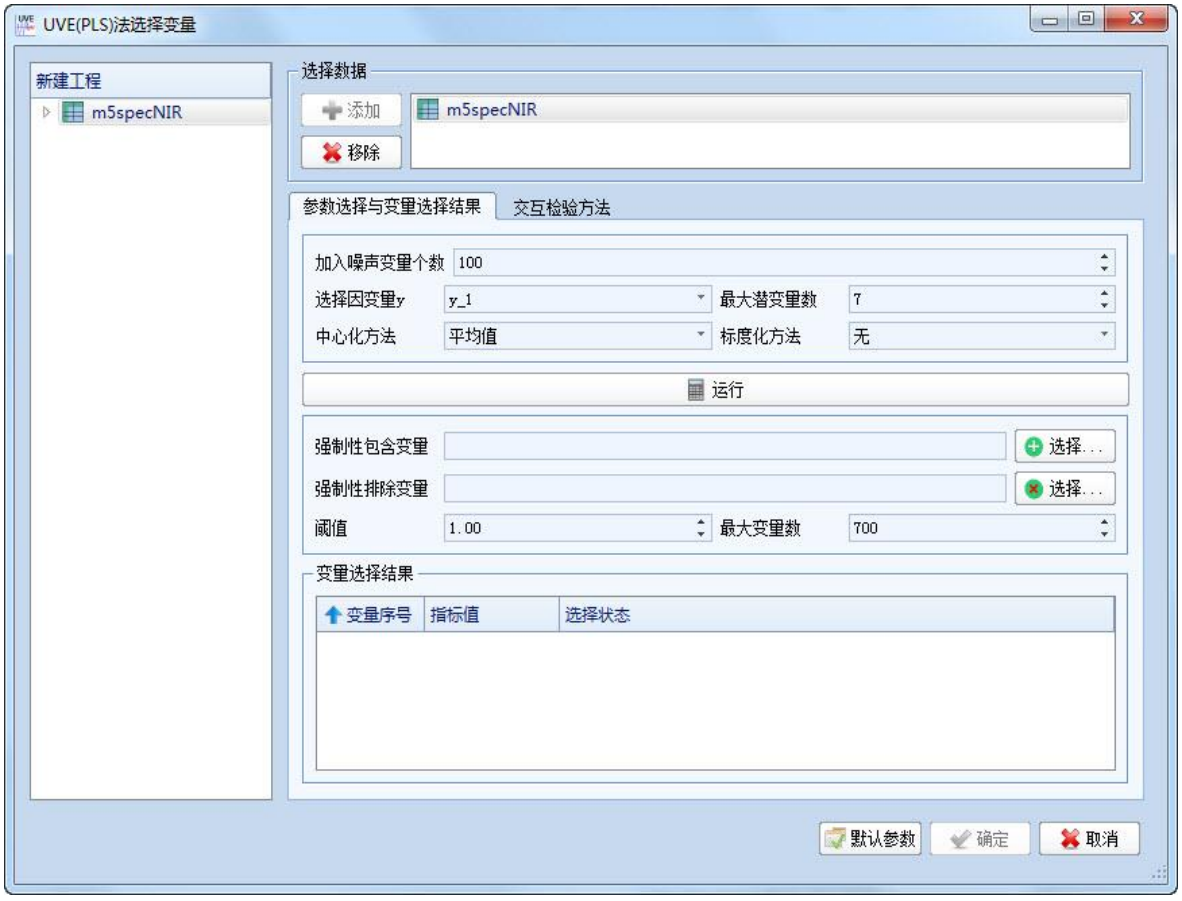
	UVE(PCR)		与上一方法雷同，基于 PCR 法建模，用于回归中的变量选择。
8	MC-UVE		
	MC-UVE(PLS)		与上一方法雷同，但基于蒙特卡罗方法构建模型。具体内容请参见对该方法的详细介绍。基于 PLS 法建模，用于回归中的变量选择。
	MC-UVE(PLS-DA)		与上一方法雷同，基于 PLS-DA 法建模，用于分类中的变量选择。
	MC-UVE(O-PLS)		与上一方法雷同，基于 O-PLS 法建模，用于回归中的变量选择。
	MC-UVE(O-PLS-DA)		与上一方法雷同，基于 O-PLS-DA 法建模，用于分类中的变量选择。
	MC-UVE(PCR)		与上一方法雷同，基于 PCR 法建模，用于回归中的变量选择。
9	MWPLS		用于回归的变量选择方法，以一定尺寸窗口扫描整个数据区域，选取不同区段预测误差较佳的数据，构建最终的模型。基于 PLS 法建模，具体内容请参见对该方法的详细介绍。
10	S-plot(O-PLS)		用于分类的变量选择方法，基于 OPLS 方法，以可视化图形的方式表征变量与类别响应间的协方差与相关性，从图中选择变量，特别适合于代谢组学研究中寻找不同

			类别间具有统计与生物化学显著性的代谢小分子标志物。具体内容请参见对该方法的详细介绍。
11	S-plot(O-PLS-DA)		与上一方法雷同，基于 PLS-DA 法建模，用于分类中的变量选择。
12	CARS(PLS)		基于模型集群分析的思想，以计算预测误差分布的方法全面评价所选择的变量，可获得更优的变量组合。具体内容请参见对该方法的详细介绍。
13	Random Frog(PLS)		基于模型集群分析的思想，结合逆跳马尔科夫蒙特卡罗的思路提出的变量选择方法，特别适合多变量数据分析中获得更优的变量组合。具体内容请参见对该方法的详细介绍。
14	Random Frog(PLS-DA)		与上一方法雷同，基于 PLS-DA 法建模，用于分类中的变量选择。
15	MIA(SVC)		基于模型集群分析的思想，专为支持向量分类分析而发展，是极少数特别适合该类方法的变量选择方法。
16	MIA(SVR)		与上一方法雷同，用于支持向量回归中变量选择。

11.1.2. 通用步骤

通常包括选择数据，设置参数，点击运行开始计算，点击确定获得结果这几个步骤。下面一一进行介绍。

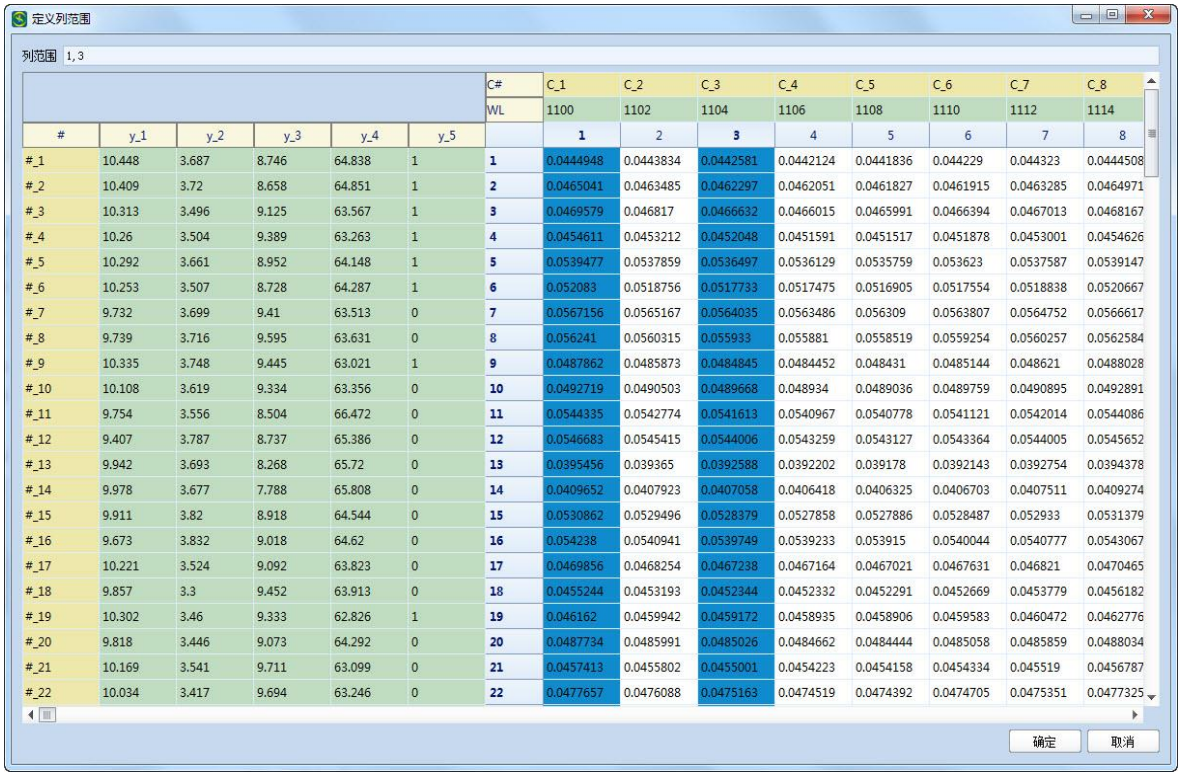
1) 选择数据：与部分雷同，请参考即可，如下图时变量选择的典型界面图形。



2) 设置参数：不同变量选择方法所涉及的参数有所不同，将在介绍各方法时详细说明。

除各方法的不同参数外，本软件亦人性化地提供二个功能：一是用户可强制性包含变量，即不管变量的实际情况如何，用户均可强制性地将任意变量作为“种子变量”予以保留。该功能特别适合针对某些变量，在变量选择前，已可通过其他先验信息判断，该变量需要加入到模型中，比如代谢组学研究中某些解释性好的代谢特征。

上述功能可通过点击图中按钮 选择... 实现。点击该按钮后，出现如下所示的页面，即所选数据界面。用户可通过此界面直接添加强制性选择的变量，Ctrl 键可用。



						C#	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
						WL	1100	1102	1104	1106	1108	1110	1112	1114
#	y_1	y_2	y_3	y_4	y_5		1	2	3	4	5	6	7	8
#_1	10.448	3.687	8.746	64.838	1	1	0.0444948	0.0443834	0.0442581	0.0442124	0.0441836	0.044229	0.044323	0.0444508
#_2	10.409	3.72	8.658	64.851	1	2	0.0465041	0.0463485	0.0462297	0.0462051	0.0461827	0.0461915	0.0463285	0.0464971
#_3	10.313	3.496	9.125	63.567	1	3	0.0469579	0.046817	0.0466632	0.0466015	0.0465991	0.0466394	0.0467013	0.0468167
#_4	10.26	3.504	9.389	63.263	1	4	0.0454611	0.0453212	0.0452048	0.0451591	0.0451517	0.0451878	0.0453001	0.0454626
#_5	10.292	3.661	8.952	64.148	1	5	0.0539477	0.0537859	0.0536497	0.0536129	0.0535759	0.053623	0.0537587	0.0539147
#_6	10.253	3.507	8.728	64.287	1	6	0.052083	0.0518756	0.0517733	0.0517475	0.0516905	0.0517554	0.0518838	0.0520667
#_7	9.732	3.699	9.41	63.513	0	7	0.0567156	0.0565167	0.0564035	0.0563486	0.056309	0.0563807	0.0564752	0.0566617
#_8	9.739	3.716	9.595	63.631	0	8	0.056241	0.0560315	0.055933	0.055881	0.0558519	0.0559254	0.0560257	0.0562584
#_9	10.335	3.748	9.445	63.021	1	9	0.0487862	0.0485873	0.0484845	0.0484452	0.048431	0.0485144	0.048621	0.0488028
#_10	10.108	3.619	9.334	63.356	0	10	0.0492719	0.0490503	0.0489668	0.048934	0.0489036	0.0489759	0.0490895	0.0492891
#_11	9.754	3.556	8.504	66.472	0	11	0.0544335	0.0542774	0.0541613	0.0540967	0.0540778	0.0541121	0.0542014	0.0544086
#_12	9.407	3.787	8.737	65.386	0	12	0.0546683	0.0545415	0.0544006	0.0543259	0.0543127	0.0543364	0.0544005	0.0545652
#_13	9.942	3.693	8.268	65.72	0	13	0.0395456	0.039365	0.0392588	0.0392202	0.039178	0.0392143	0.0392754	0.0394378
#_14	9.978	3.677	7.788	65.808	0	14	0.0409652	0.0407923	0.0407058	0.0406418	0.0406325	0.0406703	0.0407511	0.0409274
#_15	9.911	3.82	8.918	64.544	0	15	0.0530862	0.0529496	0.0528379	0.0527858	0.0527886	0.0528487	0.052933	0.0531379
#_16	9.673	3.832	9.018	64.62	0	16	0.054238	0.0540941	0.0539749	0.0539233	0.053915	0.0540044	0.0540777	0.0543067
#_17	10.221	3.524	9.092	63.823	0	17	0.0469856	0.0468254	0.0467238	0.0467164	0.0467021	0.0467631	0.046821	0.0470465
#_18	9.857	3.3	9.452	63.913	0	18	0.0455244	0.0453193	0.0452344	0.0452332	0.0452291	0.0452669	0.0453779	0.0456182
#_19	10.302	3.46	9.333	62.826	1	19	0.046162	0.0459942	0.0459172	0.0458935	0.0458906	0.0459583	0.0460472	0.0462776
#_20	9.818	3.446	9.073	64.292	0	20	0.0487734	0.0485991	0.0485026	0.0484662	0.0484444	0.0485058	0.0485859	0.0488034
#_21	10.169	3.541	9.711	63.099	0	21	0.0457413	0.0455802	0.0455001	0.0454223	0.0454158	0.0454334	0.045519	0.0456787
#_22	10.034	3.417	9.694	63.246	0	22	0.0477657	0.0476088	0.0475163	0.0474519	0.0474392	0.0474705	0.0475351	0.0477325

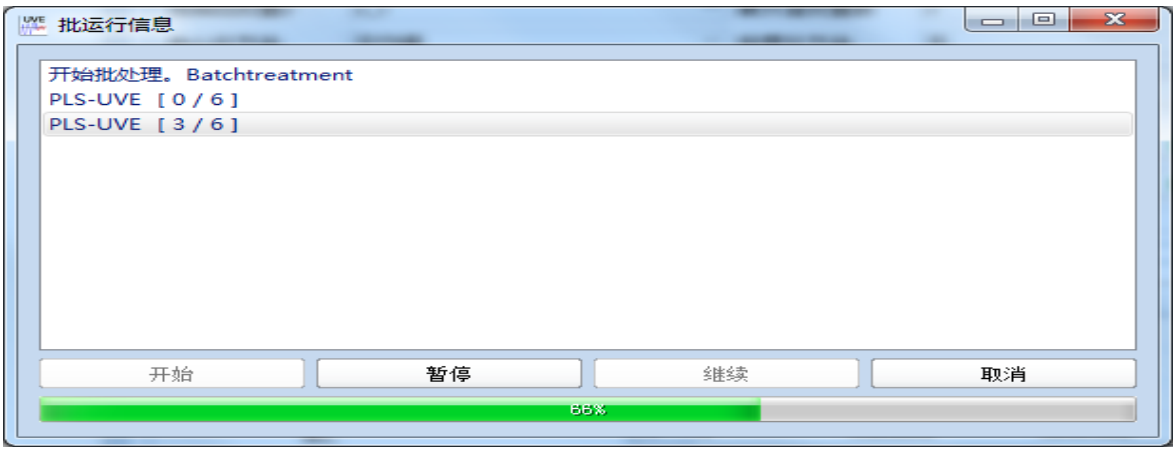
另一功能则与强制性包括某些变量相反，用户亦可强制性排除某些变量，通过点击页面中的按钮 

达到。其使用与前一功能雷同。用户亦可直接输入强制包括或排除的变量序号，可达到同样效果。除此之外，还包括选择变量的方法阈值，以及变量总数二个参数，概括为如下表。

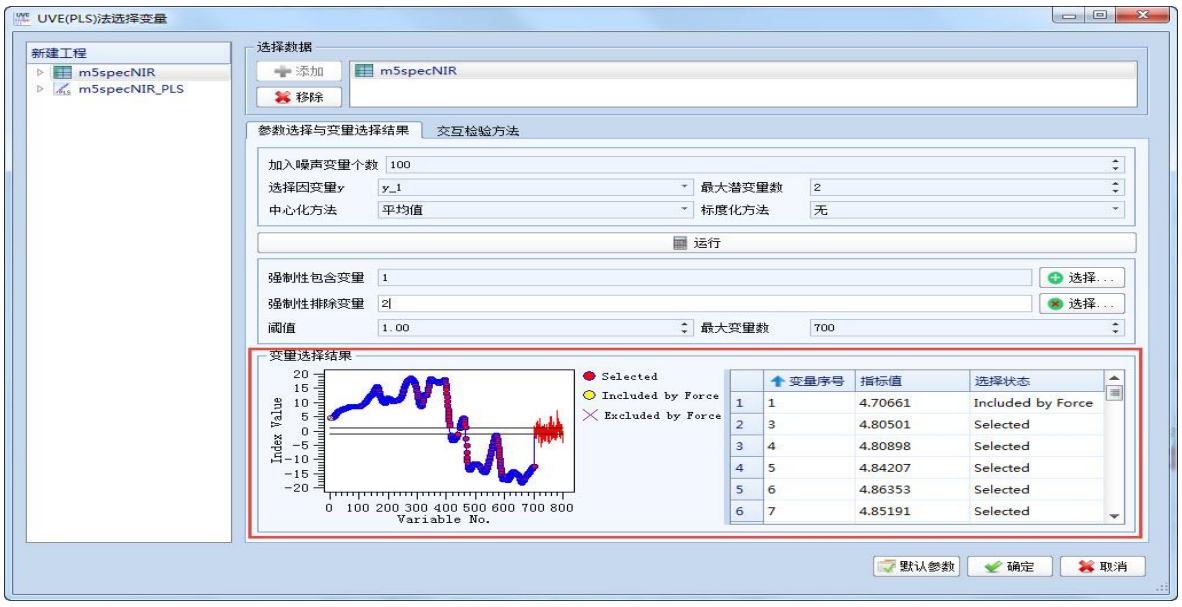
参数	范围	说明
强制性包含变量	无。	无论这个变量是好是坏，均被被选择。
强制性排除变量	无。	无论这个变量是好是坏，均不被选择。
阈值	见各变量选择方法的参数说明。	见各具体变量选择方法的参数说明。
最大变量数	[1 num]，其中 num 为所选数据的长度。	最大变量选择数。

特别需要说明的是，改变界面中运算按钮上方的参数后，一般是不会立即重新计算获得结果的。但改变表中的参数，则将立即自动重新计算，并更新结果显示。

3) 点击运行开始计算：选择合适参数后，点击按钮 运行 便开始运行，并弹出如下运行对话框:



运行成功后，变量选择结果区域图形和表格结果，如下图所示(以 UVE 方法为例，但部分方法有所不同):



上图结果表格中对变量选择状态的描述，如下表所示:

选择状态	含义	说明
Selected	该变量被选中。	被选中的变量。
Included by Force	该变量被强制性包含。	无论该变量好坏，均被选择。
Excluded by Force	该变量被强制性排除。	无论该变量好坏，均被排除。

4) 点击确定获得结果：用户认可变量选择结果后，即可点击按钮  获得结果。若计算成功，则该结果将先是在工程导航栏中，并作为新的数据节点，如下图所示；若计算失败，则提示用户计算失败。

此外点击按钮  ，则界面上的参数值改变为默认值；点击按钮  ，取消操作，关闭对话框。

接下来依次介绍各变量选择方法的使用。

11.2. 不加权法

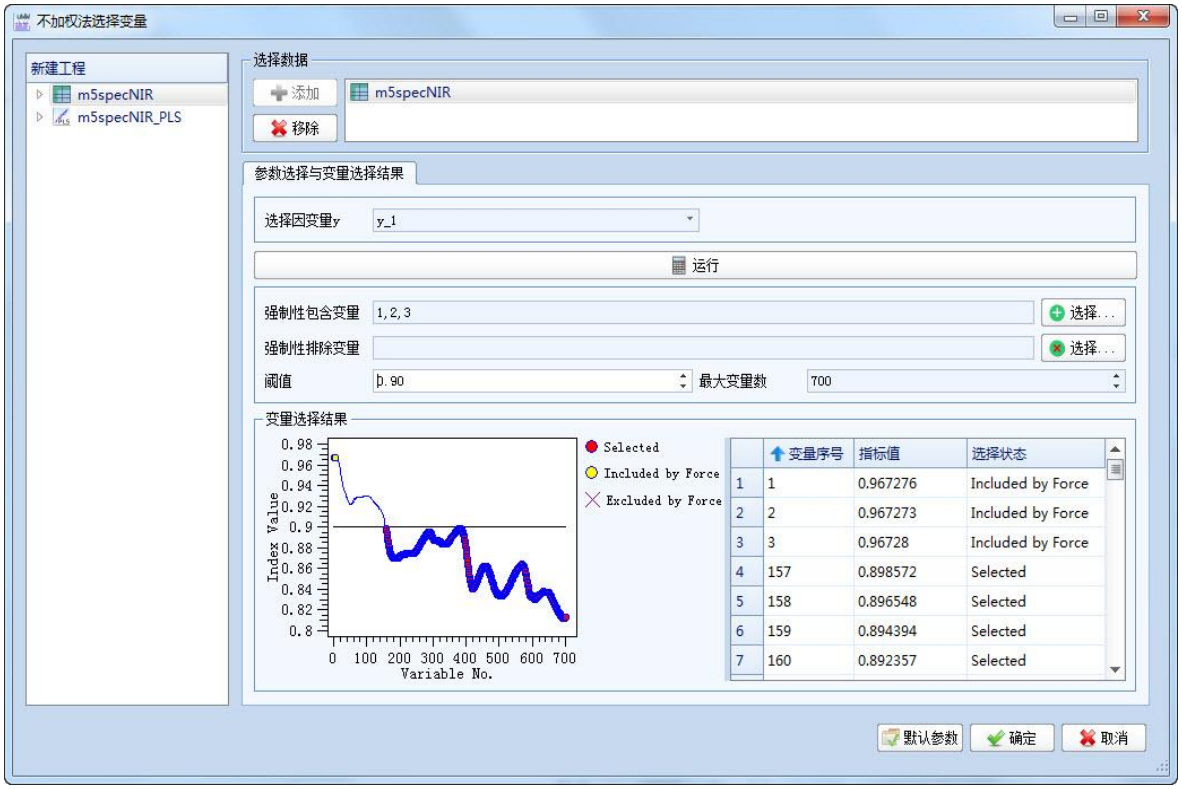
本法计算各类样本中变量的标准偏差之均值与所有样本中变量标准偏差的比值得到，具体以如下公式计算。

$$\mathrm{un\_}w(i)=\frac{\{\mathrm{average}[\mathrm{std}(i_1),\mathrm{std}(i_2),\ldots,\mathrm{std}(i_n)]\}}{\mathrm{std}(i_{\mathrm{all}})}$$

其中 std(i_n)和 std(i_{all})分别为第 i 个样本在第 n 类和所有样本中的标准偏差。用户通过设定选取合适变量的 un_*w* 阈值完成变量选择。显然，un_*w* 值越大，越需要被选择。

操作步骤:

步骤 1: 点击**变量选择** -> **不加权**，弹出如下对话框:

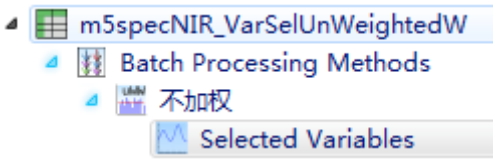


接下来的操作步骤参照变量选择之通用步骤。

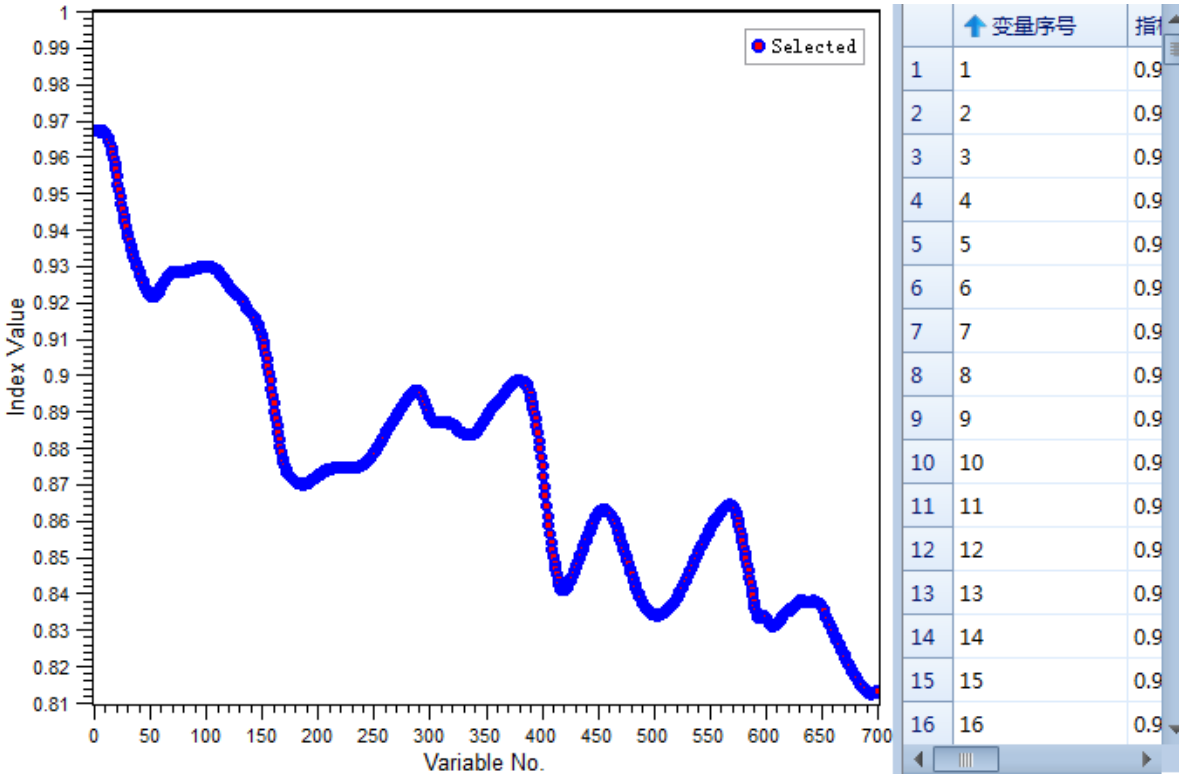
参数说明见下表:

参数	范围	说明
选择因变量 y	根据数据实际情况选择。	无类别信息则无法计算。
阈值	[0 ∞]，其中∞为变量选择的阈值。	被选变量的 Index Value 值不能大于该阈值。

变量选择完成后，将在工程导航栏中产生新的节点，示例数据的变量选择结果如下图所示。



Selected Variables 节点下所得到的结果如下图所示，图中结果的解释请参见 11.1.2.。



	变量序号	指
1	1	0.9
2	2	0.9
3	3	0.9
4	4	0.9
5	5	0.9
6	6	0.9
7	7	0.9
8	8	0.9
9	9	0.9
10	10	0.9
11	11	0.9
12	12	0.9
13	13	0.9
14	14	0.9
15	15	0.9
16	16	0.9

11.3. 加权法

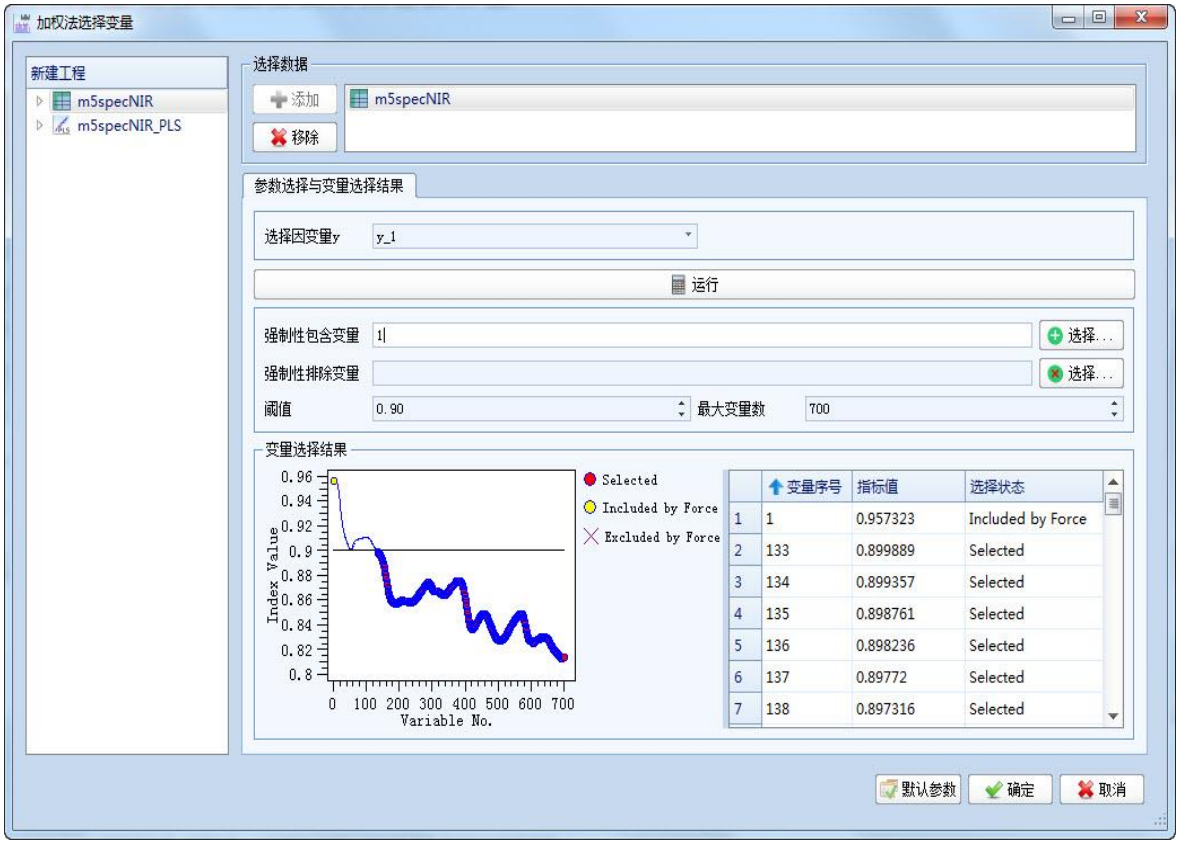
本法与上一方法对应，其差别在于在上一方法的基础上，加入权重计算变量选择指标，可以如下方程表示。

$$w(i) = \frac{\{[\mathrm{std}(i_1) \times \mathrm{g}_1 + \mathrm{std}(i_2) \times \mathrm{g}_2 + \cdots + \mathrm{std}(i_n) \times \mathrm{g}_n]\}}{[\mathrm{std}(i_{\mathrm{all}}) \times \mathrm{g}_{\mathrm{all}}]}$$

其中，g_n 和 g_{all} 分别为第 n 类和所有样本的数量。同样，用户通过设定选取合适变量的 *w* 阈值完成变量选择。显然，*w* 值越大，越需要被选择。

操作步骤:

步骤 1: 点击**变量选择** -> **加权**，弹出如下对话框:



接下来的操作步骤参照变量选择之通用步骤。

本法中所涉及的变量意义，与表中擂台，在此不再赘述。变量选择完成后，将在工程导航栏中产生新的节点，示例数据的变量选择结果与 11.2.方法雷同，不再赘述。

11.4. Fisher 比法

本法通过计算各变量在类间与类内的变化比值而得到，如下式所示。

$$\mathrm{F}(i) = \frac{\{[1/(\mathrm{g}-1)]\mathrm{SSB}(i)\}}{\{[1/(n-\mathrm{g})]\mathrm{SSW}(i)\}}$$

其中 SSB 和 SSW 项分别由如下二式得到。

$$\mathrm{SSB}(i) = \sum_{k=1}^{\mathrm{g}} n_k [i\_\mathrm{average}(y_k) - i\_\mathrm{average}(y)]^2$$

$$(k = 1,2,\ldots,\mathrm{g})$$

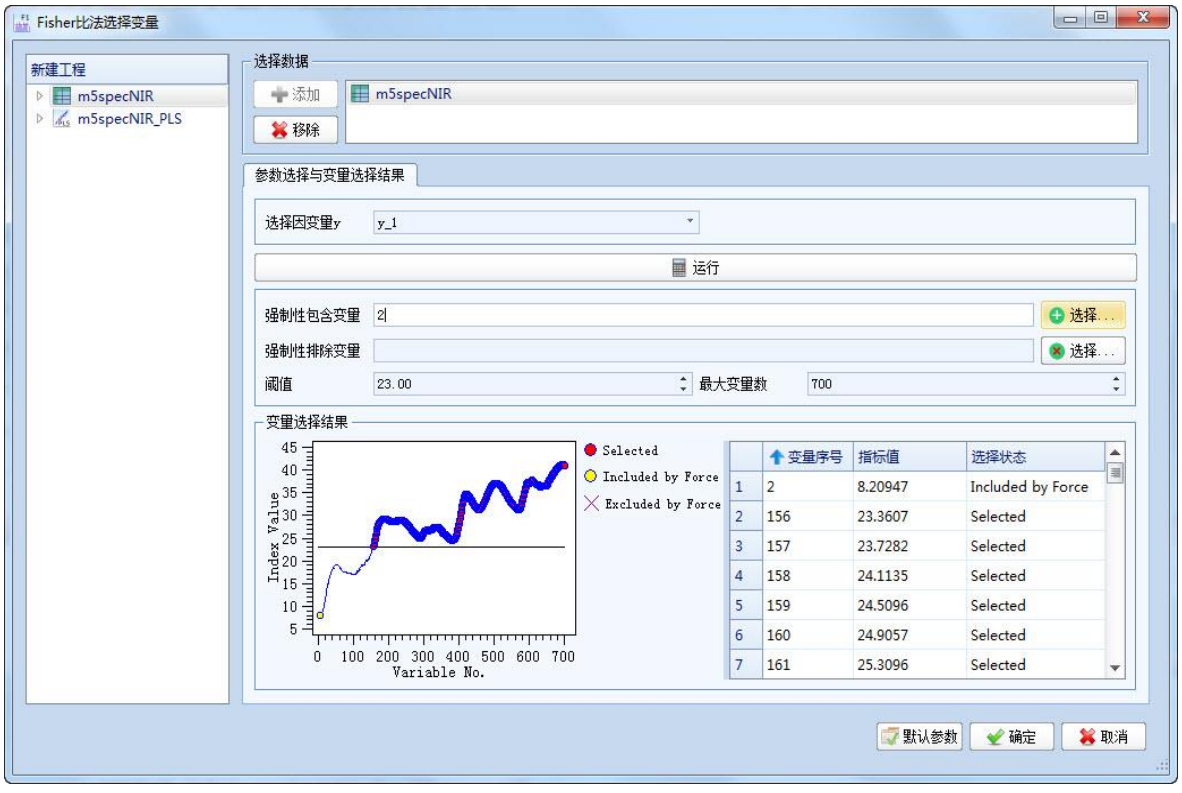
$$\mathrm{SSW}(i)=\sum_{k=1}^{\mathrm{g}}\sum_{j=1}^{n_k}\left[y_{kj}-i\_\mathrm{average}(y_k)\right]^2$$

$$(k=1,2,\ldots,\mathrm{g},j=1,2,\ldots,n_k)$$

其中 g 和 n_k 分别样本类别数，以及 k 类中的样本数；而 i_average(y_k)，i_average(y)和 y_{kj} 则分别为第 k 组中第 i 个变量的均值，第 i 个样本在所有样本中的总均值，以及第 k 组中第 j 个样本的值。用户通过设定选取合适变量的 F 阈值完成变量选择。显然，F 值越大，越需要被选择。

操作步骤:

步骤 1: 点击**变量选择** -> **Fisher 比**，弹出如下对话框:



接下来的操作步骤参照变量选择之通用步骤。本法中所涉及的变量意义，与表中擂台，在此不再赘述。变量选择完成后，将在工程导航栏中产生新的节点，示例数据的变量选择结果与 11.2.方法雷同，不再赘述。

11.5. 逐步回归法

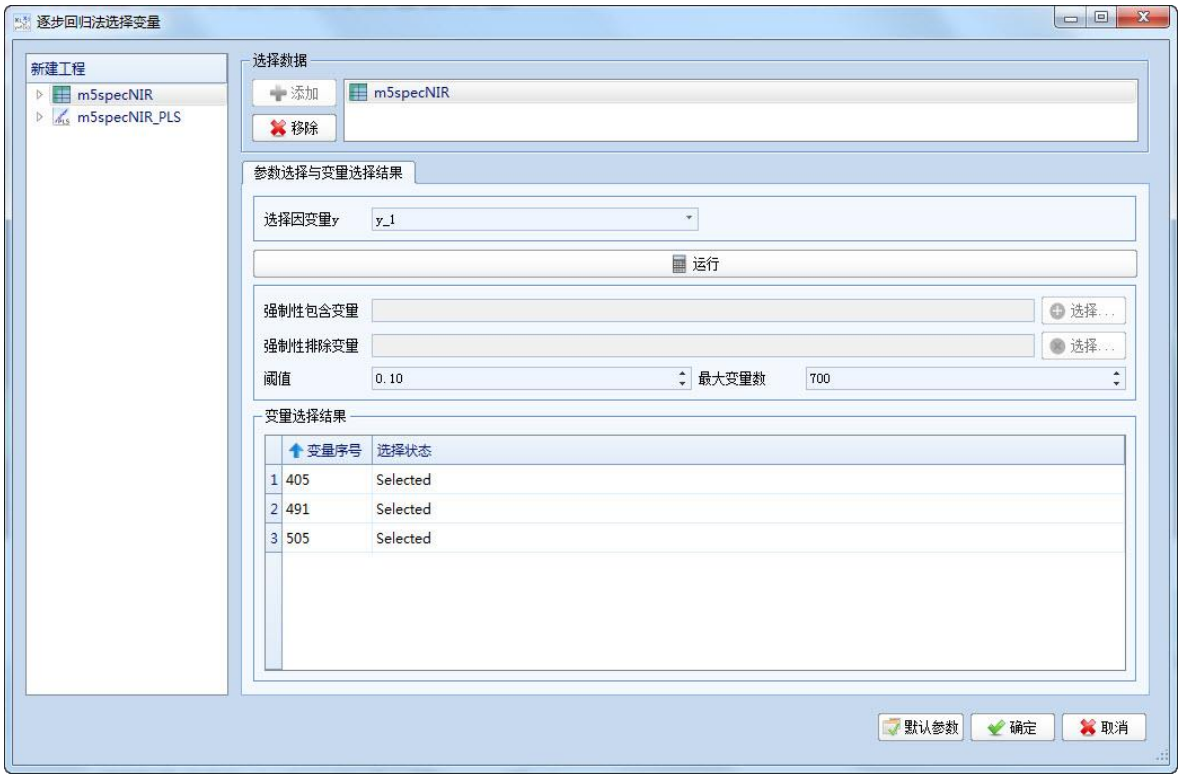
逐步回归是非常经典的方法之一，亦被发展为变量选择方法。本法选择变量建立在回归统计量进行显著性检验的基础上。将各变量分别引入到回归模型中，判断新变量是否使得模型残差平方和显著变化。若新变量使得该值显著减少(使用 F 检验判断)，则必须将该变量加入到模型中，否则就无需引入。

特别需要指明的是，没加入一个新的变量，均需对模型中已经加入的变量重新进入考察，并在必要时剔除新模型中的变量。依次方法增加或剔除变量，直至再往模型中加入新的变量，对模型残差平方和无显著变化为止。

注意本法与上变量选择方法有所不同，用户无需设置变量选择的阈值，程序将直接给出被选变量。

操作步骤:

步骤 1: 点击**变量选择** -> **逐步回归**，弹出如下对话框:



从图中可以看出，通过本法获得的变量选择结果显示无变量指标值。接下来的操作步骤参照变量选择之通用步骤。

变量选择完成后，将在工程导航栏中产生新的节点，示例数据的变量选择结果与 11.2.方法雷同，不再赘述。

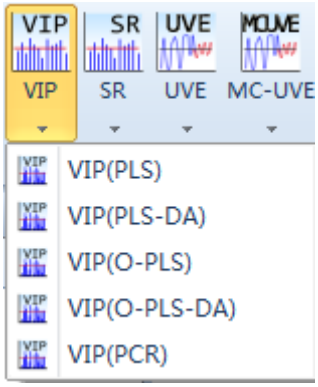
11.6. VIP 法

本法是重要的变量选择方法，得到了非常广泛的关注和使用。其核心思想是模型得分体现通过变量解释响应 **y** 的能力，若某变量对应得分的解释能力强，而且其在构造模型时亦有显著贡献，则表示该变量非常重要，需要被选择。VIP 指标值的计算如下式所示。

$$VIP_j = \sqrt{\frac{p\sum_{k=1}^{h}(\hat{c}_k^2 t'_k t_k)(w_{jk})^2}{\sum_{k=1}^{h}\hat{c}_k^2 t'_k t_k}}$$

各参数的计算均来自于偏最小二乘方法，其意义详见方法介绍。用户通过设定选取合适变量的 VIP 阈值完成变量选择。显然，VIP 值越大，越可能被选择，通常以指标值达到 1 作为引入该特征的依据。

在本软件中，VIP 法实际上是一个变量选择的策略，针对不同的分类和回归问题，均有针对性的 VIP 方法，如下图所示。

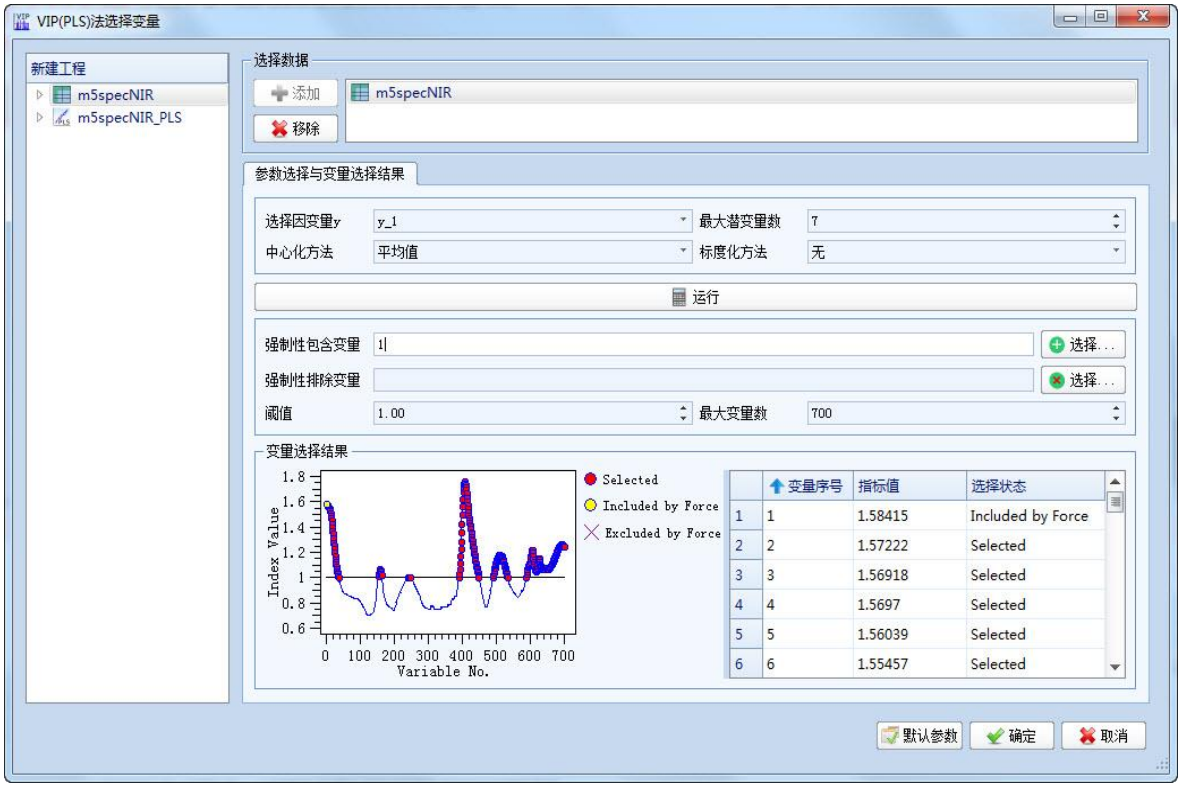


11.6.1. VIP(PLS)

本方法基于 PLS 回归构建，适合于基于该方法的变量选择。

操作步骤:

步骤 1: 点击**变量选择** -> **VIP** -> **VIP(PLS)**，弹出如下对话框:



接下来的操作步骤参照变量选择之通用步骤。

本法中所涉及参数的意义，如下表所示。

参数	范围	说明
选择因变量 y	根据数据实际情况选择。	无类别信息则无法计算。
最大潜变量数	[1 min(row, col)-1]，其中 min(row, col)表示所选数据的行数和列数中的较小值。	模型构建的关键参数，指加入模型的最大潜变量数目。

中心化方法	以下四种选其一: 1.无; 2.平均值; 3.中位数; 4.最小值。	请参见变量标度化章节。
标度化方法	以下四种选其一: 1.无; 2.标准偏差; 3.标准偏差开方; 4.四分位距(IQR)。	请参见变量标度化章节。
阈值	[0 ∞]，其中∞为变量选择的阈值。	被选变量的 Index Value 值不能大于该阈值。

变量选择完成后，将在工程导航栏中产生新的节点，示例数据的变量选择结果与 11.2.方法雷同，不再赘述。

11.6.2. VIP(PLS-DA)

本方法基于 PLS-DA 分类构建，适合于基于该方法的变量选择。

具体操作方法请参考，在此不再赘述。

11.6.3. VIP(O-PLS)

本方法基于 OPLS 回归构建，适合于基于该方法的变量选择。

具体操作方法请参考，在此不再赘述。

11.6.4. VIP(O-PLS-DA)

本方法基于 O-PLS-DA 分类构建，适合于基于该方法的变量选择。

具体操作方法请参考，在此不再赘述。

11.6.5. VIP(PCR)

本方法基于 PCR 回归构建，适合于基于该方法的变量选择。

具体操作方法请参考，在此不再赘述。

11.7. SR 法

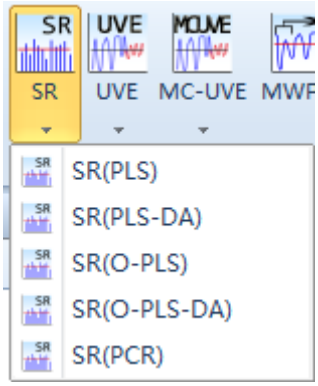
该法通过计算各变量被解释方差与残差方差比值所构造的重要性评价指标进行变量选择，显然该值越大越应该被选择。SR 值以如下方式计算。

$$\mathrm{SR}_i = v_{\mathrm{exp},i}/v_{\mathrm{res},i} \quad i = 1,2,\ldots,M$$

$v_{exp,i}$ 和 $v_{res,i}$ 分别为建模后各变量被模型所解释的方差和未被解释的方差。

更详细的介绍请参见方法介绍章节。

用户通过设定选取合适变量的 SR 阈值完成变量选择。在本软件中，SR 法实际上是一个变量选择的策略，针对不同的分类和回归问题，均有针对性的 SR 方法，如下图所示。



11.7.1. SR(PLS)

本方法基于 PLS 回归构建，适合于基于该方法的变量选择。

操作步骤:

步骤 1: 点击**变量选择** -> **SR** -> **SR(PLS)**，弹出如下对话框:

接下来的操作步骤参照变量选择之通用步骤。

本法中所涉及参数的意义，如下表所示。

参数	范围	说明
选择因变量 y	根据数据实际情况选择。	无类别信息则无法计算。
最大潜变量数	[1 min(row, col)-1]，其中 min(row, col)表示所选数据的行数和列数中的较小值。	模型构建的关键参数，指加入模型的最大潜变量数目。
中心化方法	以下四种选其一: 1.无；2.平均值；3.中位数；4.最小值。	请参见变量标度化章节。
标度化方法	以下四种选其一: 1.无；2.标准偏差；3.标准偏差开方；4.四分位距(IQR)。	请参见变量标度化章节。

阈值	[0 ∞]，其中∞为变量选择的阈值。	被选变量的 Index Value 值不能大于该阈值。

变量选择完成后，将在工程导航栏中产生新的节点，示例数据的变量选择结果与 11.2.方法雷同，不再赘述。

11.7.2. SR(PLS-DA)

本方法基于 PLS-DA 分类构建，适合于基于该方法的变量选择。

具体操作方法请参考，在此不再赘述。

11.7.3. SR(O-PLS)

本方法基于 OPLS 回归构建，适合于基于该方法的变量选择。

具体操作方法请参考，在此不再赘述。

11.7.4. SR(O-PLS-DA)

本方法基于 O-PLS-DA 分类构建，适合于基于该方法的变量选择。

具体操作方法请参考，在此不再赘述。

11.7.5. SR(PCR)

本方法基于 PCR 回归构建，适合于基于该方法的变量选择。

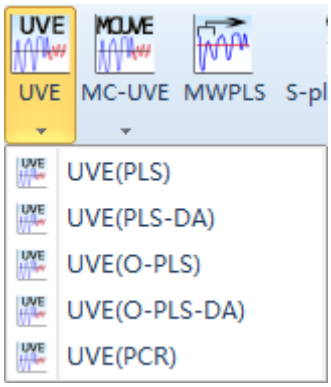
具体操作方法请参考，在此不再赘述。

11.8. UVE 法

本法通过往原始矩阵中加入一个相同大小的噪声矩阵，基于交互验证构建模型并获得回归系数，将各变量回归系数与噪声比较，并以其稳定性作为评价指标。该法同时考虑回归系数的绝对值及其方差，使用中综合利用数据矩阵，响应向量以及噪声信息，得到非常广泛

的使用和认可。具体内容请参见对该方法的详细介绍。

在本软件中，UVE 法实际上是一个变量选择的策略，针对不同的分类和回归问题，均有针对性的 UVE 方法，如下图所示。



11.8.1. UVE(PLS)

本方法基于 PLS 回归构建，适合于基于该方法的变量选择。

操作步骤:

步骤 1: 点击**变量选择** -> **UVE** -> **UVE(PLS)**，弹出如下对话框:

接下来的操作步骤参照变量选择之通用步骤。

本法中所涉及参数的意义，如下表所示。

参数	范围	说明
选择因变量 y	根据数据实际情况选择。	无类别信息则无法计算。
最大潜变量数	[1 min(row, col)-1]，其中 min(row, col)表示所选数据的行数和列数中的较小值。	模型构建的关键参数，指加入模型的最大潜变量数目。
中心化方法	以下四种选其一: 1.无；2.平均值；3.中位数；4.最小值。	请参见变量标度化章节。
标度化方法	以下四种选其一: 1.无；2.标准偏差；3.标准偏差开方；4.四分位距(IQR)。	请参见变量标度化章节。
阈值	[0 ∞]，其中∞为变量选择的阈值。	被选变量的 Index Value 值不能大于该阈值。

变量选择完成后，将在工程导航栏中产生新的节点，示例数据的变量选择结果与 11.2.方法雷同，不再赘述。

11.8.2. UVE(PLS-DA)

本方法基于 PLS-DA 分类构建，适合于基于该方法的变量选择。

具体操作方法请参考，在此不再赘述。

11.8.3. UVE(O-PLS)

本方法基于 OPLS 回归构建，适合于基于该方法的变量选择。

具体操作方法请参考，在此不再赘述。

11.8.4. UVE(O-PLS-DA)

本方法基于 O-PLS-DA 分类构建，适合于基于该方法的变量选择。

具体操作方法请参考，在此不再赘述。

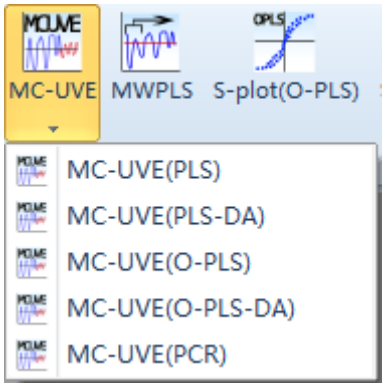
11.8.5. UVE(PCR)

本方法基于 PCR 回归构建，适合于基于该方法的变量选择。

具体操作方法请参考，在此不再赘述。

11.9. MC-UVE 法

本法与上一方法雷同，差异在于同时使用蒙特卡罗方法建模，具体使用不再赘述。本软件针对不同分类和回归情形的 MC-UVE 方法，如下图所示。



11.9.1. MC-UVE(PLS)

本方法基于 PLS 回归构建，适合于基于该方法的变量选择。

操作步骤:

步骤 1: 点击**变量选择** -> **MC-UVE** -> **MC-UVE(PLS)**，弹出如下对话框:

接下来的操作步骤参照变量选择之通用步骤。

本法中所涉及参数的意义，如下表所示。

参数	范围	说明
选择因变量 y	根据数据实际情况选择。	无类别信息则无法计算。
最大潜变量数	[1 min(row, col)-1]，其中 min(row, col)表示所选数据的行数和列数中的较小值。	模型构建的关键参数，指加入模型的最大潜变量数目。
中心化方法	以下四种选其一: 1.无；2.平均值；3.中位数；4.最小值。	请参见变量标度化章节。
标度化方法	以下四种选其一: 1.无；2.标准偏差；3.标准偏差开方；4.四分位距(IQR)。	请参见变量标度化章节。

阈值	[0 ∞]，其中∞为变量选择的阈值。	被选变量的 Index Value 值不能大于该阈值。

变量选择完成后，将在工程导航栏中产生新的节点，示例数据的变量选择结果与 11.2.方法雷同，不再赘述。

11.9.2. MC-UVE(PLS-DA)

本方法基于 PLS-DA 分类构建，适合于基于该方法的变量选择。

具体操作方法请参考，在此不再赘述。

11.9.3. MC-UVE(O-PLS)

本方法基于 OPLS 回归构建，适合于基于该方法的变量选择。

具体操作方法请参考，在此不再赘述。

11.9.4. MC-UVE(O-PLS-DA)

本方法基于 O-PLS-DA 分类构建，适合于基于该方法的变量选择。

具体操作方法请参考，在此不再赘述。

11.9.5. MC-UVE(PCR)

本方法基于 PCR 回归构建，适合于基于该方法的变量选择。

具体操作方法请参考，在此不再赘述。

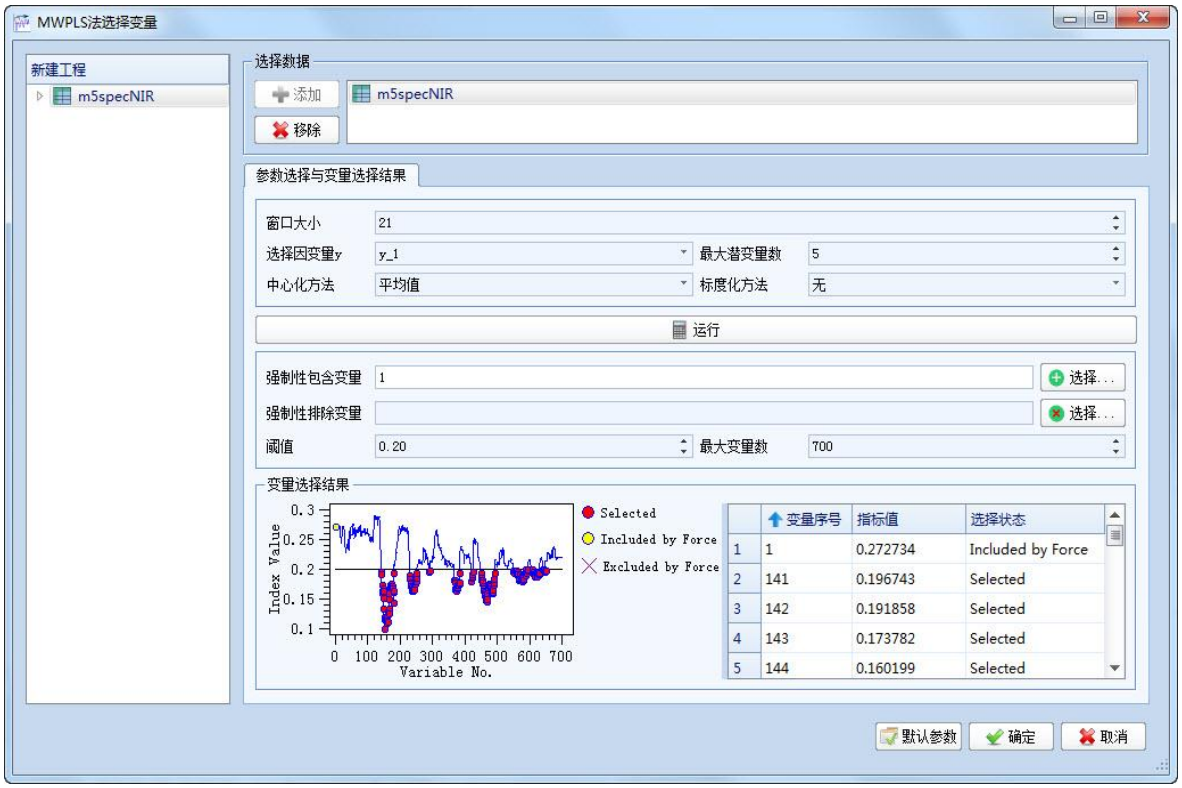
11.10. MWPLS 法

本法在光谱(如近红外)数据的分析中可获得良好的结果，其核心思想是若某光谱是信息含量高的有用信息，可帮助建立稳健可靠的模型，则该附近的光谱亦应该具有相同或相近的性质。因此，可通过移动窗口搜索的方式建立一系列的子模型，并构建组合多个子窗口光

谱数据的方法便可获得优化的光谱数据，用于模型建立。具体内容请参见对该方法的详细介绍。

操作步骤:

步骤 1: 点击**变量选择** -> **MWPLS**，弹出如下对话框:



接下来的操作步骤参照变量选择之通用步骤。

本法中所涉及参数的意义，如下表所示。

参数	范围	说明
选择因变量 y	根据数据实际情况选择。	无类别信息则无法计算。
最大潜变量数	[1 min(row, col)-1]，其中 min(row, col)表示所选数据的行数和列数中的较小值。	模型构建的关键参数，指加入模型的最大潜变量数目。

中心化方法	以下四种选其一: 1.无; 2.平均值; 3.中位数; 4.最小值。	请参见变量标度化章节。
标度化方法	以下四种选其一: 1.无; 2.标准偏差; 3.标准偏差开方; 4.四分位距(IQR)。	请参见变量标度化章节。
阈值	[0 ∞]，其中∞为变量选择的阈值。	被选变量的 Index Value 值不能大于该阈值。

变量选择完成后，将在工程导航栏中产生新的节点，示例数据的变量选择结果与 11.2.方法雷同，不再赘述。

11.11. S-plot(O-PLS)法

S-plot 法以 S 形可视化图形的方式(数据矩阵 X 有强度变化)，综合表征变量与类别响应间的协方差与模型相关性。其 X 轴的值依下式计算，描述数据矩阵中各变量值的大小变化。

$$Cov(t_1, \mathrm{X}) = \frac{\mathrm{t}_1^{\mathrm{T}} \times \mathbf{X}}{N-1} = p[1]$$

在上式中，t_1 和 X 分别为模型第一得分，以及原始矩阵的中心化数据。

Y 轴的值则依下式计算，表征数据矩阵中各变量的可靠性，其值大小从-1 到+1。

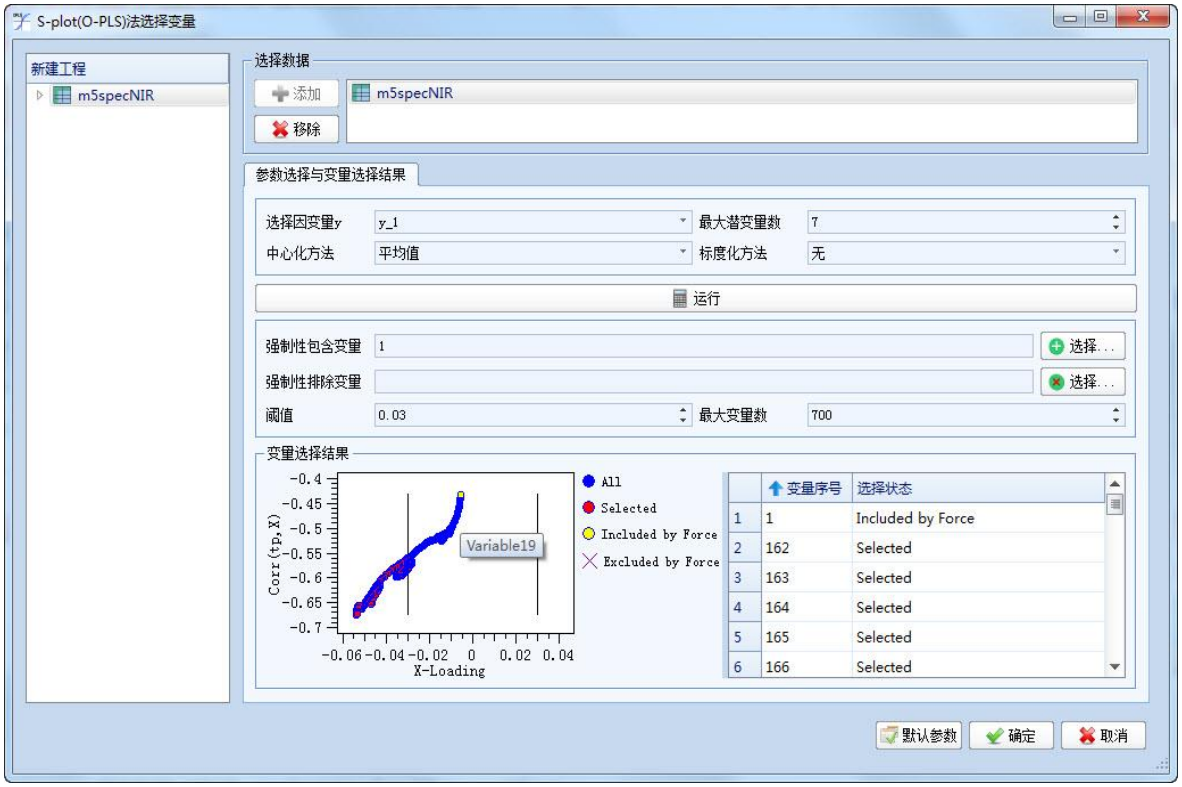
$$Corr(t_1, \mathrm{X}) = \frac{Cov(\mathrm{t}_1, \mathrm{X})}{\sigma_{\mathrm{t}1}\sigma_{\mathrm{x}}} = \frac{p[1]}{\sigma_{\mathrm{t}1}\sigma_{\mathrm{x}}} = p(corr)[1]$$

其中，σ_{t1} 和 σ_X 分别为 t_1 和数据矩阵 X 中各变量的标准偏差。

具体内容请参见对该方法的详细介绍。

操作步骤:

步骤 1: 点击**变量选择** -> **S-plot(O-PLS)**，弹出如下对话框:



本法中的变量选择并非基于指标阈值来决定，而是从图形中可视化选择关键变量。用户可通过鼠标点击图中的点，即可显示该点的变量序号，以方便用户选择。当然，通过该法选择变量，其结果中无“指标值”这一列。接下来的操作步骤参照变量选择之通用步骤。

本法中所涉及参数的意义，如下表所示。

参数	范围	说明
选择因变量 y	根据数据实际情况选择。	无类别信息则无法计算。
最大潜变量数	[1 min(row, col)-1]，其中 min(row, col)表示所选数据的行数和列数中的较小值。	模型构建的关键参数，指加入模型的最大潜变量数目。
中心化方法	以下四种选其一：1.无；2.平均值；3.中位数；4.最小值。	请参见变量标度化章节。

标度化方法	以下四种选其一: 1.无；2.标准偏差；3.标准偏差开方；4.四分位距(IQR)。	请参见变量标度化章节。
阈值	[0 ∞]，其中∞为变量选择的阈值。	被选变量的 Index Value 值不能大于该阈值。

变量选择完成后，将在工程导航栏中产生新的节点，示例数据的变量选择结果与 11.2.方法雷同，不再赘述。

11.12. S-plot(O-PLS-DA)法

本方法基于 O-PLS-DA 分类构建，适合于基于该方法的变量选择。

具体操作方法请参考，在此不再赘述。

11.13. CARS(PLS)法

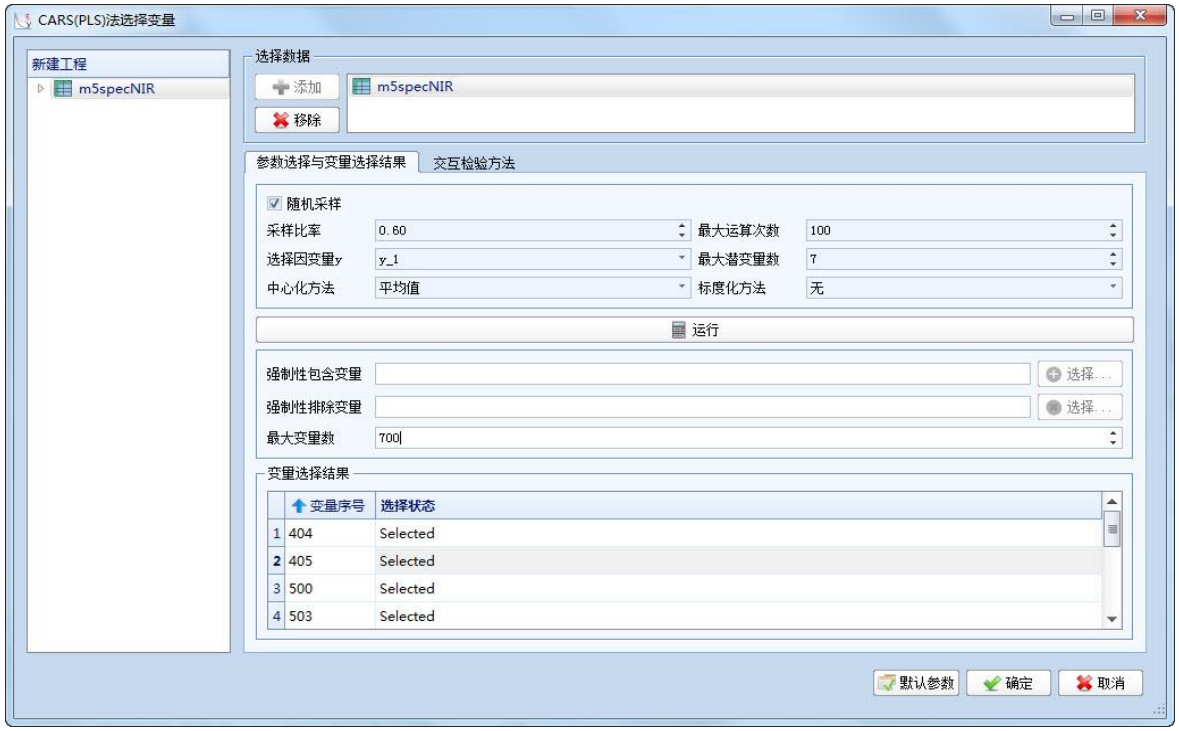
本法建立在模型集群思想的基础上，先基于所谓“适者生存”的原则选择最优变量子集，再随机划分样本构建模型，计算变量子集的交互检验误差，最后以其平均值最小的变量子集作为输出结果。

具体来说，本法由如下几步构成。首先随机选取一定比例的样本构建训练集并建模，以回归系数计算权重与变量的保留比例； 基于保留比与以自适应加权采样选取变量，计算交互检验误差；循环上述步骤达到预先设定的迭代次数，并记录每次迭代得到的结果与分布；最后选择交互检验误差值的平均值最小作为最终结果。

具体内容请参见对该方法的详细介绍。

操作步骤:

步骤 1: 点击**变量选择** -> **CARS(PLS)**，弹出如下对话框:



本法涉及交互检验相关的内容，详情请参见。此外，本法的变量选择结果亦无“指标值”项目。接下来的操作步骤参照变量选择之通用步骤。

参数说明见下表:

参数	范围	说明
采样比率	[0.1 1]	子模型构建采样变量。
最大运算次数	[10 ∞]，其中∞表示无穷大	子模型构建最大次数。
选择因变量 y	根据数据实际情况选择。	无类别信息则无法计算。
最大潜变量数	[1 min(row, col)-1]，其中 min(row, col)表示所选数据的行数和列数中的较小值。	模型构建的关键参数，指加入模型的最大潜变量数目。
中心化方法	以下四种选其一: 1.无；2.平均值；3.中位数；4.最小值。	请参见变量标度化章节。

标度化方法	以下四种选其一: 1.无；2.标准偏差；3.标准偏差开方；4.四分位距(IQR)。	请参见变量标度化章节。
	[0 ∞]，其中∞为变量选择的阈值。	被选变量的 Index Value 值不能大于该阈值。

变量选择完成后，将在工程导航栏中产生新的节点，示例数据的变量选择结果与 11.2.方法雷同，不再赘述。

11.14. Random Frog(PLS)法

本法基于序贯方法获得不同子模型，计算各变量的选择频率，以此评价变量的重要性。即先采用改进的逆跳马尔科夫链进行模型采样，得到多个模型。统计分析各变量在模型中出现的概率，作为选择重要变量的指标和依据。结果表明基于正态分布随机维数转换机制的逆跳马尔科夫改进方法，计算速度快，且模型的预测能力亦得到改善。具体内容请参见对该方法的详细介绍。

操作步骤:

步骤 1: 点击**变量选择** -> **Random Frog(PLS)**，弹出如下对话框:

本法涉及交互检验相关的内容，详情请参见 12.1.2.2.。此外，本法的变量选择结果亦无“指标值”项目。接下来的操作步骤参照变量选择之通用步骤。

本法中所涉及参数的意义，如下表所示。

参数	范围	说明
运算次数	[10 ∞]，其中∞表示无穷大。	子模型构建运算次数。
初始化变量个数	[2 col]，其中 col 表示所选数据的长度。	程序运行预设初始变量数目。
评价指标	以下两种选其一: 1.回归系数 2.SR 值。	获得变量的评价指标。
选择因变量 y	根据数据实际情况选择。	无类别信息则无法计算。
最大潜变量数	[1 min(row, col)-1]，其中 min(row, col)表示所选数据的行数和列数中的较小值。	模型构建的关键参数，指加入模型的最大潜变量数目。
中心化方法	以下四种选其一: 1.无；2.平均值；3.中位数；4.最小值。	请参见变量标度化章节。
标度化方法	以下四种选其一: 1.无；2.标准偏差；3.标准偏差开方；4.四分位距(IQR)。	请参见变量标度化章节。
阈值	[0 ∞]，其中∞为变量选择的阈值。	被选变量的 Index Value 值不能大于该阈值。

变量选择完成后，将在工程导航栏中产生新的节点，示例数据的变量选择结果与 11.2.方法雷同，不再赘述。

11.15. Random Frog(PLS-DA)法

本方法基于 PLS-DA 分类构建，适合于基于该方法的变量选择。

具体操作方法请参考，在此不再赘述。

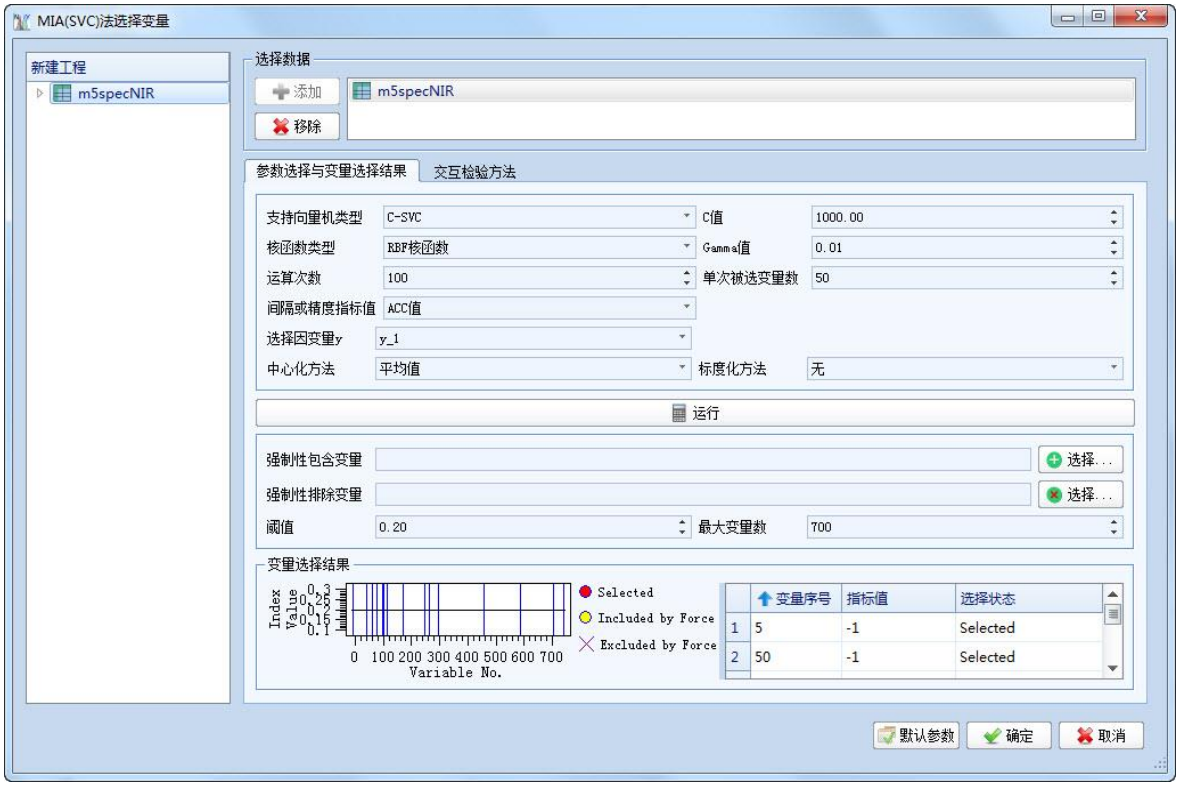
11.16. MIA(SVC)法

支持向量机以结构风险最小化为目标，以最优化方法在高维空间中训练得到达到最大间隔的分类模型，保证很好的模型泛化能力。本法则是完全基于支持向量机的特点而针对性地提出来的方法，包括如下几个步骤。

首先以蒙特卡罗方法从数据变量方向随机采样，获得子数据集，构建模型；构建格各子数据的支持向量机模型，记录其所用的变量和间隔值；统计分析各模型间隔值，将模型是否保护某变量分为二组，获得间隔分布，删除二组分布均值 <0 的变量，再采用 Mann-Whitney U 方法计算其余各变量 P 值，以 P 值作为指标判断间隔增值的变化显著性，并以该值是否大于设定的阈值作为引入变量的依据。同样地，P 值越大，显然越需要被选择；反之亦然。

操作步骤:

步骤 1: 点击**变量选择** -> **MIA(SVC)**，弹出如下对话框:



本法涉及交互检验相关的内容，详情请参见 12.1.2.2.。此外，本法的变量选择结果亦无“指

标值”项目。接下来的操作步骤参照变量选择之通用步骤。

本法中所涉及参数的意义，如下表所示：

参数	范围	说明
支持向量机类型	二其一: 1.C-SVC；2.nu-SVC。	选择 SVC 建模类型。
C 值	[1 ∞]，其中∞表示无穷大。	惩罚系数，仅当选择支持向量机类型为 C-SVC 的时候出现。
nu 参数	[0 1.0]	SVC 建模参数，仅当选择支持向量机类型为 nu-SVC 时出现。
核函数类型	四选一: 1.线性核函数；2.多项式核函数；3.RBF 核函数；4.Sigmoid 核函数。	选择核函数类型。
Gamma 值	[0 ∞]，其中∞表示无穷大。	SVC 建模参数，仅当选择核函数类型为多项式核函数或 RBF 核函数或 Sigmoid 核函数时出现。
阶数	[2 ∞]，其中∞表示无穷大。	多项式阶数，仅当选择核函数类型为多项式核函数时出现。
运算次数	[100 ∞]，其中∞表示无穷大。	程序运算的最大次数。
单次被选变量数	[2 col-2]，其中∞表示无穷大。	选择用于建模的变量数。
间隔或精度指标值	二选一: 1.间隔值；2.ACC 值。	模型评价指标。
选择因变量 y	根据数据实际情况选择。	无类别信息则无法计算。
中心化方法	以下四种选其一: 1.无；2.平均值；3.	请参见变量标度化章节。

	中位数；4.最小值。	
标度化方法	以下四种选其一: 1.无；2.标准偏差；3.标准偏差开方；4.四分位距(IQR)。	请参见变量标度化章节。
阈值	[0 ∞]，其中∞为变量选择的阈值。	被选变量的 Index Value 值不能大于该阈值。

变量选择完成后，将在工程导航栏中产生新的节点，示例数据的变量选择结果与 11.2.方法雷同，不再赘述。

11.17. MIA(SVR)法

本方法基于 SVR 回归构建，适合于基于该方法的变量选择。

具体操作方法请参考，在此不再赘述。

第十二章　建模

建模是本软件的关键功能，该步骤既是前面几步的主要目的，亦是下一章预测(或验证)的重要基础。在这一章里，我们将同时介绍探索性分析，分类和回归的建模方法，关于这些方法的初步介绍，详情请参见第十八章。

在介绍具体的建模方法之前，先概述这些方法中涉及的一些数据处理基础性问题，如模型验证方法等，以及建模中的通用操作步骤。

12.1. 基础介绍

12.1.1. 模型验证方法

模型验证是指基于已知模型估计其对未知数据预测的不确定性，获得未来可能的表现与泛化能力结果；若这种不确定性较低且在合理的范围内，则该模型可视为有效，能用于未知数据的预测，获得可靠结果。

估计模型稳定性与预测能力的方法很多，主要包括如下表所示的三大类：

序号	方法名称	说明
1	独立测试集	将整个数据划分为测试集和预测集，前者用于构造模型，而后者则包括所有剩余样本，使用前述模型获得预测统计结果。
2	交互检验法	同一样本同时用于模型估计与检验，是使用非常广泛的一类方法。
3	杠杆校正法	该法是交互检验方法的一个近似估计，即在没有实际对样本进行预测而估计得到预测残差，以此获得模型的评价结果。

独立测试集数据应占全部数据的 20-40%，且校正集和测试集应具有最大代表性，尽可能均

包括不同类别的样本，重复量测获得样本数据不能同时出现校正集与测试集中。

通常地，获得测试集的方法包括手动数据选择，随机数据选择，以及成组数据选择(如基于变量的数值范围等)。本软件提供丰富的数据载入和划分功能，比如单个或批量载入数据，在数据处理的不同阶段载入数据，将载入的数据添加到已有的基本数据表，或者建立新的数据矩阵，以及对数据进行行、列或任意划分等(详情请参看)，可满足用户对数据划分和使用的各种需求。

交互检验的方法则是首先将整个数据的一部分用于校正，另一部分则用于测试并记录所有结果，然后循环此过程，直到每个样本均被测试一次为止。基于上述基本原则，已经发展了一系列交互检验方法。杠杆校正法则是基于已知模型及其结果的一个估计，如下式所示：

预测残差 = 校正残差 /(1 – 样本杠杆)

当所有样本杠杆较低，即对模型的影响较低时，模型的预测残差将非常靠近校正残差；反之当样本杠杆较高时，由于校正残差与更小的数值相除，预测残差便会更大。

本软件中所用到的模型验证方法，如下表所示。

序号	方法名称	说明
1	留一法	每次构建校正模型时，留出一个样本，将建立好的模型预测被留出样本，直至每个样本有且只有一次被预测到为止，最后统计预测结果以评价模型。
2	随机选择	基于用户自定义的样本分割数与每个分割中所包含的样本数，系统随机自动产生交互检验样本集。
3	系统性选择 (112233 模式)	用户先自定义样本分割数与每个分割中所包含的样本数，而每个分割中的样本序号则是根据其所应包含的样本数，由第一个分割开始，按照顺序依次选择和排列。

4	系统性选择 (123123 模式)	与 112233 模式的差异在于样本按照分割优先的顺序排列，即不同上一模式中序号为 1，2，3…号样本位于同一分割中，而是位于不同分割中，循环往复排列，直至最后一个样本。
5	用户自定义	在随机选择的基础上，用户可手动任意设定不同分割中包含的样本，以及需要被排除的样本。

其中留一法尤其获得了广泛应用，尽管其往往对模型预测误差(Root Mean Squared Errors of Cross Validation, RMSECV)的估计过低，获得过于乐观的结果，交互验证均方根偏差如下式所示：

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

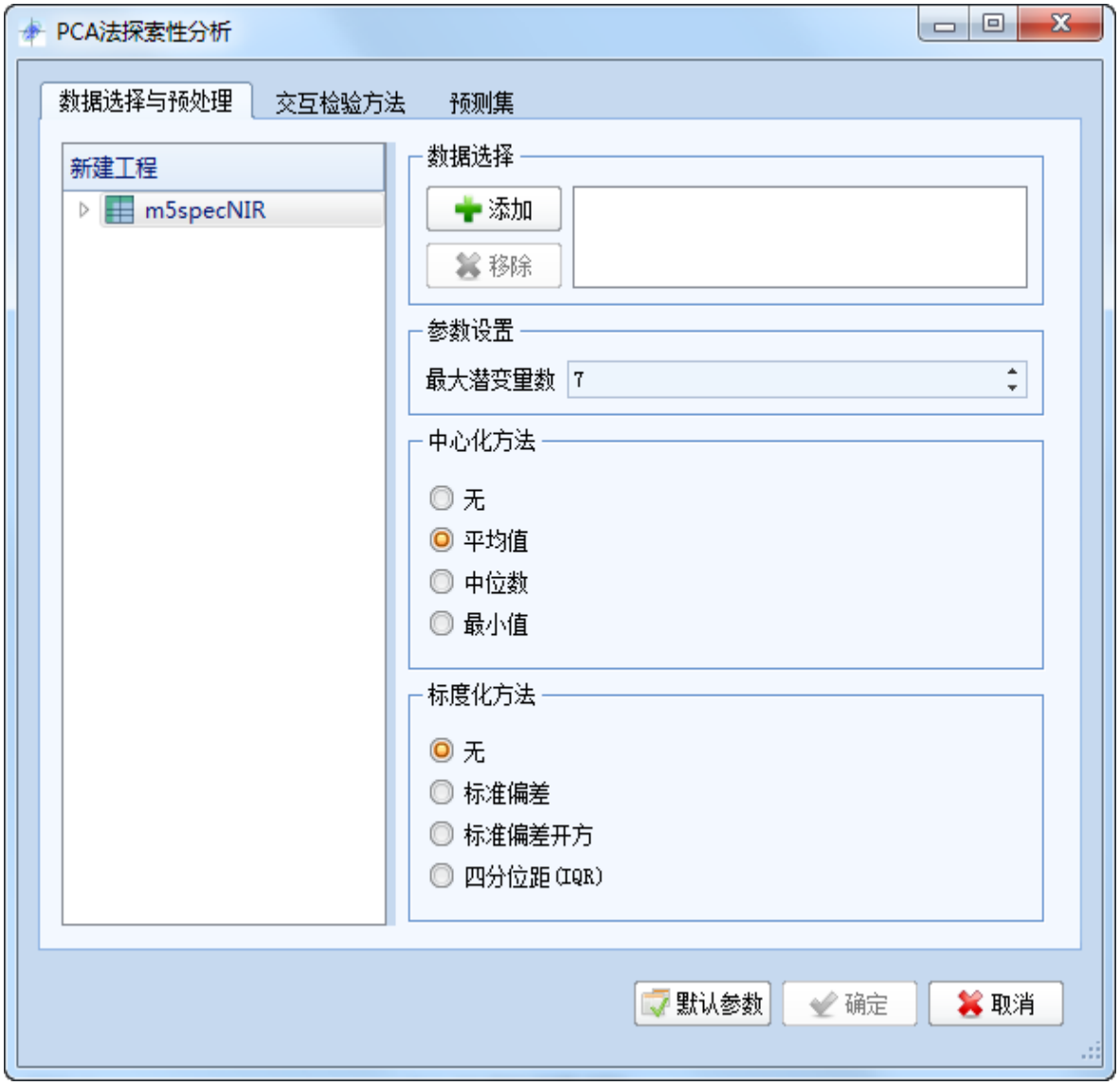
此外，用户可任意定义当前划分中的样本序号，以及被排除的样本序号等(若有)。详情请参见方法介绍中的内容。

ⓘ 上述方法在本软件中亦用于确定最优潜变量数等。

12.1.2. 通用步骤

12.1.2.1.数据选择与预处理

本部分与上述数据预处理以及变量选择雷同，即在进行数据分析前，需要添加目标数据，以便分析，如下图所示(以 PCA 分析为例)。



特别需要强调的是，本章关于建模的内容(以及上述预处理和变量选择)，是指独立使用本功能的情形。若用户通过构造算法流实现数据的批处理以及模型构建，则需使用部分所述的功能。

选择目标数据后，用户需设置方法参数，不同方法的参数可能有所不同，不同的内容将在介绍具体方法时详述。以图中所述 PCA 方法为例，用户需输入最大主成分数。

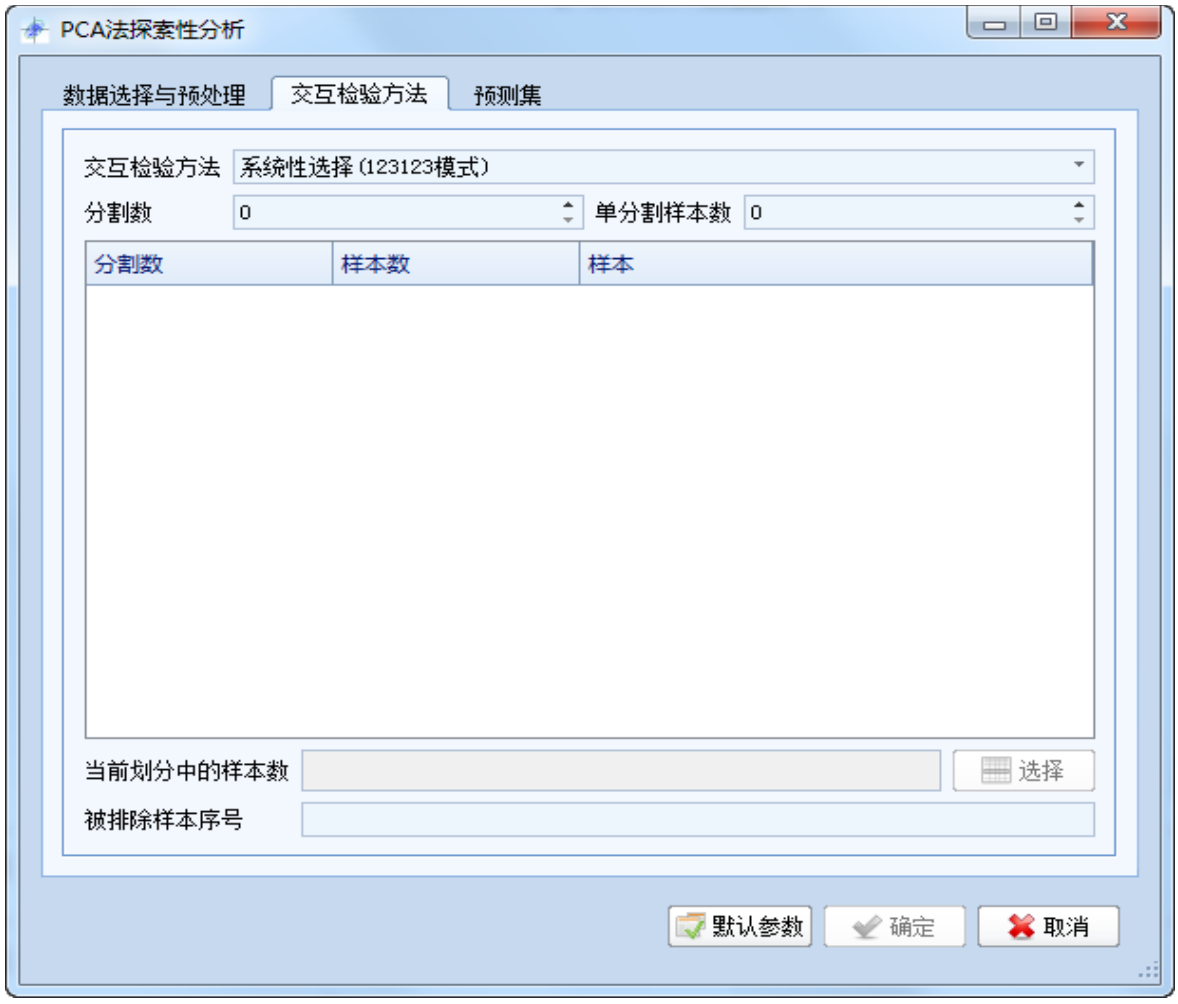
完成数据选择与参数设置后，则还需要设定变量标度化方法，系统默认状态中的中心化方法为平均值，而标度化方法为标准偏差开方。用户可在主页的参数设置中修改默认值。关

于这些方法的详细介绍，详情请参见第十八章，在此不再赘述。

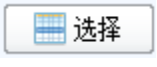
如前所述，用户亦可使用默认参数功能恢复系统默认值。

12.1.2.2.交互检验方法

交互检验是建模过程中的重要内容，以 PCA 分析为例，其交互检验窗口如下图所示。



其中具体方法的介绍，详见如下表。但使用中需注意如下几点：

- 若选择“留一法”，该标签页中其他内容则均无需再设置，按钮和文本框呈灰色非激活状态。
- 仅选择自定义交互检验方法时，选择按钮 

才可用，其他情形均呈灰色

非激活状态。

- 分割数与单分割样本数具有关联性，仅需设定其中的一个参数，另一个便可自动获得。

其他参数的具体含义，如下表所示。

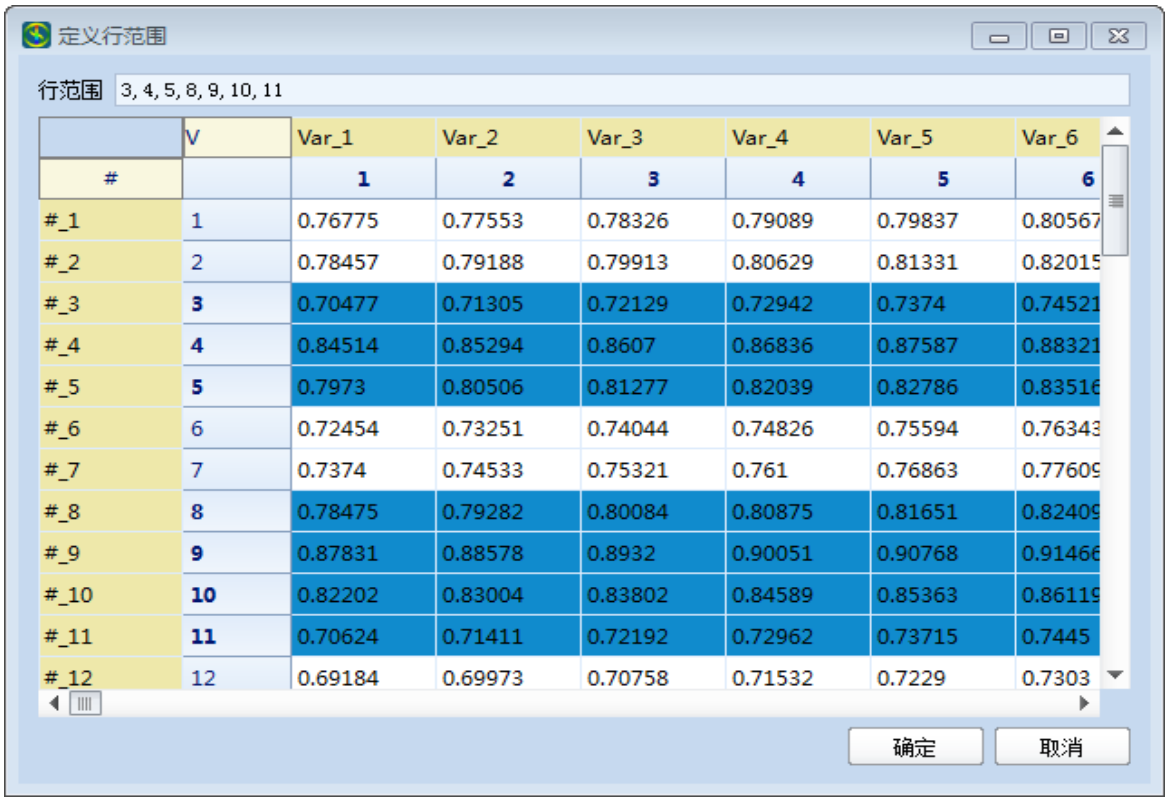
序号	交互检验方法	说明
1	分割数	指将整个数据分割为一定的段数，每段数据分别作为测试集，剩余数据作为校正集，以此构建模型并获得测试数据结果，直至所有测试集完成计算为止。
2	单分割样本数	与分割数对应，指每个分割数中可得到的样本数。实因数据样本数量的原因，每个分割里的样本数不一定完成相等(相差一个样本)。
3	当前划分中的样本数	指当前划分中的样本，可点击选择按钮选择；若点击界面中的不同分割，亦会显示当前分割下的样本序号。
4	被排除的样本序号	指用户排除的样本序号，可根据实际情况输入。

用户选择合适的交互检验方法和参数后，将在图中界面中间区域，详细显示各分割数序号，该分割中的样本数，以及实际样本序号等，如下图所示。

分割数	样本数	样本
1	6	1,10,19,28,37,46
2	6	2,11,20,29,38,47
3	5	3,12,21,30,39
4	5	4,13,22,31,40
5	5	5,14,23,32,41
6	5	6,15,24,33,42
7	5	7,16,25,34,43

如上所述，若选择用户自定义，则可点击选择按钮，并得到如下图所示的界面。用户可在

此界面选择样本(Ctrl 键可用)，同时在行范围中显示被选的样本序号，如下图所示。与此同时，若用户在行范围框中输入样本序号，并在相邻样本间以逗号分开，则系统亦将自动地动态选中对应的样本。



	V	Var_1	Var_2	Var_3	Var_4	Var_5	Var_6
#		1	2	3	4	5	6
#_1	1	0.76775	0.77553	0.78326	0.79089	0.79837	0.80567
#_2	2	0.78457	0.79188	0.79913	0.80629	0.81331	0.82015
#_3	3	0.70477	0.71305	0.72129	0.72942	0.7374	0.74521
#_4	4	0.84514	0.85294	0.8607	0.86836	0.87587	0.88321
#_5	5	0.7973	0.80506	0.81277	0.82039	0.82786	0.83516
#_6	6	0.72454	0.73251	0.74044	0.74826	0.75594	0.76343
#_7	7	0.7374	0.74533	0.75321	0.761	0.76863	0.77609
#_8	8	0.78475	0.79282	0.80084	0.80875	0.81651	0.82409
#_9	9	0.87831	0.88578	0.8932	0.90051	0.90768	0.91466
#_10	10	0.82202	0.83004	0.83802	0.84589	0.85363	0.86119
#_11	11	0.70624	0.71411	0.72192	0.72962	0.73715	0.7445
#_12	12	0.69184	0.69973	0.70758	0.71532	0.7229	0.7303

12.1.2.3. 预测集设置

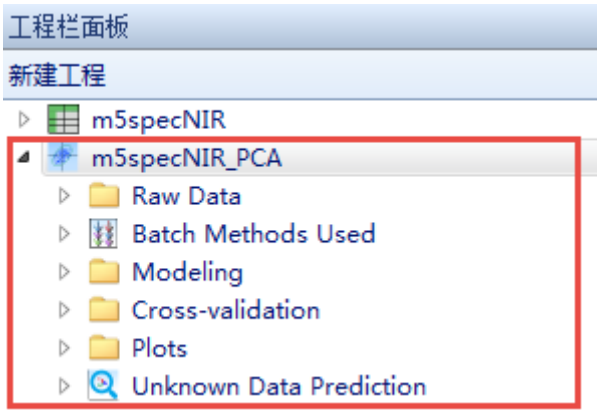
通过前述设置便可构建模型。而获得需要的模型后，便可对未知数据进行预测。本部分主要实现预测集的添加，这样便在点击确定后，可同时得到建模和预测结果，如下图所示。

需要注意的是，当使用应用批功能时间，用户可在同一界面中同时设置训练集，验证集，以及预测集。预测集的列数(即变量个数)必须与用于建模的数据矩阵列数相等，否则将无法成功添加预测矩阵。此外，预测功能除在建模时选择数据并获得预测结果之外，亦可对已有的模型做出预测(在工具栏的预测目录下选择)。

设置完成上述数据和参数后，点击确定及开始运算，并快速获得结果；若点击取消，则关闭界面，不产生任何结果。

12.1.3. 模型结果概述

通过上述步骤即可获得模型结果，并在工程导航栏中以节点文件夹和节点的形式列举这些结果，如下图所示。模型结果的节点文件夹名称为校正数据文件名加上建模方法名称，中间加上下划线，如图中的“m5specNIR_PCA”。

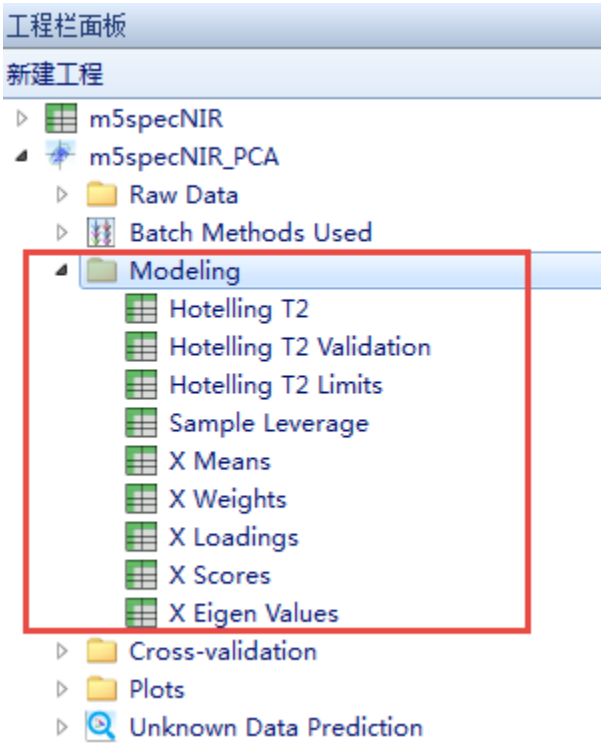


在模型结果节点文件夹下，包含如下表中的子节点文件夹。

序号	节点文件夹名称	说明
1	Raw Data	建模时所选数据的一个副本，即将建模时所用的数据重新复制一份置于结果文件夹下，以保持节点的完整性。
2	Batch Methods Used	批处理方法，即保存建模时的一系列方法，包括参数设置等。本软件将单个数据处理方法的使用亦自动以批的形式体现，以

		保证其可比性与完整性。
3	Modeling	建模所产生的系列表格结果。
4	Cross-validation	建模过程中涉及交互检验时，所产生的结果。
5	Plots	图形结果的节点文件夹，主要是对 Modeling 和 Cross-validation 节点文件下数据的绘图。
6	Unknown Data Prediction	未知数据的预测结果节点文件夹。若建模时并未添加预测数据，则忽略该节点。

各节点文件夹下具体节点所对应的表格和图形结果，则在介绍具体建模方法时详述，如 Modeling 下的节点如下图所示。



下面依次介绍具体的建模方法，并详述模型结果，以及结果的解释等，以求用户可透彻理解，并通晓方法的应用。

12.2. PCA 法

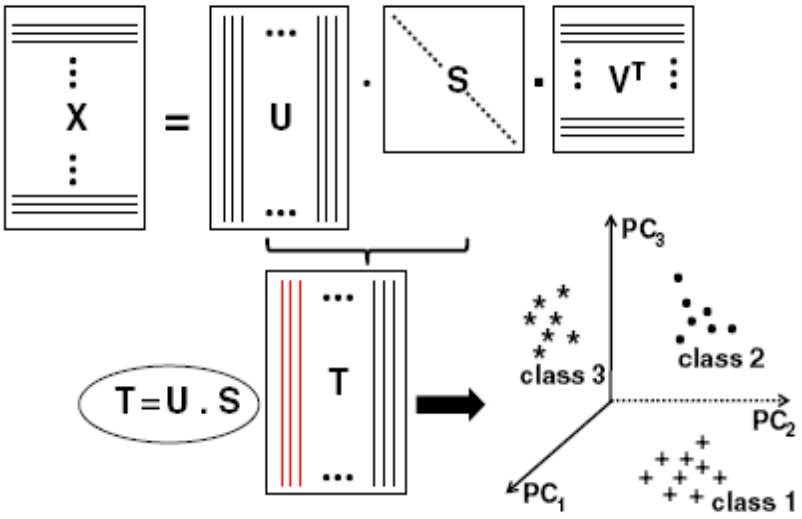
探索性数据分析辅助获得初步数据信息与结果，通常以可视化的形式表达。PCA 法是其中最重要的方法之一，亦是其他一系列多变量数据分析方法的基础(如 PCR，PLS 和 SIMCA 等)。本法可很好地展示高维数据的隐含结构信息，如可视化获得样本和/或变量间的关系，并解释究竟哪些变量导致样本间的相似性或差异性。如前所述，使用本方法时同样约定数据矩阵的行为样本或研究对象，而列则为变量或描述符。

该法通过将原始数据矩阵 **X** 分解为不同信息含量的正交主成分(PCs)，每个主成分解释原始数据中一定量的数据信息(变化)，第一主成分信息含量最丰富，其他主成分则依次递减，如下式所示：

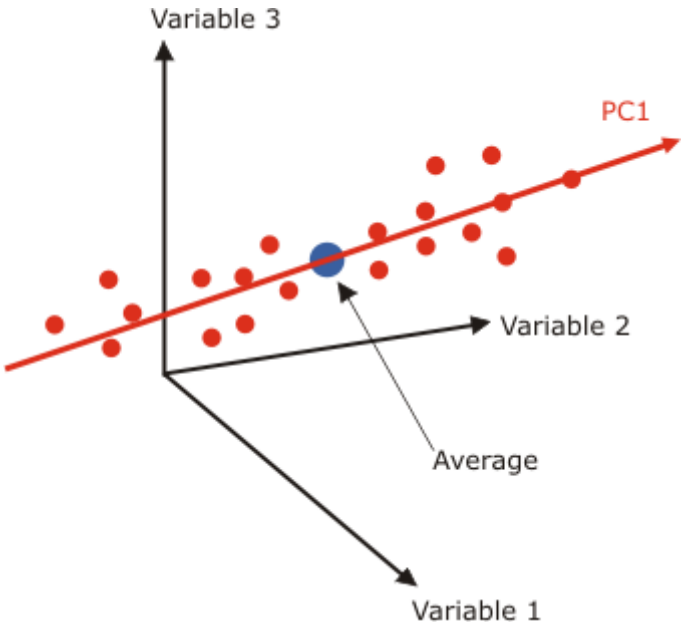
$$\mathbf{X} = \mathbf{USV}^T + \mathbf{E} = \mathbf{TP}^T + \mathbf{E}$$

在上式中，矩阵 **T**、**P** 和 **E** 分别为得分、载荷和误差矩阵。

如下图所示，实因我们能观察的最高维度仅为三维，在样本和变量数较多时，样本亦以点的形式分布在由更多变量组成的高维度空间中。在上图中，不同类别的样本点显然可由 PC1、PC2 和 PC3 组成的图形度量。若图中二样本在三个主成分上的投影很接近，则可将他们视为相似样本，如图中第一、第二或第三类样本；若二样本具有明显差异，则这些主成分中的一个或多个同样应具有显著差异。显然上述图形结果的解释，可拓展到更高维度的数据矩阵中。



综上所述，该法的原理是依次寻找样本所构成的空间中距离(散布度)最大的方向，如下图中所示的 PC1，更多主成分可依次类推。



若以图形的方式绘出多个主成分，则可解释样本间的相互关系。具体来说，PCA 法可达致如下目标:

- 哪些变量(如光谱长波)描述了不同样本间的差异?
- 哪个或哪些样本对描述上述差异的贡献最大?
- 哪些样本相互关联，即他们以相似的方式解释样本间的相似性与差异性?

基于此，PCA 法可用于检测样本的整体分布模式，发现显著或非显著性的奇异样本，并定量描述上述数据信息。

特别需要指出的是，PCA 法是通过正交双线性矩阵分解，获得不同主成分以解释数据中最大方差，在此条件下可获得唯一解。然而 PCA 法计算所得到的是所谓抽象解，可用于解释数据结构和数据中的变化，但不是导致数据实际变化的真实解，而是他们的线性组合。

计算主成分分析中主成分的方法主要有非线性迭代偏最小二乘法(Non-linear Iterative

Partial Least Squares)，以及奇异值分解法(Singular Value Decomposition)二种。前者每次计算得到一个主成分，且可处理含有却是值的数据，而后者再一次计算中同时获得所有主成分，但不能处理存在缺失值的情形。

12.2.1. 操作说明

具体内容与 12.1.3.章节相同，不再赘述。

12.2.2. 模型结果概述

关于模型结果的初步介绍，请参见 12.1.3.，下面一一各节点文件夹的详细内容。

12.2.3. Raw Data 节点

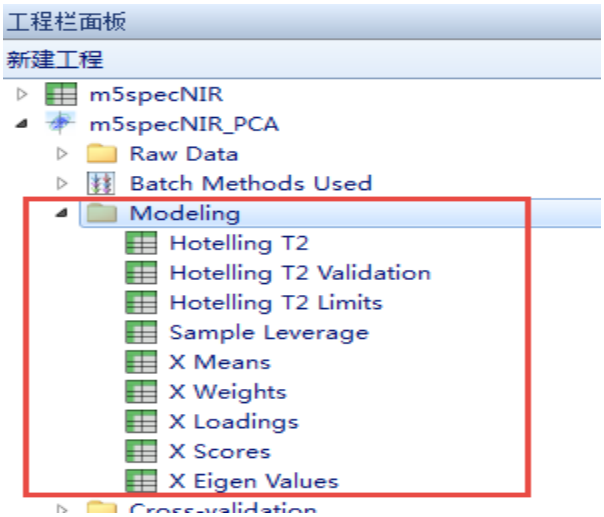
请参见 12.1.3.。

12.2.4. Batch Methods Used 节点

请参见 12.1.3.。

12.2.5. Modeling 节点

打开该节点，可得到如下图所示的模型结果节点。

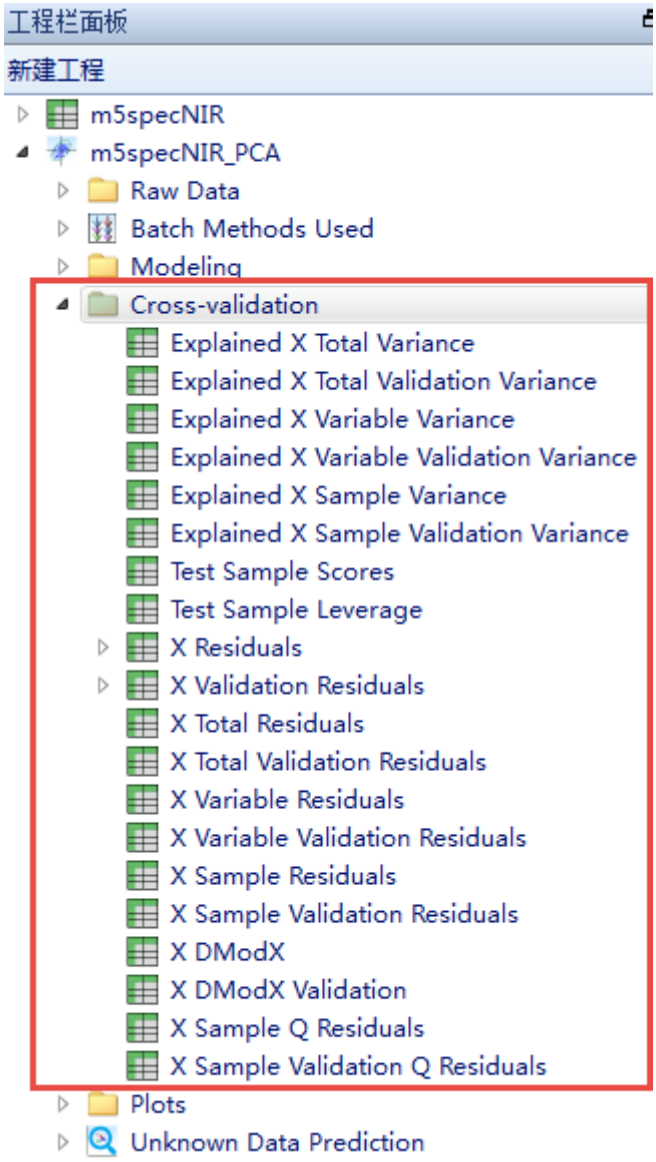


上图中的节点以表格的形式，完整记载 PCA 分析所得到的结果。各节点的详细信息如下表(假设原始数据矩阵 X 的大小为 *m*×*n*，即含有 *m* 个样本，*n* 个变量)。

序号	节点名称	说明
1	Hotelling T2	模型内部变化的统计量，表示每个采样点在变化趋势和幅值上偏离模型的程度。本软件中则计算不同主成分数下，各样本到模型中心的距离，从而构成 *m*×*A* 的矩阵，其中 *m* 和 *A* 分别指样本和主成分数。 其的定义为：$T_i^2 = t_i\lambda^{-1}t_i^T = x_iP_K\lambda^{-1}P_k^Tx_i^T$
2	Hotelling T2 Validation	校正集的 Hotelling T2 值。
3	Hotelling T2 Limits	定义 Hotelling T2 的极限值，比如 5%则意味着 95%样本到模型的距离在此限制之内，而不能包括在此限制内样本，则很可能是奇异值。
4	Sample Leverage	量测样本投影到模型中心距离的指标，可度量某样本与另一类样本的差异性，而不管该样本是否能被模型描述。 $Leverage_i = H_{i,i}$，　其中 $\mathbf{H} = \mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T$ 注意，样本杠杆值与 Hotelling T2 有线性关系。
5	X Means	数据矩阵 **X** 变量方向的均值，大小为 *1*×*n*。
6	X Weights	数据矩阵 **X** 变量方向的权重，大小为 *1*×*n*。
7	X Loadings	经 PCA 分解后得到的载荷矩阵，大小为 *m*×*A*。
8	X Scores	经 PCA 分解后得到的载荷得分，大小为 *m*×*A*。
9	X Eigen Values	经 PCA 分解后得到的特征值，大小为 *1*×*A*。

12.2.6. Cross-validation 节点

本软件所涉及的主要交互检验方法，已经在 12.1.1.中做出详细介绍。本节主要介绍进行 PCA 分析时，由交互检验步骤所得到的结果，该节点文件夹的展开后如下图。



交互检验节点文件夹下各具体节点的意义如下表示，用户可基于结果解释，更加深入地理解被分析的实际数据，以此得出正确的判断结果(同样假设原始数据矩阵 X 的大小为 $m \times n$，即含有 m 个样本，n 个变量)。

序号	节点名称	说明
1	Explained X Total Variance	总被解释的 **X** 数据方差，用于量测原始数据中被模型所解释的数据变化，即被模型所解释的数据结构百分比。其计算可由下式表示， 100×(初始方差–残差方差)/初始方差大小为 *A*×*1*。
2	Explained X Sample Total Validation Variance	验证集总被解释的样本方差，其大小为 *A*×*1*。
3	Explained X Variable Variance	被解释的变量方差，其大小为 *A*×*n*。
4	Explained X Variable Validation Variance	验证集被解释的变量方差，其大小为 *A*×*n*。
5	Explained X Sample Variance	被解释的样本方差，其大小为 *m*×*A*。
6	Explained X Sample Validation Variance	验证集被解释的变量方差，其大小为 *m*×*A*。
7	Test Sample Scores	预测样本得分，其大小为 *m*×*A*。
8	Test Sample Leverage	预测样本杠杆值，其大小为 *m*×*A*。
9	X Residuals	数据矩阵残差，由 A 个矩阵构成，每个矩阵的大小均为 *m*×*n*。
10	X Validation Residuals	验证集数据矩阵残差，由 A 个矩阵构成，每个矩阵的大小均为 *m*×*n*。
11	X Total Residuals	数据矩阵 **X** 的总残差，其大小为 *A*×*1*。
12	X Total Validation Residuals	验证集矩阵 **X** 的总残差，其大小为 *A*×*1*。

13	X Variable Residuals	数据矩阵 **X** 的变量残差，其大小为 *A*×*n*。
14	X Variable Validation Residuals	验证集矩阵 **X** 的变量残差，其大小为 *A*×*n*。
15	X Sample Residuals	数据矩阵 **X** 的样本残差，其大小为 *m*×*A*。
16	X Sample Validation Residuals	验证集矩阵 **X** 的样本残差，其大小为 *m*×*A*。
17	X DModX	数据样本到模型中心的距离，其大小为 *m*×*A*。
18	X DModX Validation	验证集样本到模型中心的距离，其大小为 *m*×*A*。
19	X Sample Q Residuals	量测新样本偏离已知模型程度，是模型外部数据变化的量度。其计算公式为，$Q_i = \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} = \mathbf{x}_i^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_a\mathbf{P}_a^{\mathrm{T}})\mathbf{x}_i$，其中 $\mathbf{e}_i$ 为残差矩阵 **E** 的第 *i* 行，而 $\mathbf{P}_a$ 为前 *a* 个主成分载荷，**I** 则为单位矩阵。
20	X Sample Validation Q Residuals	验证样本的 Q 统计量。

在上表中，主要涉及方差和残差的计算，其计算方式如下。

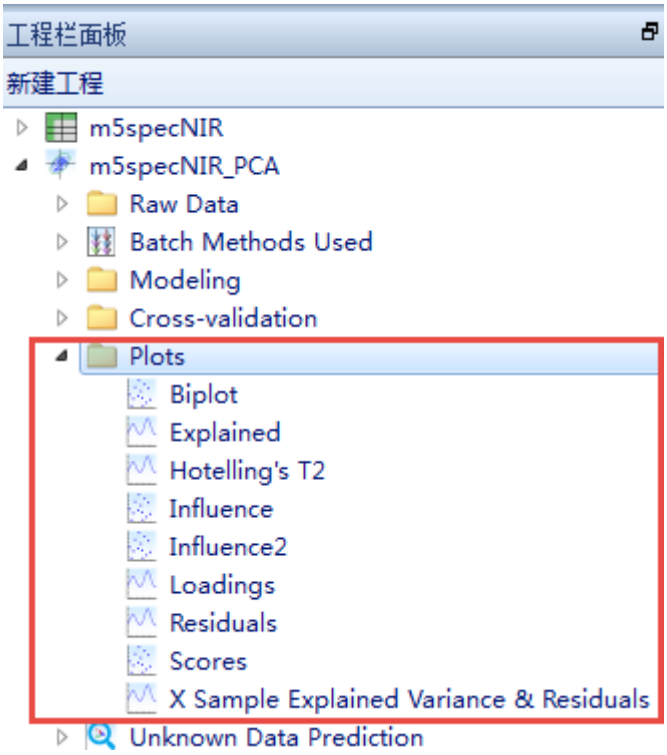
$$Var(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x})^2$$

$$Res(\mathbf{x}) = \mathbf{x} - \mathbf{x}_{mod}\ (\text{原始值} - \text{模型值})$$

12.2.7. Plots 节点

以可视化图形的形式表达 PCA 分析结果，是该法的核心内容。本软件涵盖不同结果的可视化表达形式，通过这些图形，用户可很好地获得不同样本，样本与变量，以及变量间的关系，解释实际问题。

Plots 节点文件夹下所包含的图形结果节点，如下图所示。



12.2.7.1 图形数据来源

上述图形所对应的 PCA 分析结果数据，其来源总结于下表，用户可以此更好地理解图形涵义及解释。

序号	图形节点名	说明
1	Bi-plot	此图是 Scores 和 Loadings 的综合表达，前者取自 X Scores，而后者则取自 X Loadings。
2	Explained	对校正集，若选中被解释信息项，则表示 Explained X Total Variance，若选中残差项，则表示 X Total Residuals。 对验证集，若选中被解释信息项，则表示 Explained X Total Validation Variance，若选中残差项，则表示 X Total Validation Residuals。
3	Hotelling's T2	若选中 T2 值，则表示 Hotelling's T2，若选中杠杆值，则表示 Sample Leverge。当选中 T2 值时，红色水平线显示 Hotelling's T2 Limits 值。

4	Influence	校正集：其中 X 取自 Sample Leverge。若选中被解释信息，y 取自 Explained X Sample variance；若选中残差，则 y 取自 X Sample Residuals。
		验证集：其中 X 取自 Test Sample Leverage。若选中被解释信息，y 取自 Explained X Sample Validation Variance；若选中残差，则 y 取自 X Sample Validation Residuals。
5	Influence2	校正集：对 X 轴，若选中 T2 值，为 Hotelling's T2；若选中杠杆值，则为 Sample Leverage。而对 y 轴则为 X sample Q Residuals。
		验证集：对 X 轴，若选中 T2 值，为 Hotelling's T2 Validation；若选中杠杆值，则为 Test Sample Leverage。而对 y 轴则为 X Sample Q Residuals。
6	Loadings	即为 X Loadings 载荷。
7	Residuals	即为 X Sample Q Residuals。
8	Scores	对校正集为 X Scores，对验证集即为 Test Sample Scores。
9	X Sample Explained Variance & Residuals	对校正集，若选中被解释信息，为 Explained X Sample Variance；若选中残差，则为 X Sample Residuals。
		对验证集，若选中被解释信息，为 Explained X Sample Validation Variance；若选中残差，则为 X Sample Validation Residuals。

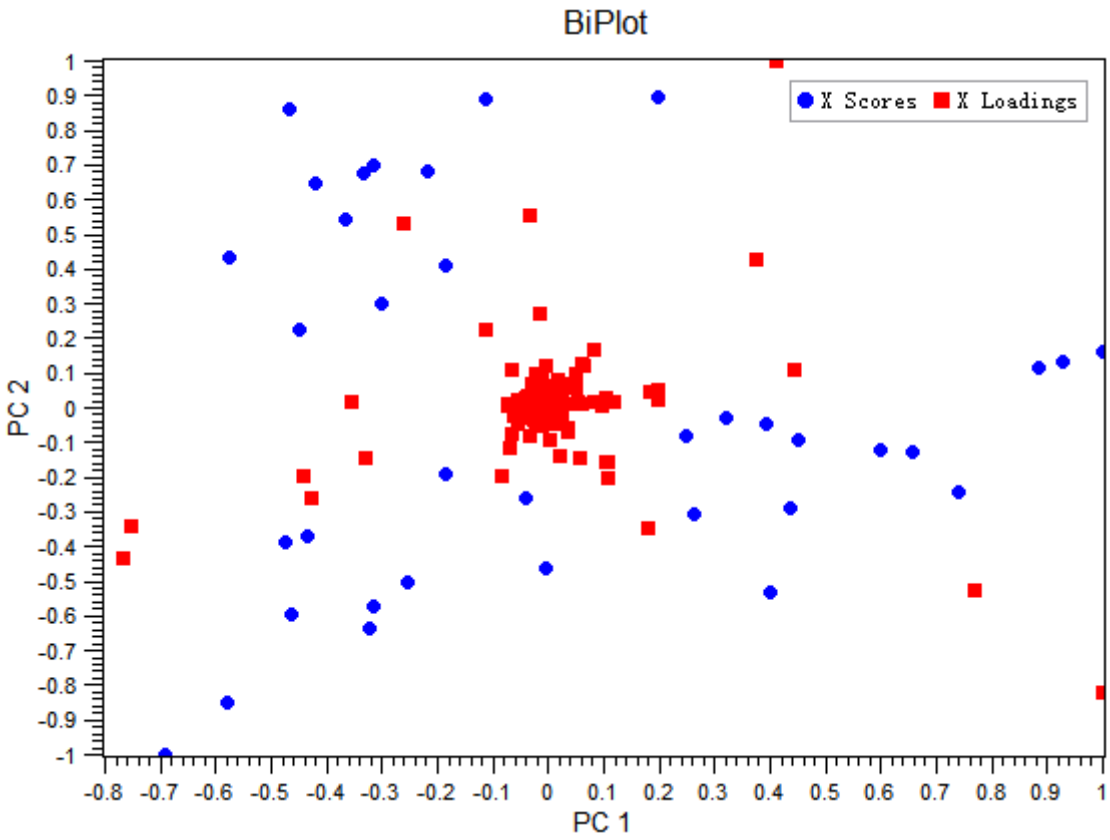
下面以一个代谢组学为例获得结果，以对上述图形一一做出说明，特别是图形所表达的意义及其解释。

12.2.7.2. Bi-plot

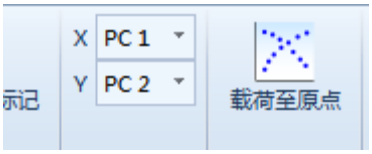
Bi-plot 图因其所含有的丰富信息，得到广泛的使用。通过该图，可清楚看出各变量对主成分的贡献，以及样本被不同主成分所解释的程度。

（一）操作说明

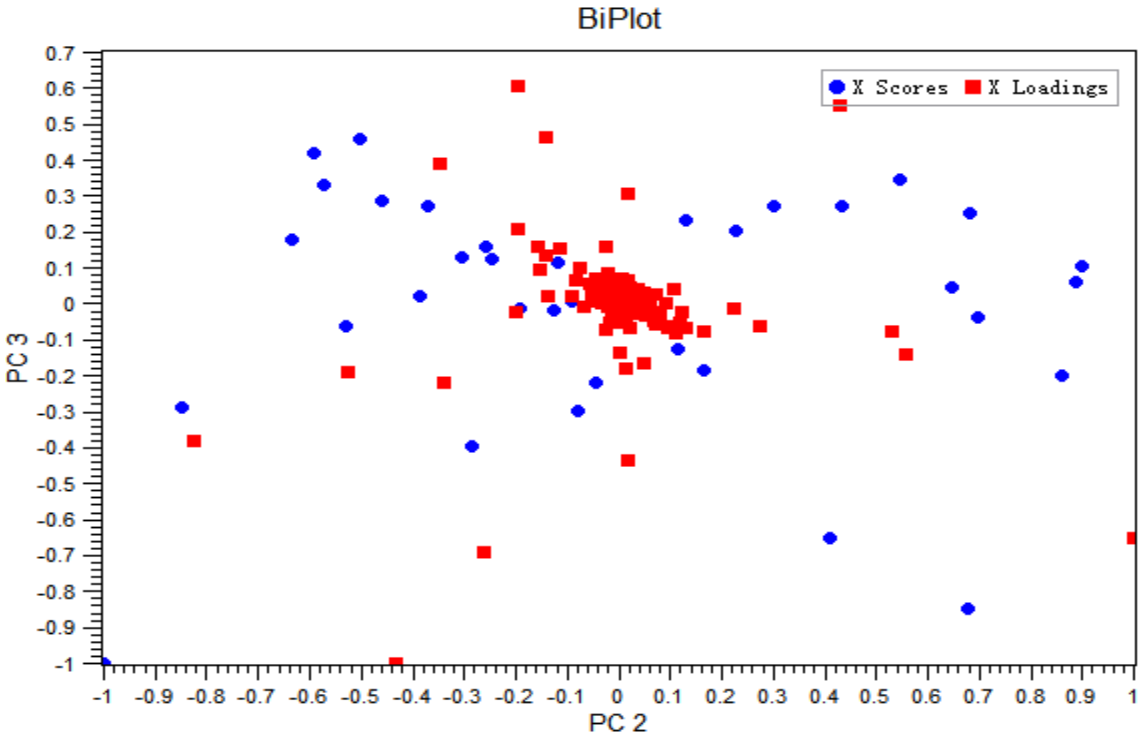
点击该节点后，可得到如下图所示的图形。



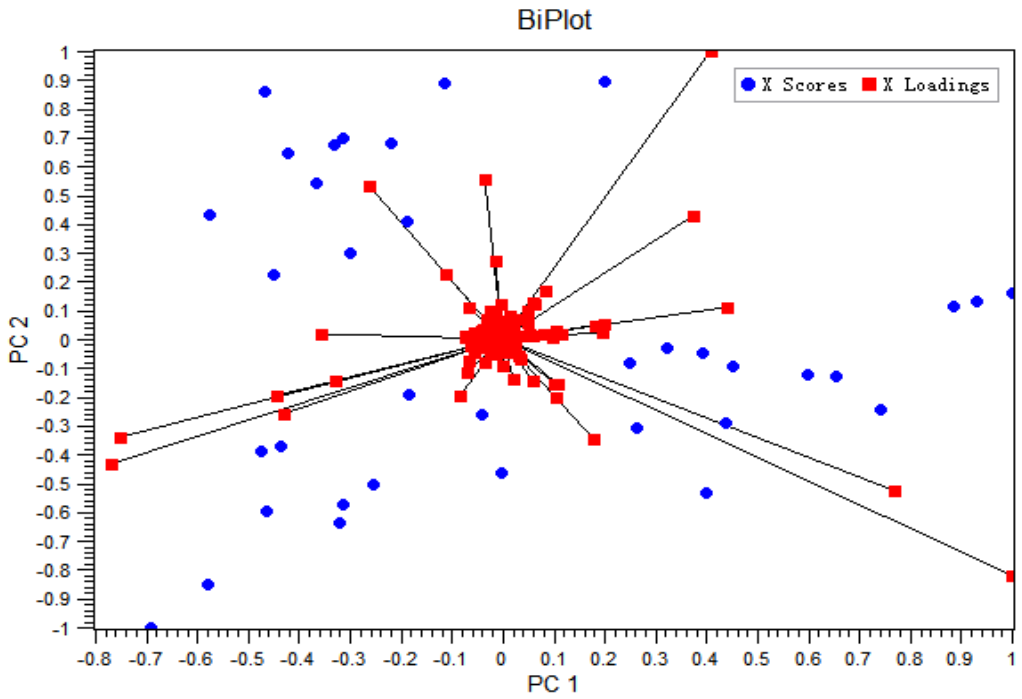
如上所述，在上图中同时绘出得分和载荷结果，并以颜色和点的符号差别显示。在图形的工具栏，除出现 9.4.章节详述的图形功能外，同时增加如下图所示的功能。



上图提供用户修改绘图的主成分数，并在图中得到即时更新，如若将 X 和 Y 坐标分别修改为 PC2 和 PC3，则得到如下图形。



若点击载荷至原点按钮使其处于选中状态，则所有载荷点将绘制至原点的直线，如下图所示。

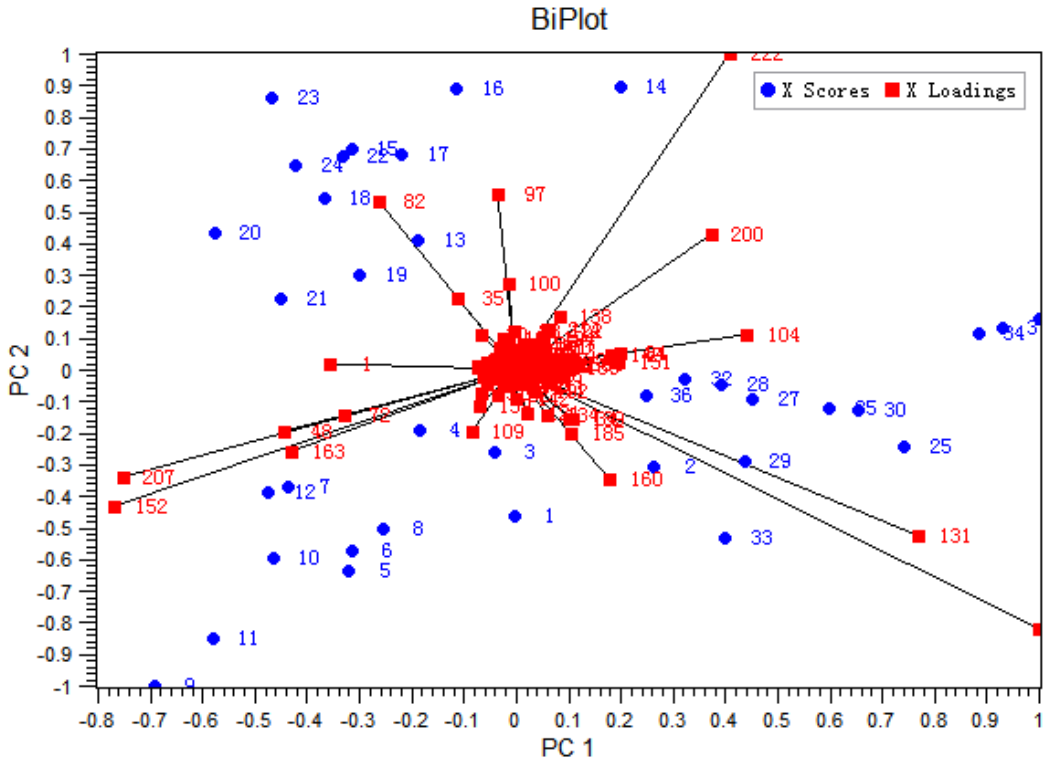


从上图可以清楚得出不同样本类别间的分布与聚类信息，变量间的相关性，以及各变量对分类的影响与贡献，即对样本分类有重要贡献的变量信息。

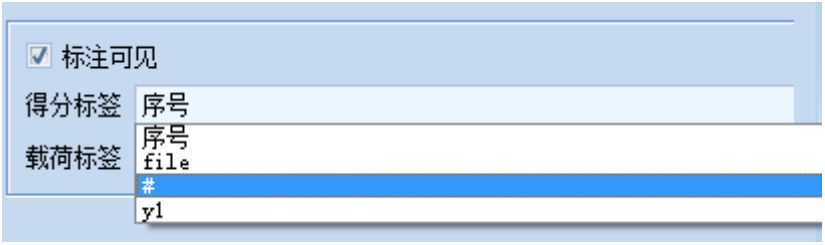
（二）属性修改

除已在 9.2.章节详细介绍的图形属性修改外，同时针对模型结果图形的个性化属性修改，以使用户更好地理解，解释和应用模型结果。其操作与 9.2.章节相同。属性修改界面的初始状态如下图所示。

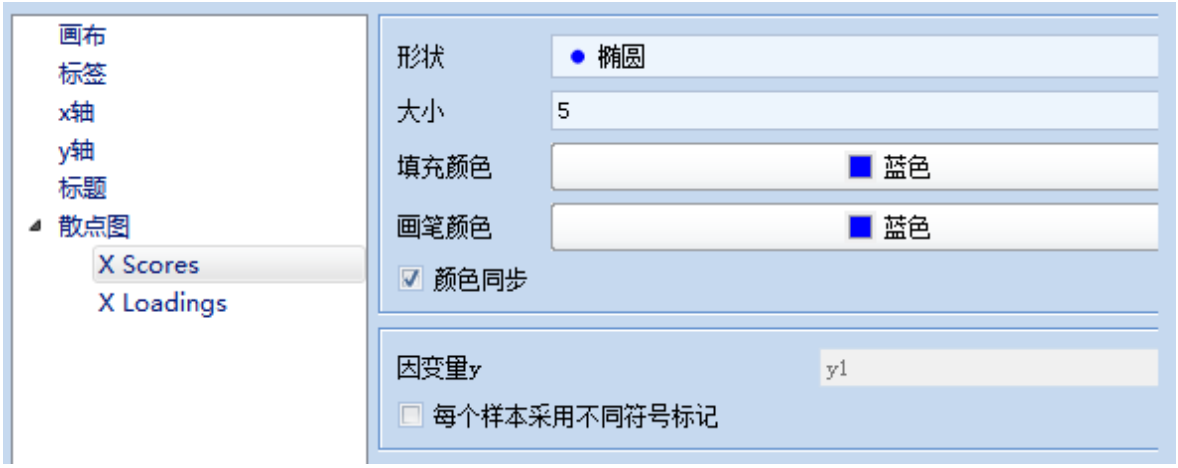
若勾选上图中的标注可见功能，则在 Bi-plot 图形中同时加上得分和载荷的序号，以帮助用户知道具体的样本与变量信息，如下图所示。



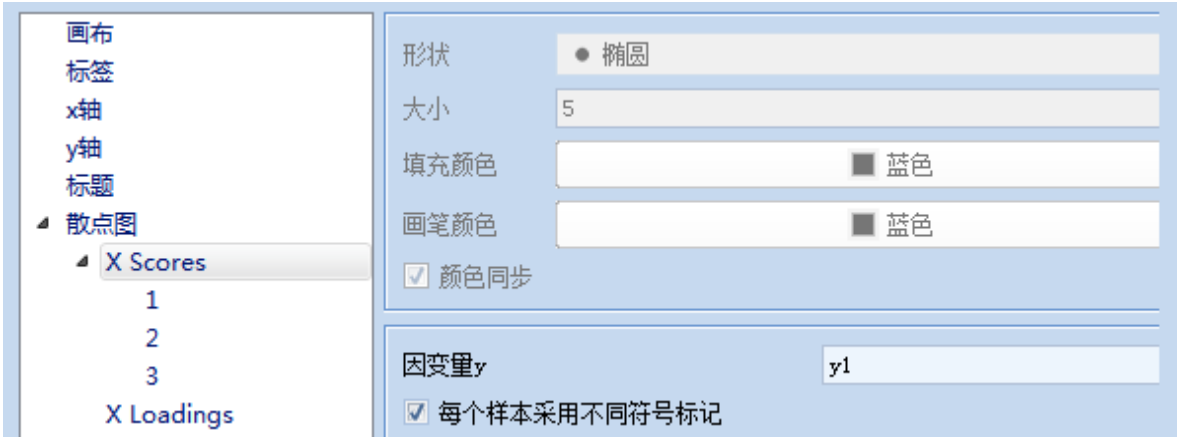
与此同时，用户亦可选择标注得分与载荷的其他标签(若有)，基于此功能，用户可在被处理的数据中添加任意的信息，并添加到模型结果图形中。



点击 X Scores 可得到如下图所示的图形。



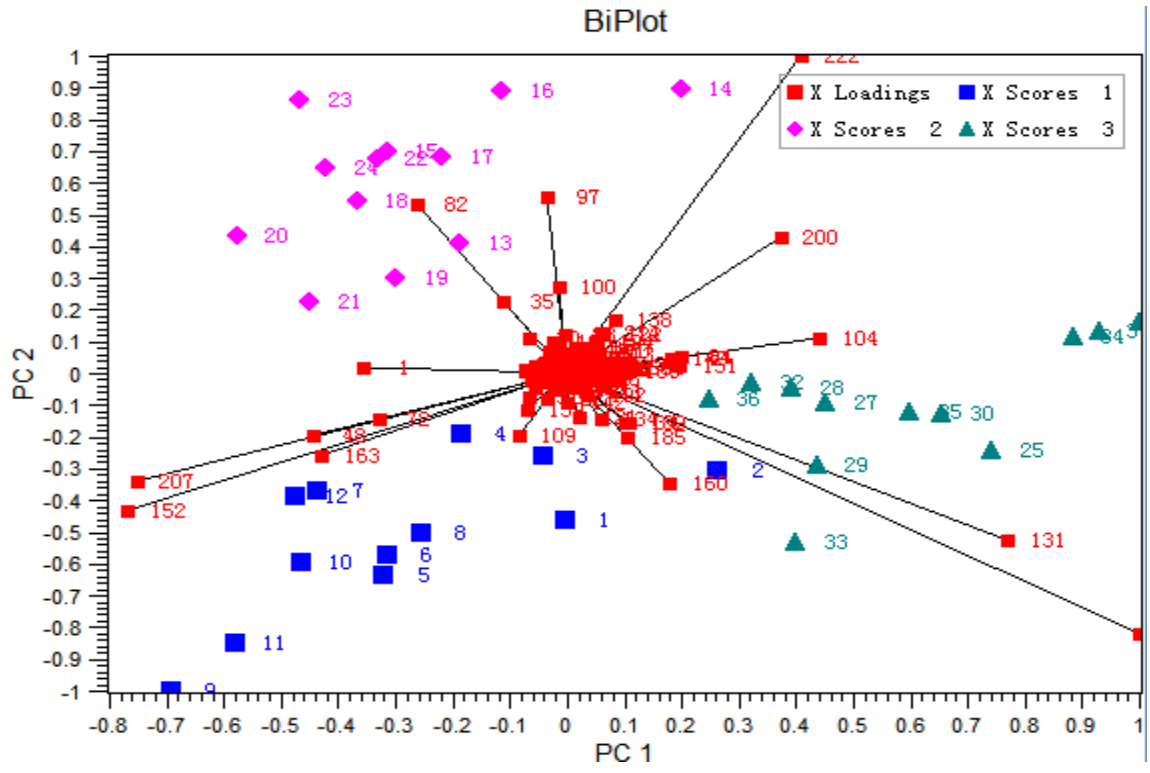
用户可通过上图修改 Bi-plot 图中得分点的属性，包括其形状、大小、填充颜色，以及画笔颜色等。若勾选上图中的每个样本采用不同符号标记功能，则可选择不同的响应 y 值，根据其类别属性，修改 Bi-plot 中不同类别样本的属性，如下图所示。



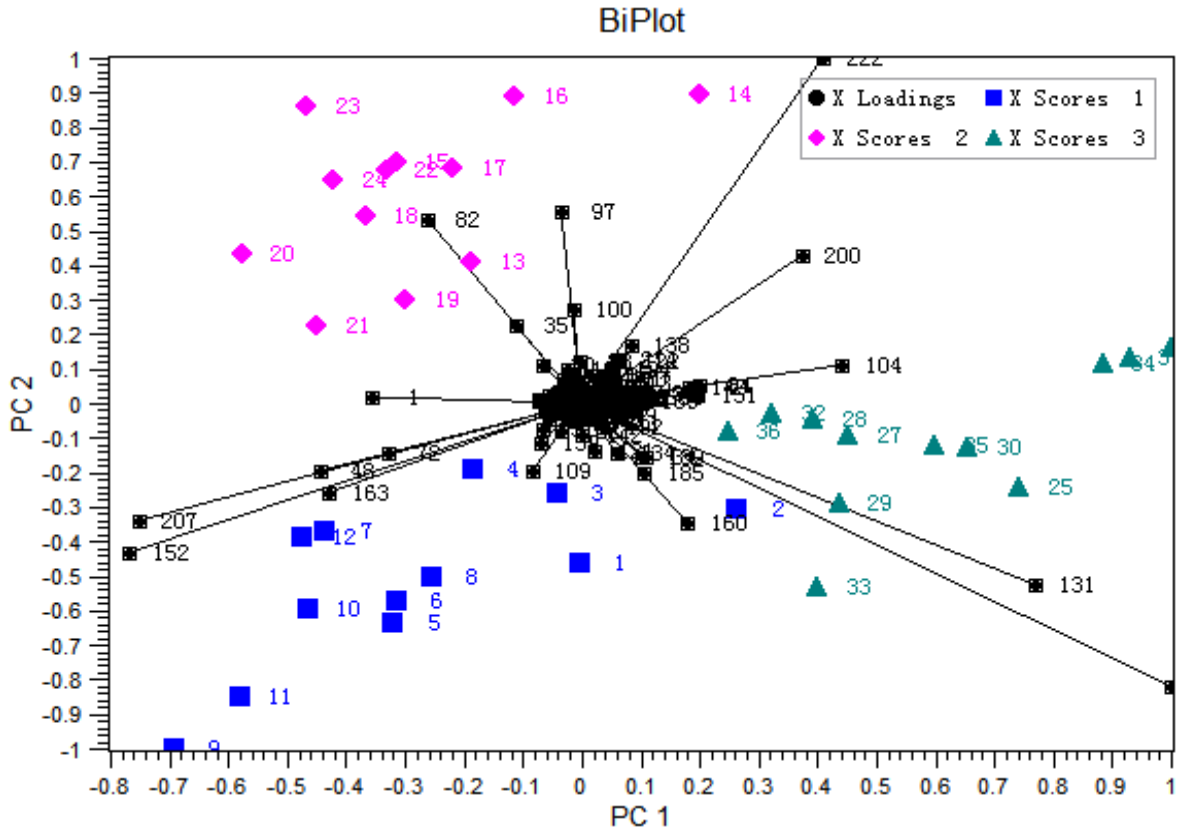
上图中的上面部分的属性变灰失效，但左侧 X Scores 下增加显示了样本类别的信息，即此时不同类别样本共有的属性编辑失效，以实现不同类别样本的个性化修改。

若点击选中 1，即样本类别 1(根据因变量 y 的选择动态变化)，得到如下图所示的图形，用户可进行合适的属性修改。

同时对各类别样本的属性进行编辑，可得到如下图所示的图形，非常清楚地显示不同类别样本的聚类分布情况，并在图形注释中显示对应的信息。



同样地，可对 X Loadings 的属性进行修改，以得到同时显示得分，载荷，以及其相互间关系的图形，如下图所示。



（三）图形解释

如前所述，Bi-plot 图信息含量非常丰富，是 PCA 分析最重要的结果之一。综合相同主成分下的得分图与载荷图，可得到究竟哪些变量的贡献，导致样本聚类结果的信息，这也是 Bi-plot 图相比于其他图形所能得到的额外信息。

PCA 模型所得到的结果，主要包括如下三个部分，其中得分矩阵描述样本属性及其模式，通常以不同主成分图形的形式表示样本间的相互关系，亦可以曲线图的形式表达，以表示不同样本的时间演进过程；载荷矩阵则描述变量间的相互关系，可以曲线或散点图的形式表示；被解释(或残差)方差则度量到底各主成分分别考虑到多少信息。其中，被解释的方差是指当前主成分所能度量的总方差百分比，而残差方差是指去除某主成分后，还有多少数据变化保留剩余的数据之中。

需要注意的是，得分和载荷相辅相成，不能相互脱离对方来独立解释，即没有得分无法解释载荷，反之亦然。在相同主成分方向，得分接近的样本为相似样本，同理得分相差很大时，对应的样本亦差别明显。基于此，Bi-plot 图可很好地用于解释样本、变量，以及相互间的关系。

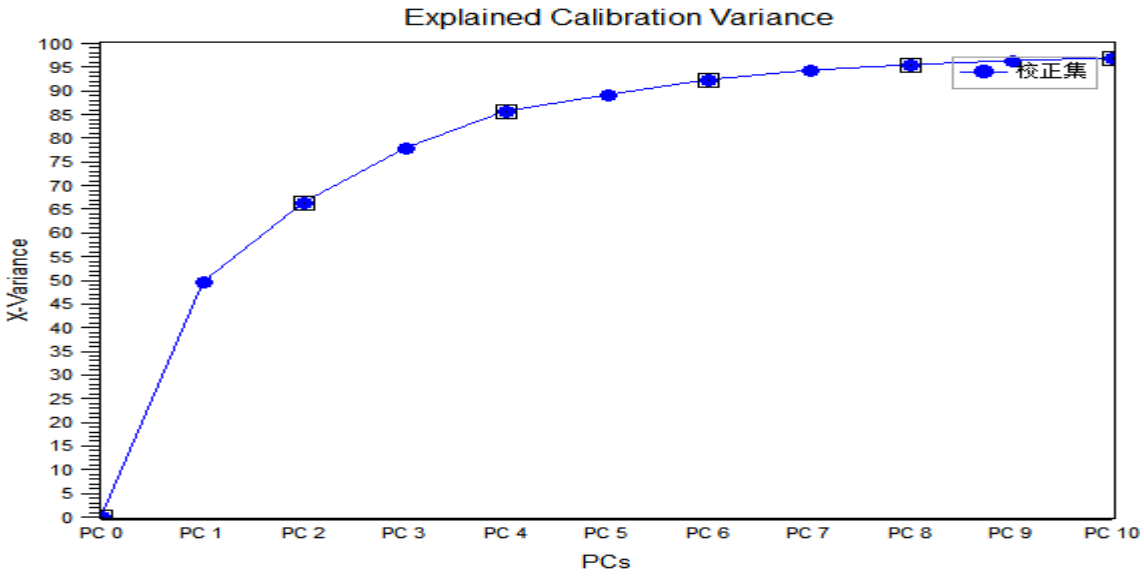
样本越靠近，则在当前主成分下越相似，反之亦然；从载荷图则可发现对于获得样本聚类有贡献的变量信息，如载荷图中越右侧的变量，对得分图中右侧样本的解释性也越高；反方向象限内的变量，则可能具有负相关的趋势；而靠近图形原点中心的变量，则在该图中难于被解释。而样本与变量间的关系，则是变量投影到样本上其投影越大则对聚类的贡献亦越大，反之亦然。

12.2.7.3. Explained

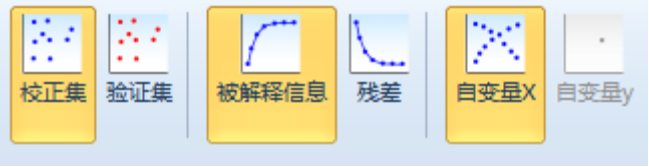
具体操作步骤等内容不再赘述，用户可参考 9.2.1.1.，以及 12.2.7.2.部分。

（一）结果介绍

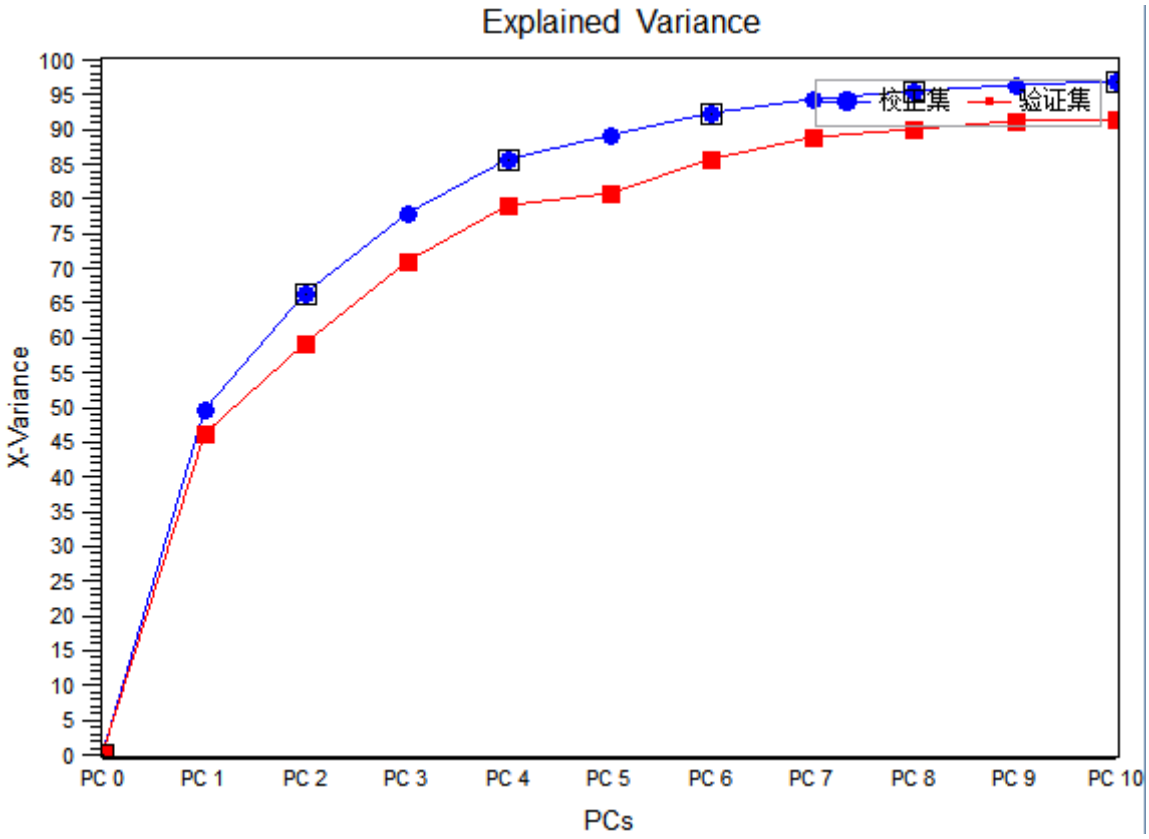
该图的初始状态如下图所示。



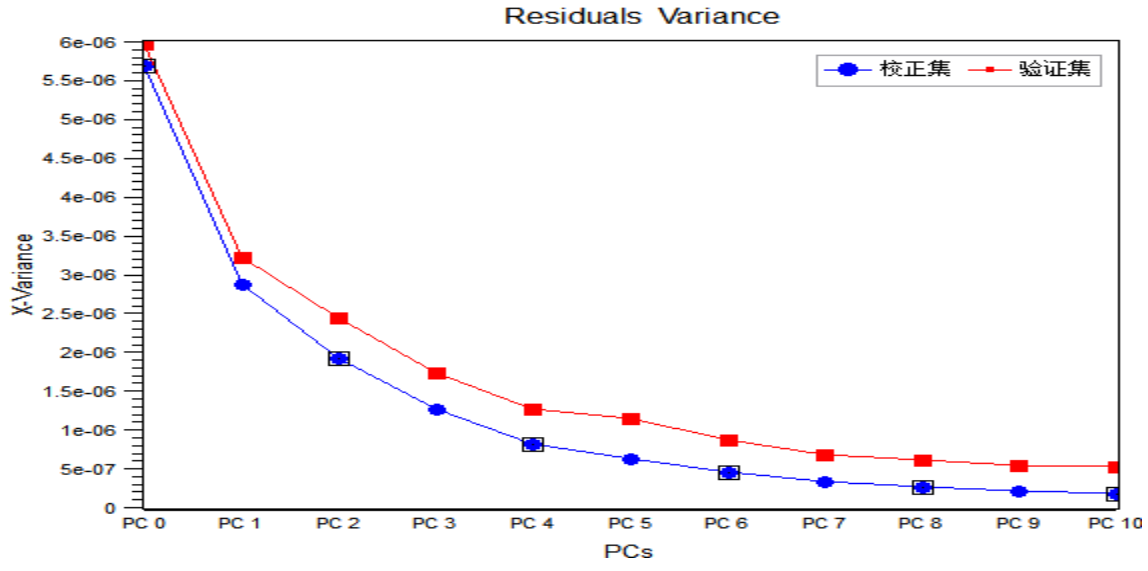
图中所显示的主成分数，与参数设置有关。除图形的基本工具外，在图形工具栏中同时增加如下图所示的功能。



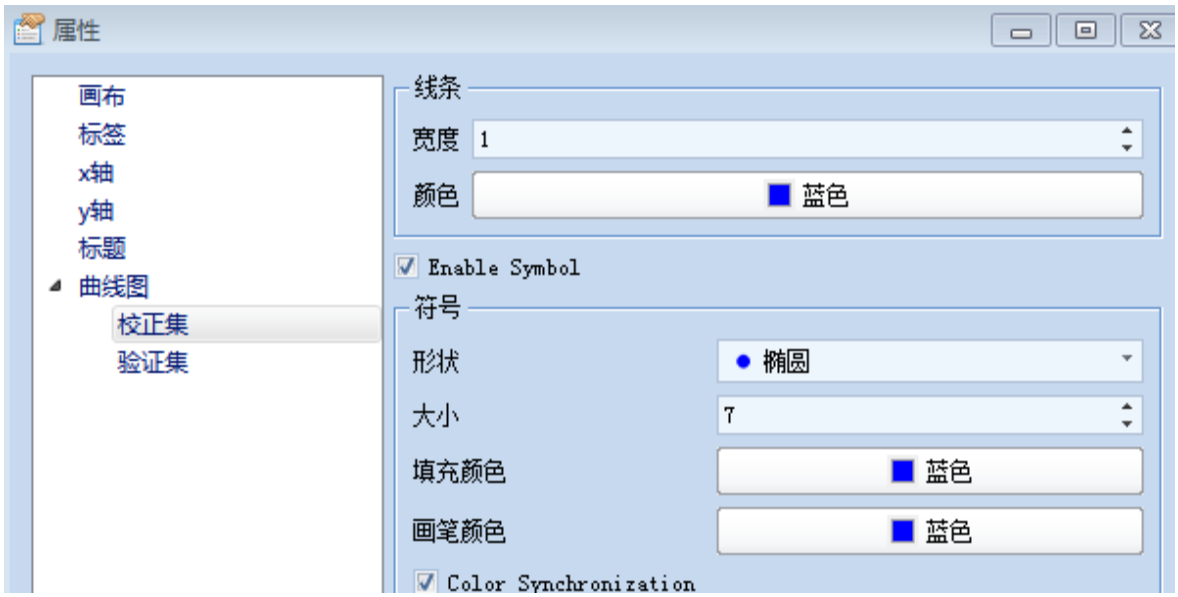
若同时选中校正集和验证集，则得到如下图所示的结果。



若将被解释信息的选择变成残差，则得到如下图所示的结图形果。



此外，若用户同时选中校正及与验证集，则图形属性修改中亦将同时显示，以方便修改各自的图形信息，如下图所示。



（二）图形解释

该图主要表述不同的主成分解释原始数据变化的能力，其计算已 12.2.5.中描述。好的结果是以简单模型(尽可能少的主成分数)便可使得残差方差达到 0，被解释信息尽可能大。

在解释 PCA 模型结果前，需先对模型质量进行有效评价。然而要达到这一目标，首先需确定主成分数，并知道这些主成分究竟包含多少信息，然后需要判别并找到奇异值，且这二步有时需多次操作才能获得理想的模型。

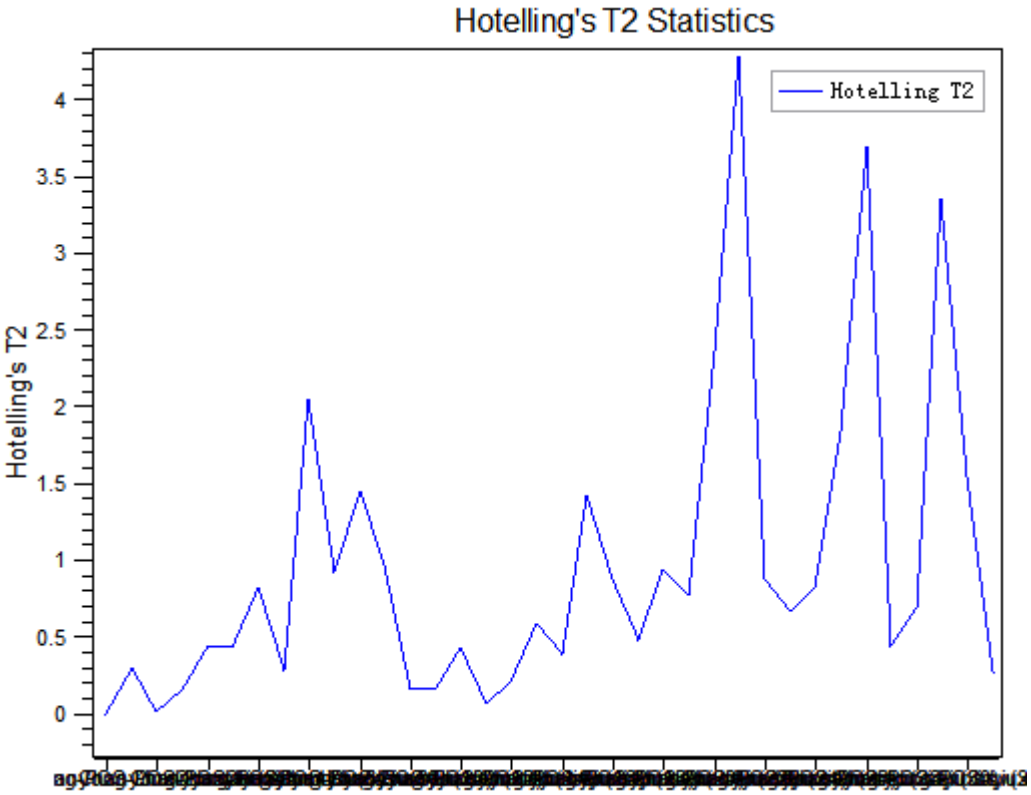
总残差与被解释方差表征模型拟合数据的好坏程度。具有更小总残差方差(如接近于 0)，或者更大的被解释方差(如接近 100%)的模型，意味原始数据变化被解释得越好。理想情况是在最简单的模型中以尽可能少的主成分数，便可使模型残差方差尽可能达到 0。

12.2.7.4. Hotelling's T2

具体操作步骤等内容不再赘述，用户可参考 9.2.1.1.，以及 12.2.7.2.部分。

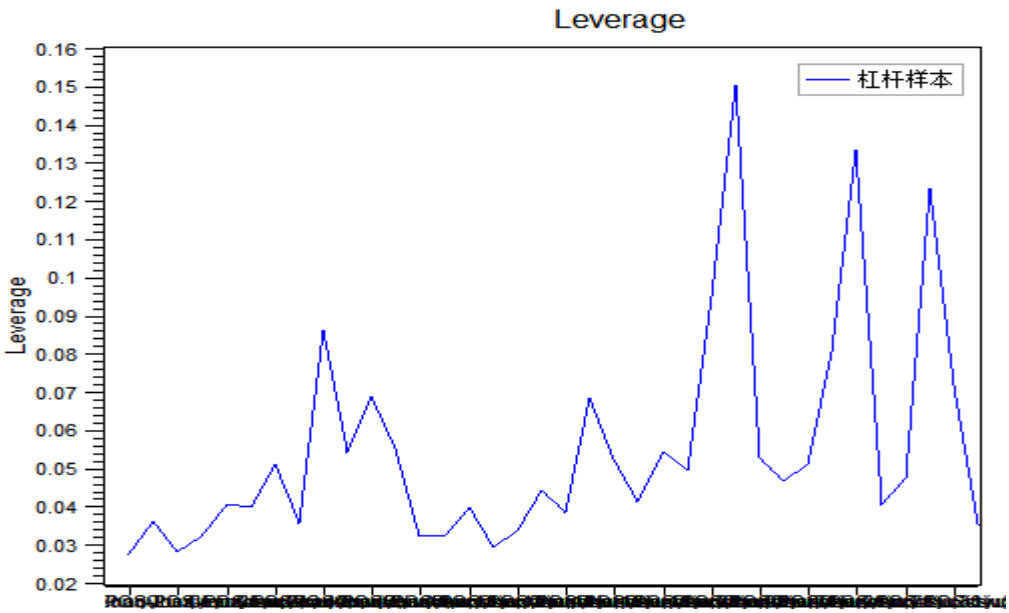
（一）结果介绍

该图结果的显示可从 T2 值和杠杆值中选择，图形的初始状态如下图所示。



图形工具栏中增加的功能如下图所示。

若用户选中杠杆指，则可得到如下所示的图形，显示样本的杠杆值。

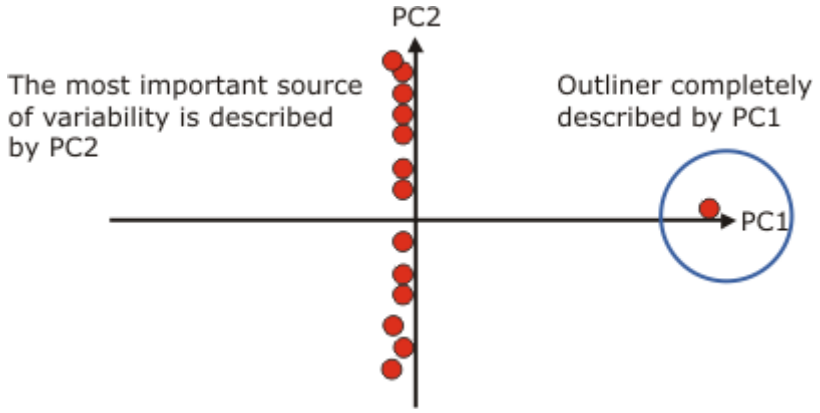


用户可选择不同的主成分数，以展示其不同的 T2 值和杠杆值图形。在任一主成分数下，通过点击建议主成分数按钮，可返回到程序推荐的结果图形。

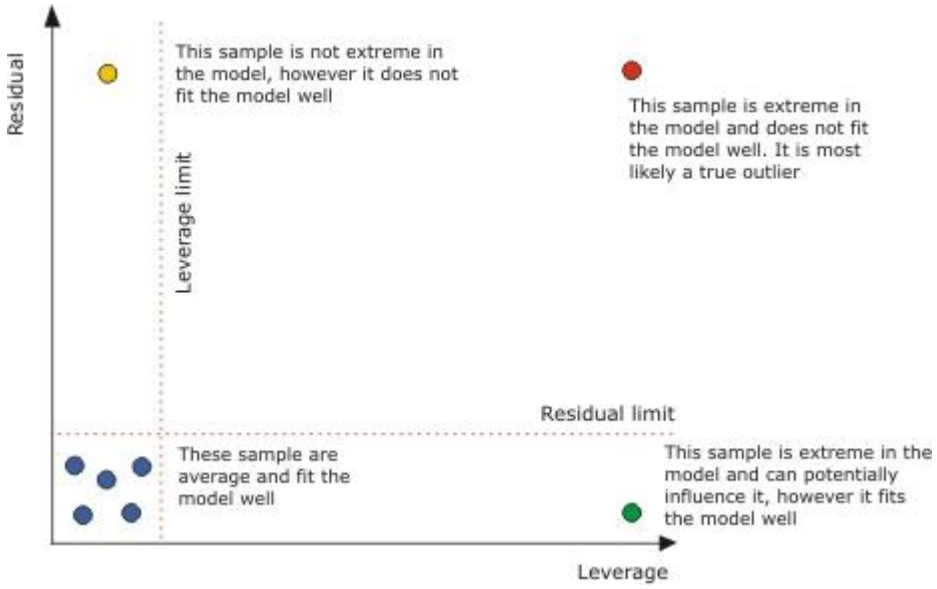
（二）图形解释

奇异值的检测是多变量分析的重要内容，PCA 计算所得到的得分图，残差或杠杆值等，可对奇异值进行有效的检测，如基于 Hotelling’s T2 椭圆。

奇异值是指与其他样本差异明显，没有被模型很好解释或者极大影响模型的样本。如下图则清楚地显示了奇异值对模型的影响，即第一主成分描述奇异值，而第二主成分则描述数据中最重要的变化。



从得分图中来看，若某样本与其他样本相比，分布特别离散，距离很远，则该样本很可能是奇异样本；样本残差越高，则同样意味着该样本没有被模型很好地解释，是可能的奇异样本；杠杆值则量测被投影样本到模型中心的距离，样本的杠杆值越高则其对模型的影响越大，它们不一定就是奇异样本，但肯定是强影响样本，而具有高残差和高杠杆值的样本，则是强影响奇异样本，这样的样本对构建稳健可靠的模型是非常不利的，好在基于样本影响力图形，可很容易发现这些奇异样本，如下图所示。



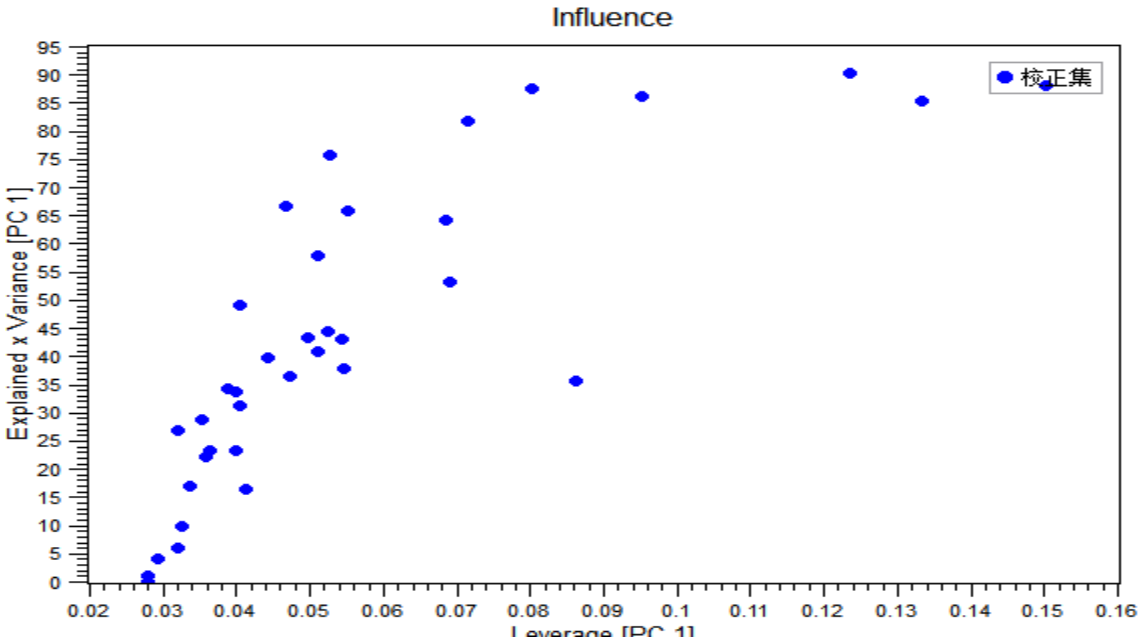
在得分图中，投影样本以不同的颜色表示，可通过样本分组、分布趋势以及散度等比较其与校正样本的差异性和相似性；影响图则检测某些被投影的样本没有被原始模型很好地解释；Hotelling's T^2及其阈值则可告诉使用者哪些被投影与模型或原始样本存在差异。

12.2.7.5. Influence

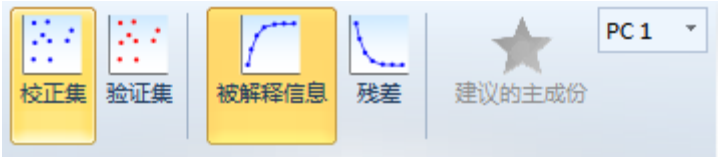
具体操作步骤等内容不再赘述，用户可参考 9.2.1.1.，以及 12.2.7.2.部分。

（一）结果介绍

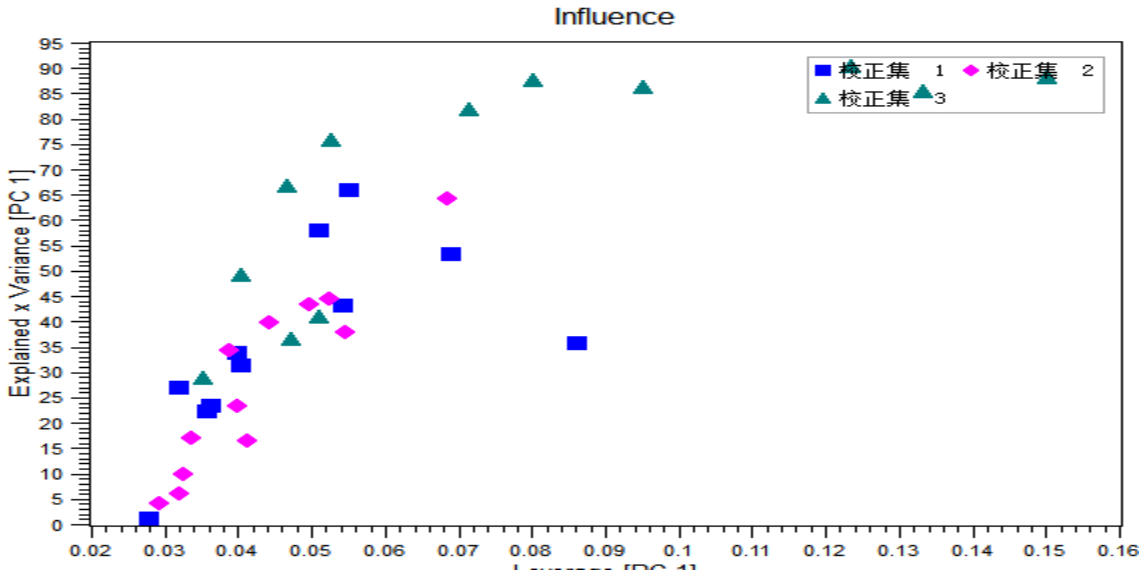
初始状态图形如下图所示。



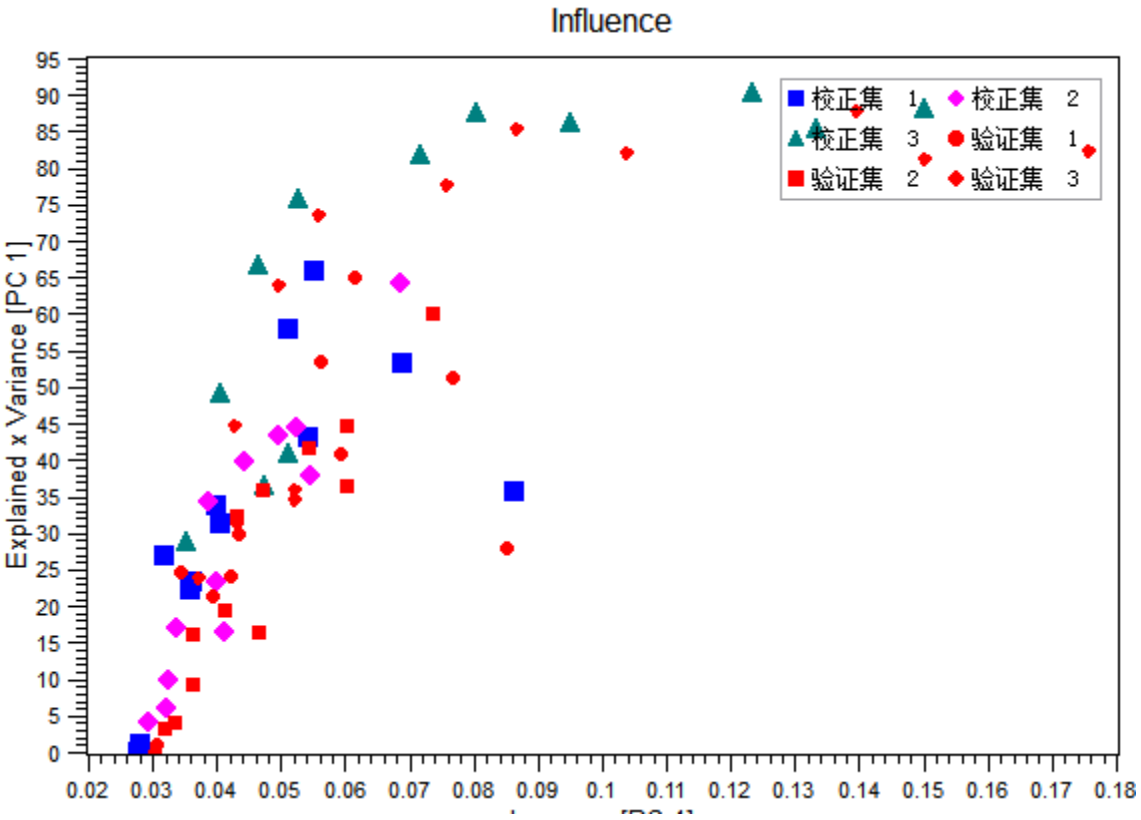
图形工具栏中增加的功能如下图所示。



上图工具栏的介绍，不再赘述，用户可参考 12.2.7.2 – 12.2.7.4 中各部分。该图的属性修改，用户亦可根据不同类别的样本，分别修改属性，可得到如下图所示的结果。



若再选中验证集按钮，则可得到如下图所示的对比图。



（二）图形解释

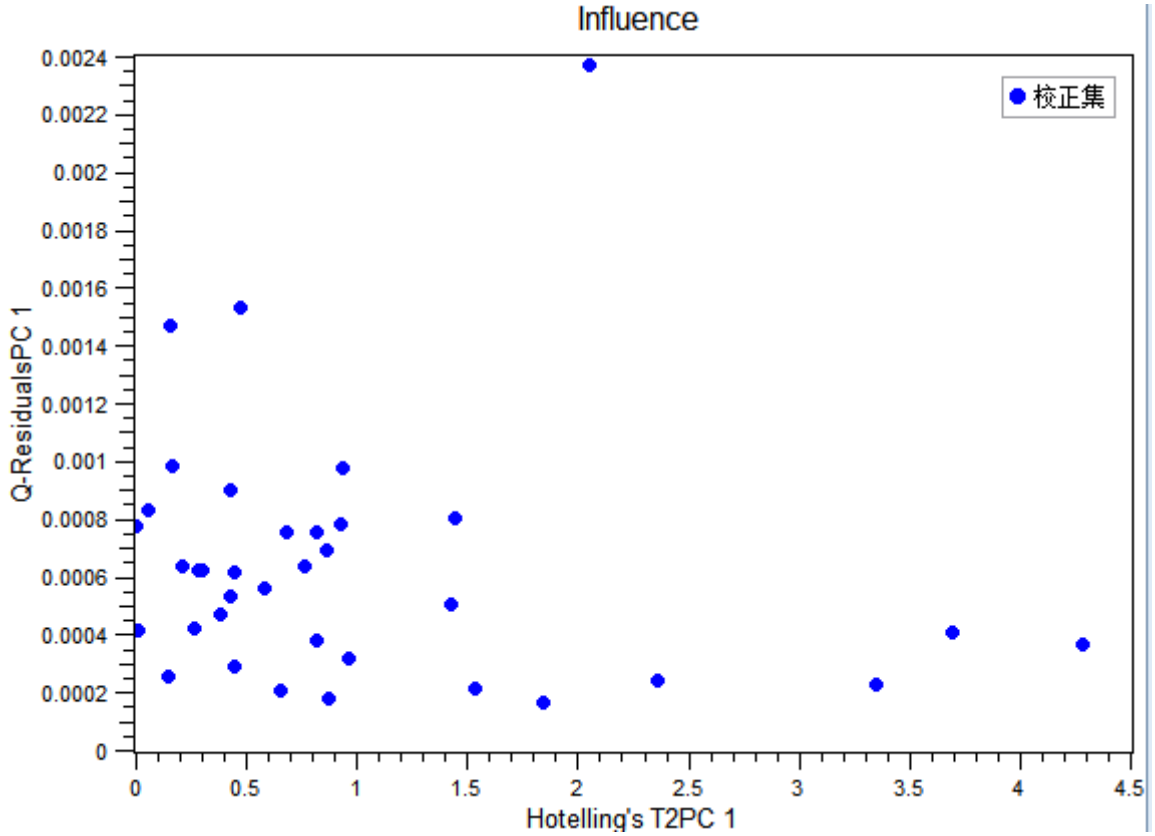
该图对于检测奇异样本、强影响样本和危险样本(奇异值)非常有效。位于图形上方，残差方差较大的样本很可能是奇异样本；位于图形右侧，杠杆值大的样本则是可能的强影响样本。强影响性样本并不一定是奇异样本；而那些残差方差和杠杆值均较大的样本，则是危险的奇异样本，不能被模型很好地描述和表征。此时模型更多地用于描述特殊样本间的差异，而不是关注普遍样本间的共性。

12.2.7.6. Influence2

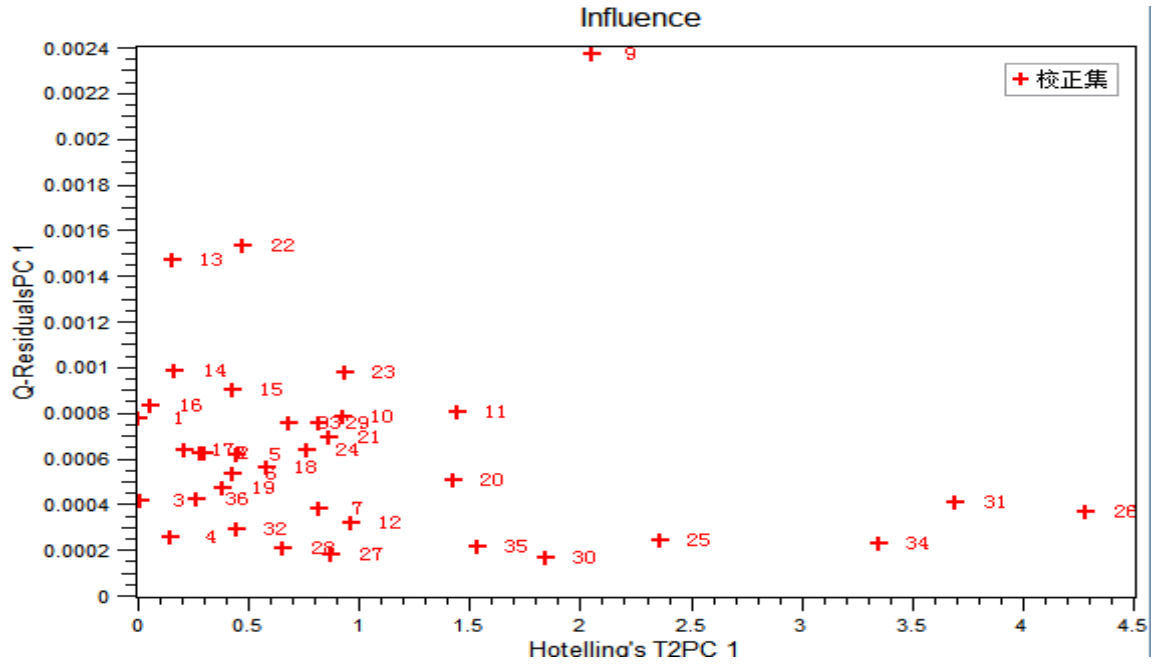
具体操作步骤等内容不再赘述，用户可参考 9.2.1.1.，以及 12.2.7.2.部分。

（一）结果介绍

初始状态图形如下图所示。

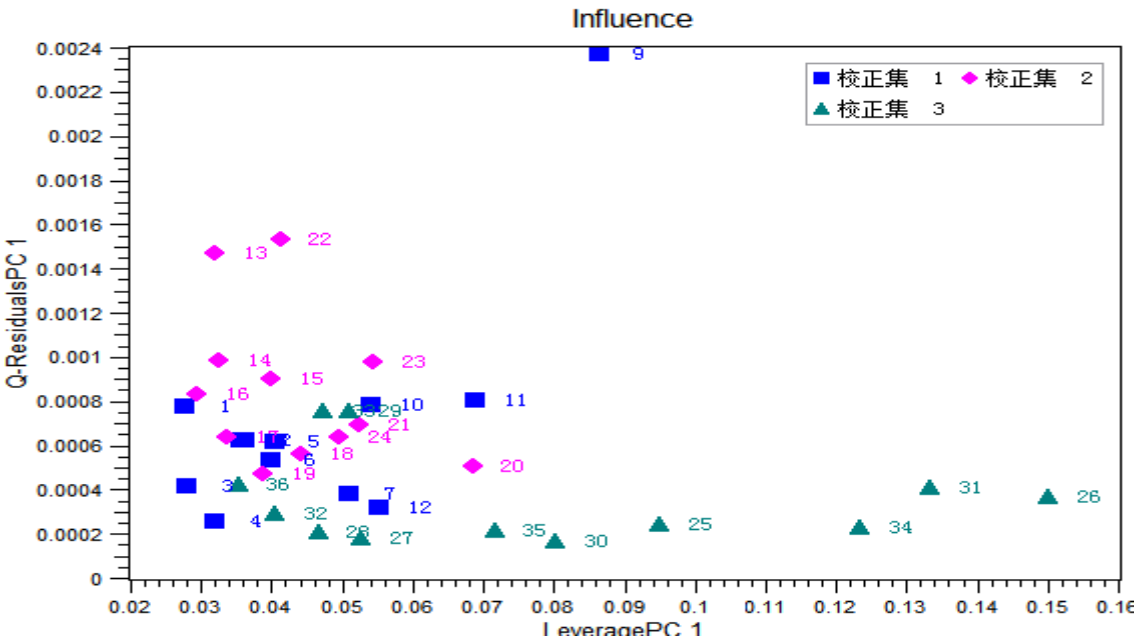


用户可修改图形属性，添加样本标注，得到如下所示的图形。

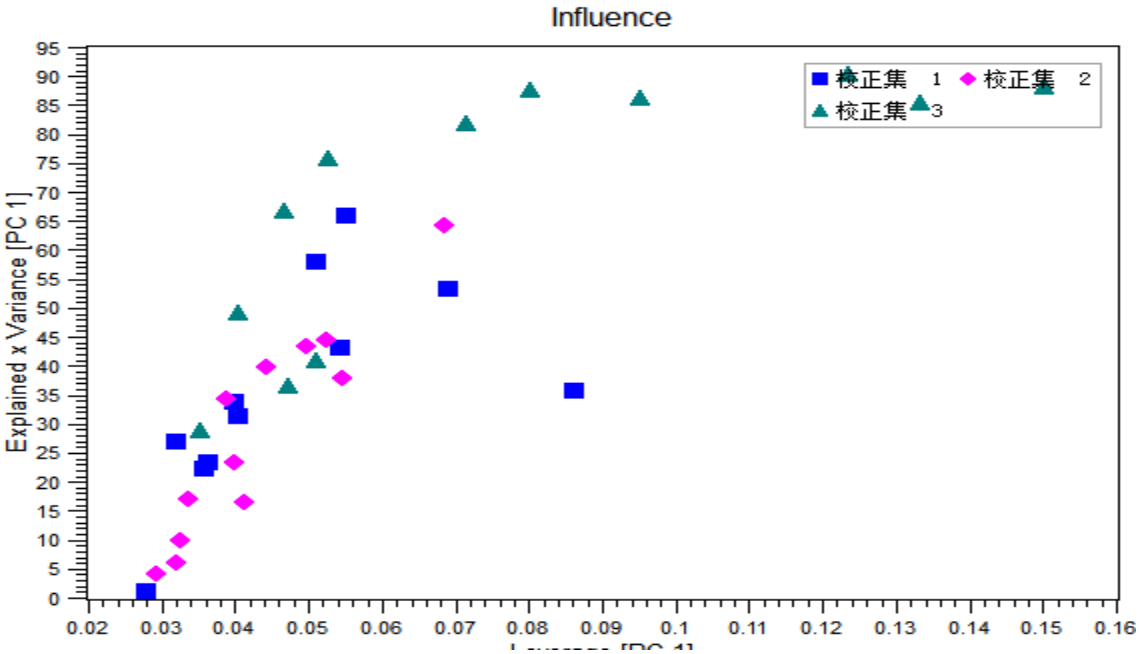


图形工具栏中增加的功能如下图所示。

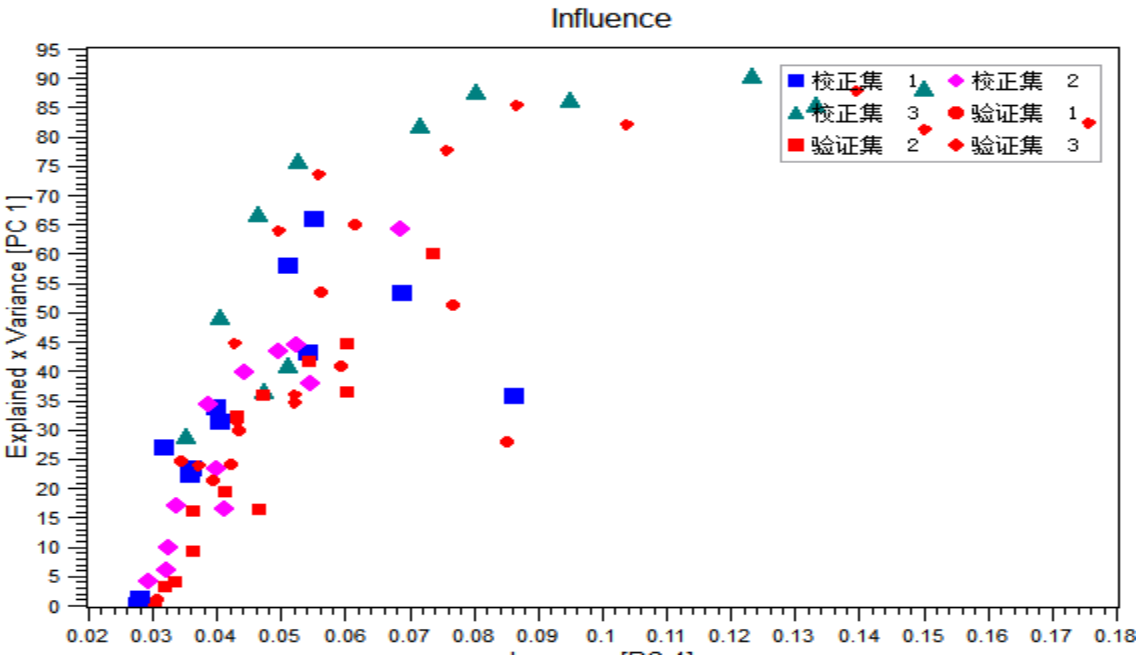
上图工具栏的介绍，用户可参考 12.2.7.2–12.2.7.4 中各部分。若选中杠杆值并修改图形属性，则得到如下图所示的图形。



该图的属性修改，用户亦可根据不同类别的样本，分别修改属性，可得到如下图所示的结果。



若再选中验证集按钮，则可得到如下图所示的对比图。

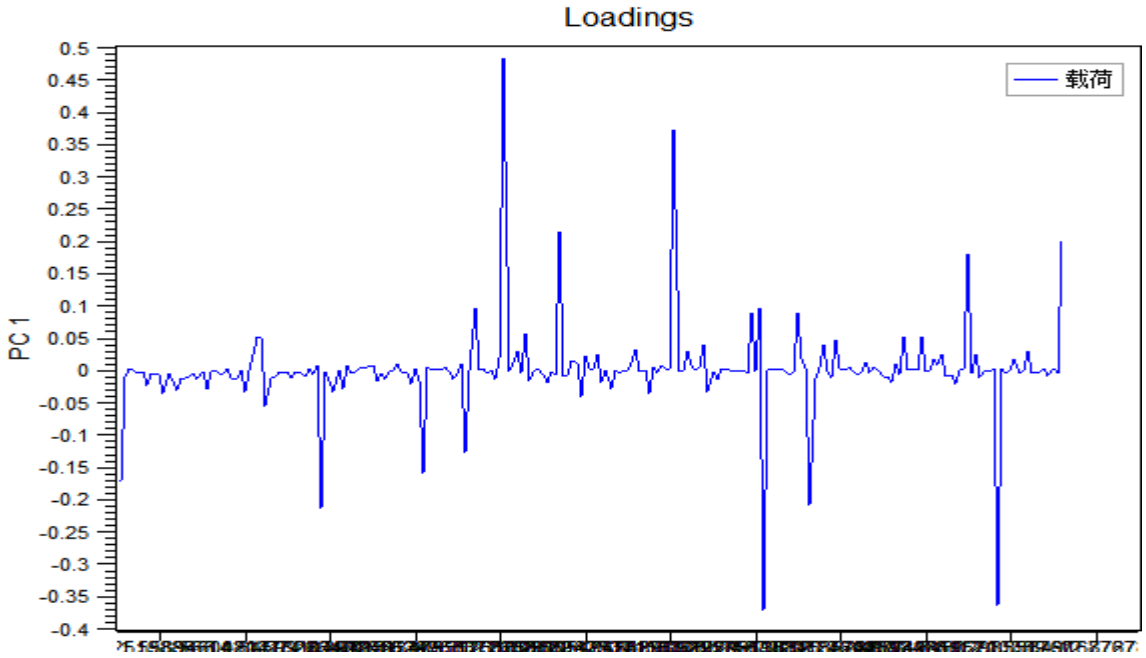


12.2.7.7. Loadings

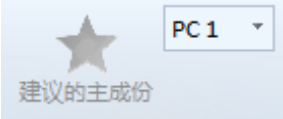
具体操作步骤等内容不再赘述，用户可参考 9.2.1.1.，以及 12.2.7.2.部分。

（一）结果介绍

初始状态图形如下图所示。



图形工具栏中增加的功能如下图所示。这些功能可以方便用户选择不同的主成分数，并得到对应的载荷图形。



（二）图形解释

PCA 载荷图是检测重要变量的有效手段。如前介绍 Bi-plot 时所述，该图最好与得分图联合起来使用，以达到最好的解释性效果。若二主成分解释大部分的数据方差，则他们在得分图中越靠近，或者在同一象限内，靠近同一直线上，则他们具有越高的正相关性；相反若变量在反方向象限内，则可能具有负相关的趋势；而靠近图形原点中心的变量，则在该图中难于被解释。

载荷是变量与当前主成分间的夹角余弦，该夹角越小，即变量与主成分间的关联性越高，载荷则越大。二变量间 $\mathbf{x}_1$ 与 $\mathbf{x}_2$ 的相关性 r 可定义为：

$$r(x1,x2) = Cov(x1, x2) / (x1 \times x2)$$

上式中，Cov 为变量间的协方差。若某变量与第一主成分间的角度接近于 0，则第一主成分完全描述该变量；同理若某变量与第二主成分间的角度接近于 0，则第二主成分完全描述该变量；若变量 1 与变量 2 间的角度为 90°，则表示变量 1 和 2 完全不相关；若某变量与第一主成分间的角度大于 180°，而与第二主成分间的角度大于 90°，则该变量同时与第一、第二主成分负相关；若某变量位于第一主成分与第二主成分的交叉点处，则该变量不能被这二个主成分很好第表达。

主成分的重要性可由其能解释的原始数据方差来表征。载荷基于变量贡献和相关性描述数据结构，每个变量在每个主成分方向上均有一个得分，该值即为其对应主成分的贡献，以及包含在该变量中数据变化。

在载荷图中，首先检查载荷值大的变量。获得变量间的相关性，需研究载荷空间中变量间的相对位置即变量越靠近在一起，则越相关，如二变量沿相同主成分具较大载荷，则意味着它们间的夹角越小，即二变量越相关；若二载荷符号相同，则它们间正相关，反之亦然。

研究变量间的变化与差异，一个很直观的问题是究竟哪些变量对数据间的这种变化起着关键作用，即:

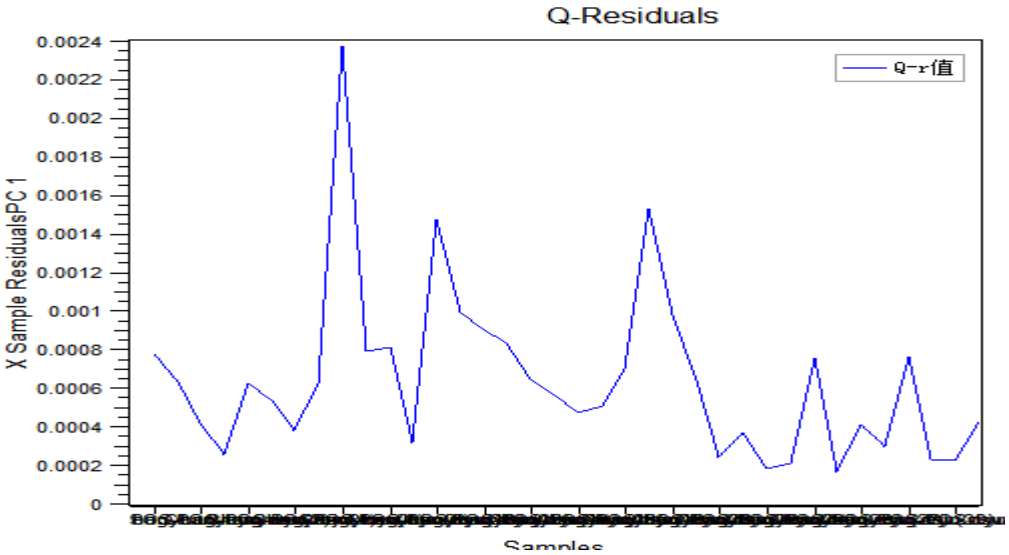
- 使用哪些充分描述样本必不可少的变量?
- 哪些样本特别相似度?
- 数据中有哪些特别的样本组?
- 样本间的这种分布模式究竟表示什么意义?
- 通过使用 PCA 法分解数据便可回答上述问题。

12.2.7.8. Residuals

具体操作步骤等内容不再赘述，用户可参考 9.2.1.1.，以及 12.2.7.2.部分。

（一）结果介绍

初始状态图形如下图所示。



图形工具栏中增加的功能与 12.2.7.7.雷同，可参考。

（二）图形解释

Q 统计量量测新样本偏离已知模型程度，是模型外部数据变化的量度。其计算公式为，$Q_i = \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} = \mathbf{x}_i^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_a\mathbf{P}_a^{\mathrm{T}})\mathbf{x}_i$，其中 $\mathbf{e}_i$ 为残差矩阵 **E** 的第 *i* 行，而 $\mathbf{P}_a$ 为前 *a* 个主成分载荷，**I** 则为单位矩阵。

原始样本与变量模型空间组份由其到模型空间的投影表示，二者间的差异则是样本或变量残差；样本残差变化则是所有模型成分的残差平方和；变量残差方差则是所有模型成分残差平方的均值；总残差方差则是所有变量残差方差之和；被解释方差则是残差方差的补充，指可被模型解释的模型方差站数据总方差的百分比；总被解释方差则是原始数据中的变化被模型解释的量。

ⓘ 变量的残差方差越小，则意味着其被对应的模型解释得越好，而在前三个最重要主成分，甚至所有主成分中均具有较大残差方差的变量，与其他变量则具有较小或中等程度的关联性。若某些变量与其他变量相比，在最重要前面 3-4 各主成分(或所有主成分)中均含有更大的残差方差，则这样的变量应被剔除重新计算，新的模型很可能更容易被解释。

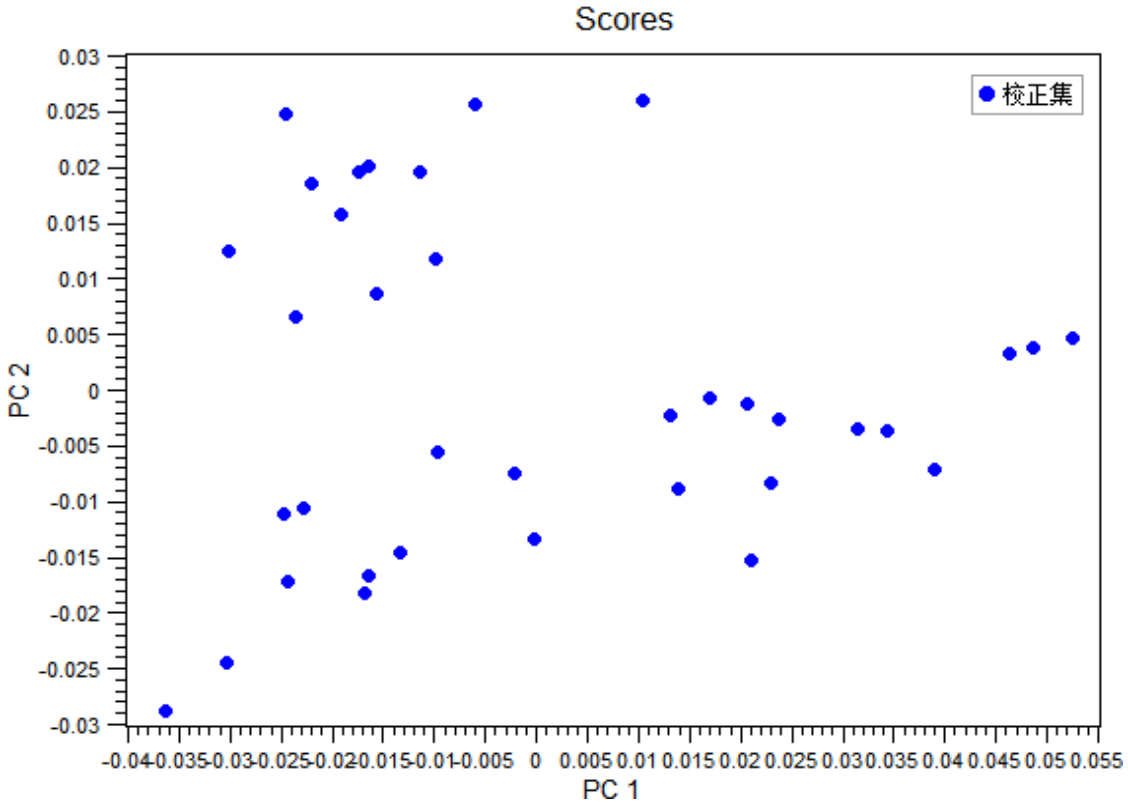
ⓘ 校正方差基于校正集数据的拟合计算获得，而验证方差则基于未参与建模的测试集数据来计算。若二者间的差异很大，则需要考虑这些数据是否具有足够的代表性。奇异值则是产生较大残差方差的原因。

12.2.7.9. Scores

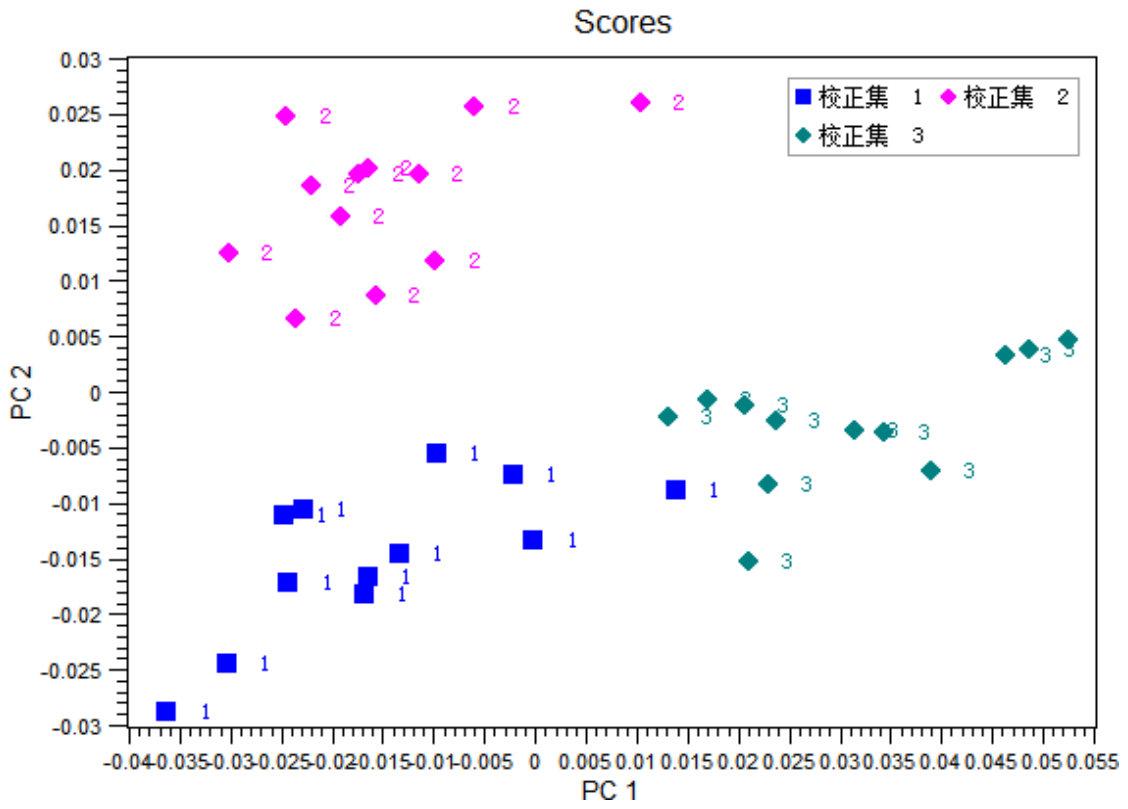
具体操作步骤等内容不再赘述，用户可参考 9.2.1.1.，以及 12.2.7.2.部分。

（一）结果介绍

初始状态图形如下图所示。

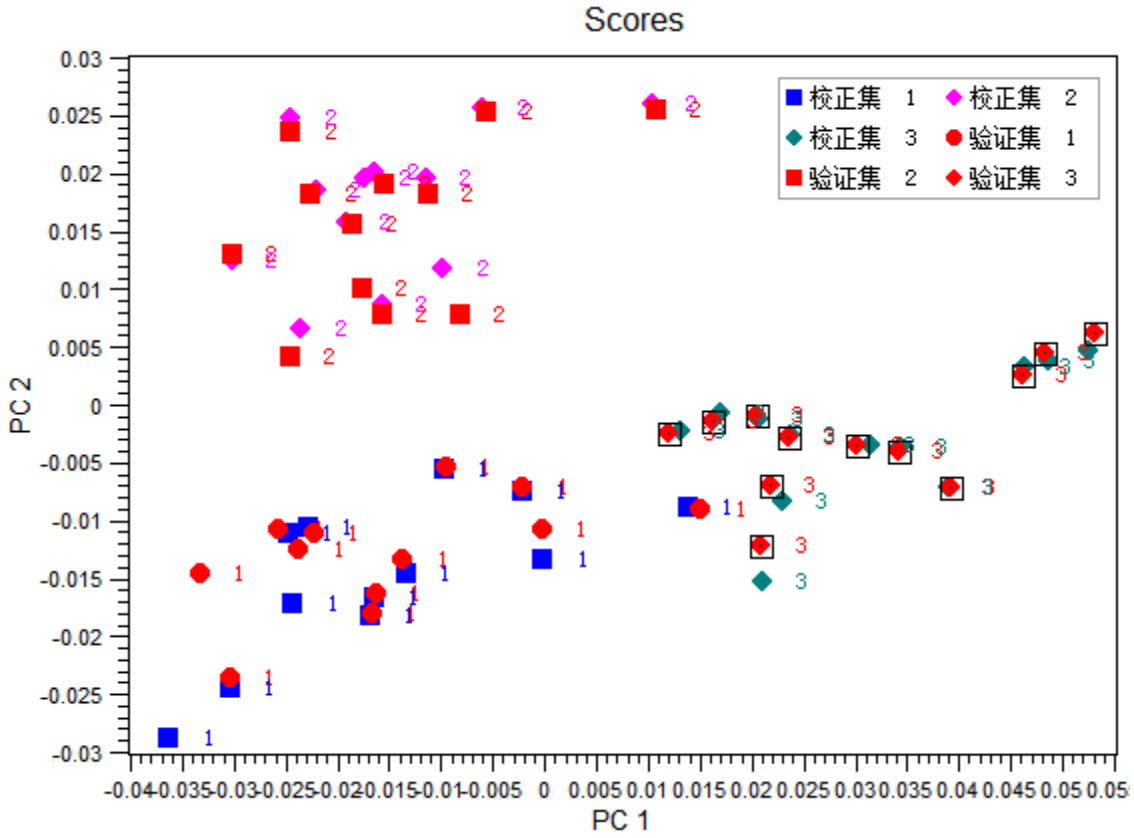


用户可修改图形属性，根据样本类别添加样本标注，得到如下所示的图形。



图形工具栏中增加的功能如下图所示。

上图工具栏的介绍，用户可参考 12.2.7.2–12.2.7.4 中各部分。若再选中验证集，则得到如下图所示的图形，很好地比较结果。



（二）图形解释

得分图给出样本分布模式的信息。在不同主成分所构成的得分图中，PC1-PC2 所提供的信息最丰富。样本在得分图中靠得越近，则表示在当前主成分下越相似，反之亦然，因而基于得分图，可解释样本间的相似性与差异性。若 PCA 分析时同时涉及校正集和验证集或预测集，则可从得分图中，综合评价验证机或预测集是否可覆盖校正集的整个模型空间，从而得到结果可靠性的信息。

如前所述，通过对得分图的研究，亦可达到此目的。此外，理解得分与载荷图，首先应明白其值可正可负，沿坐标轴分布的方向性，主成分通过得分和载荷将样本和变量建立很好的连接。对得分和载荷的理解，主要包括如下几个方面：

- 若某变量的载荷值很小，则应忽略该变量而重点关注其他载荷值更大的变量，不可用于数据解释(无论其值为正或负)。
- 样本的正得分越高，则其值对应于正载荷变量亦越大，反之亦然。
- 样本负得分越低，则其正载荷变量中的对应值越小。
- 变量载荷越大，则样本值随得分增加越快。

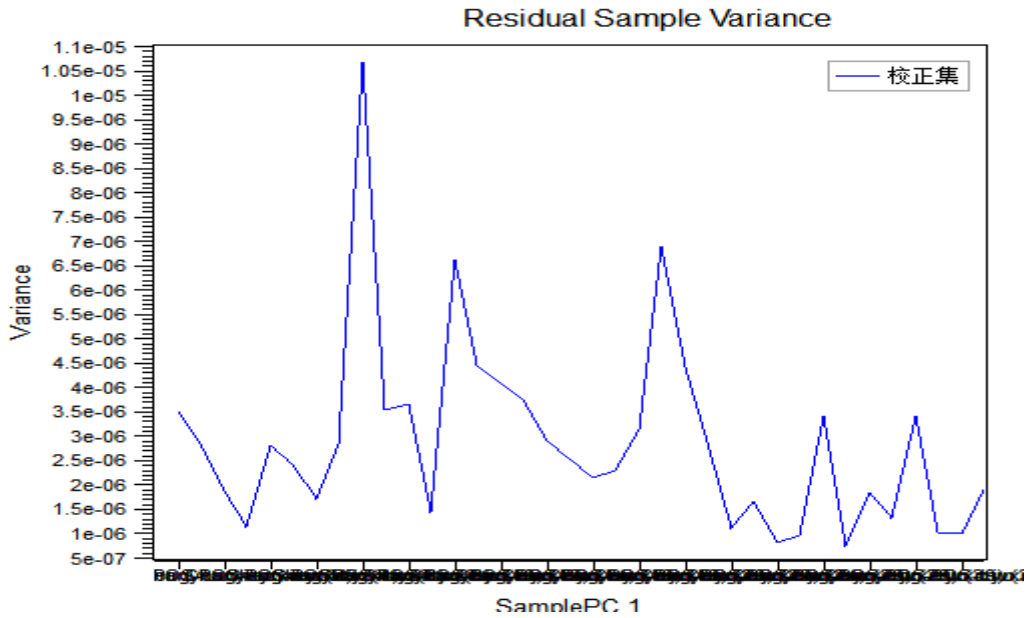
简言之，若在某主成分方向，样本得分与变量载荷符号相同，则与平均变量相比，该样本与变量越相关，反之亦然；样本得分与变量载荷的值越大，则他们间越相关。

12.2.7.10. X Sample Explained Variance & Residuals

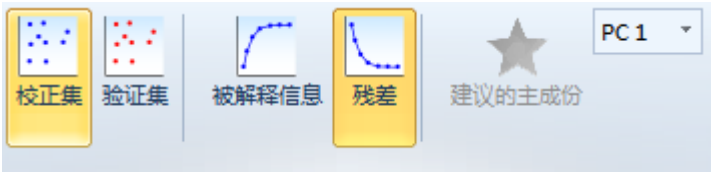
具体操作步骤等内容不再赘述，用户可参考 9.2.1.1.，以及 12.2.7.2.部分。

（一）结果介绍

初始状态图形如下图所示。



图形工具栏中增加的功能如下图所示。



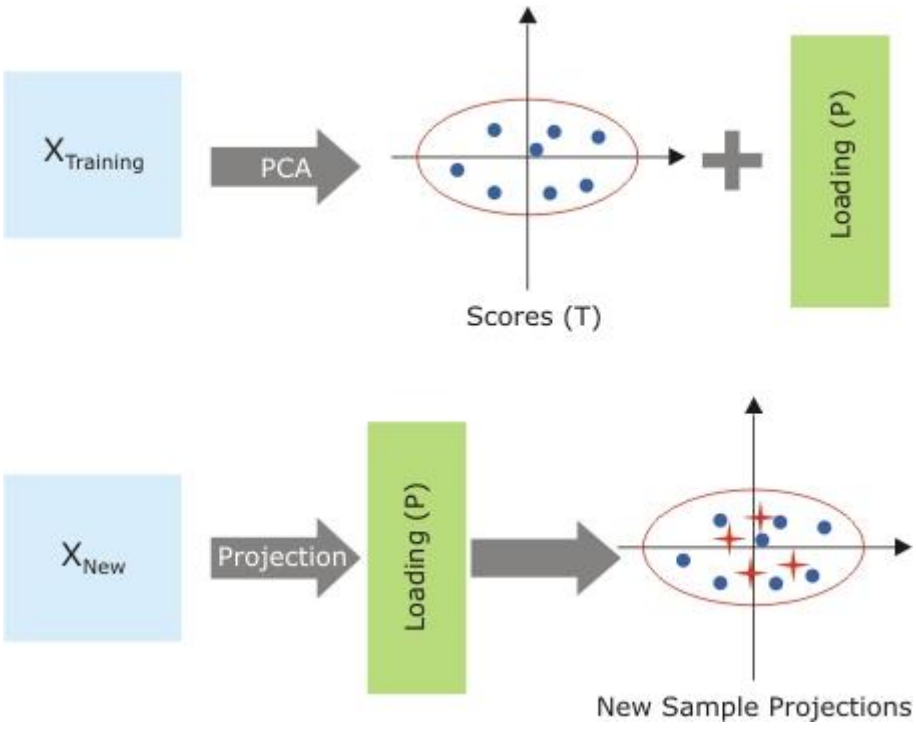
上图工具栏的介绍，不再赘述，用户可参考 12.2.7.5。改图属性修改、加入验证集、选择被解释信息或残差，以及改变主成分等得到的结果，亦不再介绍，用户完全可参考上述部分操作。

（二）图形解释

请参考 12.2.7.3.章节。

12.2.8. Unknown Data Prediction 节点

若用户在 12.1.2.3.中，并没有选择预测集数据，则不产生本部分结果。PCA 中投影分析等同于回归分析中的新样本预测，如下图所示。



典型地，PCA 投影分析可用于解决如下问题：

a) 产品生产中原材料供应商的变化是否导致产品特征发生了变化：将供应商变化后得到的新产品样本数据，投影到使用过往原材料生产所得样本数据所构建的 PCA 模型上，便可得到原材料变化后产品是否发生改变的结论。

b) 生产设备维修后所生产的产品，其质量是否发生了改变?

c) 如何比较由不同厂商 A 和 B 生产的产品样本，其相似性和差异性如何：将新样本(如厂商 B)投影到参考样本(如厂商 A)所构建的 PCA 模型上，以查看他们在得分图上是否有重叠来判断。

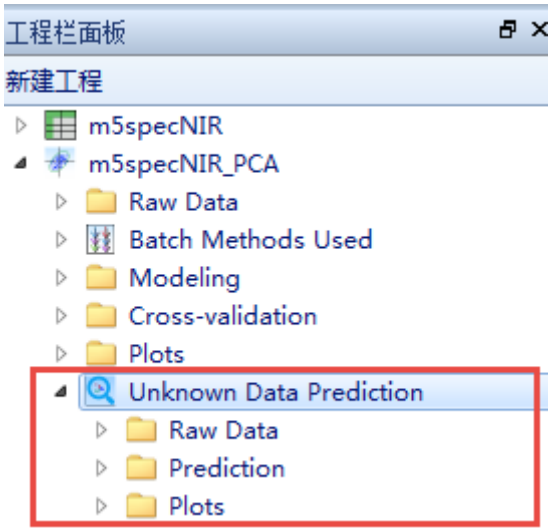
d) 一年以前构建的模型，是否还能分析或描述最新样本：同样将新样本投影到旧模型上，通过查看样本分布的漂移，平均得分、散度的增加，以及残差的增大来判断模型的适应性。

新样本投影结果的解释与上述 PCA 模型的解释雷同，区别在于固定载荷为已知 PCA 模型的结果，而新样本则投影到该载荷，产生新的得分，计算方式没有其他变化，如图所示。通过投影所得到的载荷、方差、残差、杠杆值和统计量 Hotelling's T^2等则表征投影样本。

比较投影方差曲线与校正或验证曲线，若它们很相似，则模型得到投影样本的进一步确认。

12.2.8.1. 预测结果概述

若在 12.1.2.3.中，用户选择了预测集数据，则 PCA 建模完成后，将得到如下图所示的预测结果节点文件夹。



上述各节点文件夹的意义如下表所示。

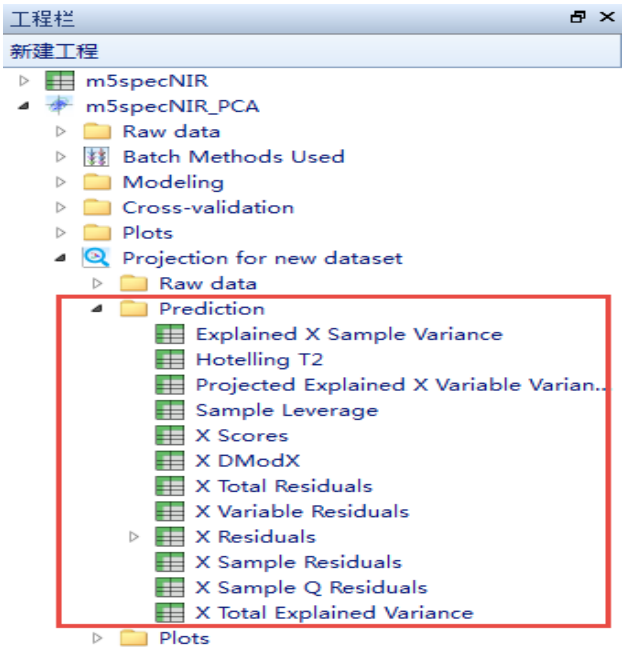
序号	节点名称	说明
1	Raw Data	建模时所选数据的一个副本，即将建模时所用的数据重新复制一份置于结果文件夹下，以保持节点的完整性。
2	Prediction	预测集得到的数据结果。
3	Plots	图形结果的节点文件夹，主要是对 Modeling 和 Cross-validation 节点文件下数据的绘图。

12.2.8.2. Raw Data 节点

如前所述，不再赘述。

12.2.8.3. Prediction 节点

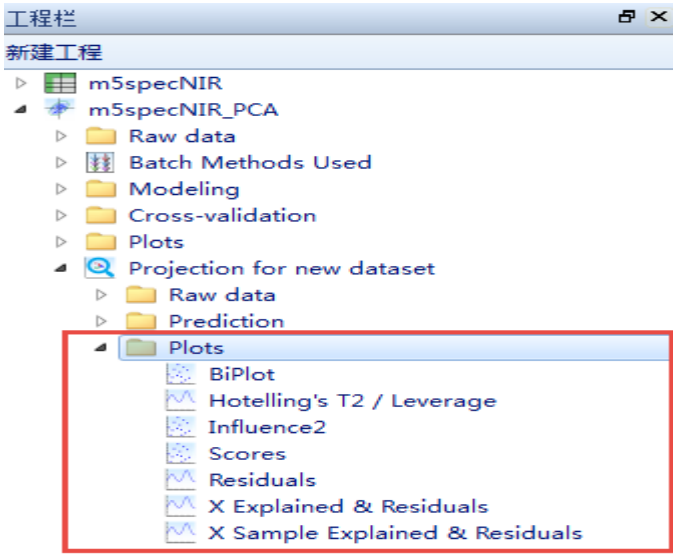
类似上述 12.2.5.和 12.2.6.的结果，该节点文件夹包含预测集所得到的各种结果。具体所包含的节点如下图所示。



关于各节点下结果的详细介绍，在此亦不再赘述，请参考表 12.2.5.，其差异在于本处的结果是基于上述已经构建的 PCA 模型，针对预测集得到的结果。

12.2.8.4. Plots 节点

类似上述 12.2.7.的结果，该节点文件夹包含预测集所得到的各种图形结果。具体所包含的节点如下图所示。



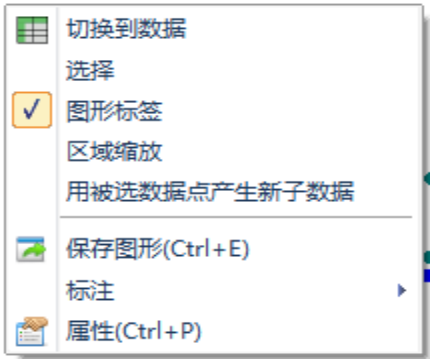
上述图形所对应的 PCA 分析结果数据，以及各图形的操作和意涵，其来源亦请参考 12.2.7.1.，不再赘述，需要了解的是本节结果对应预测集数据。

12.2.9. 产生新数据

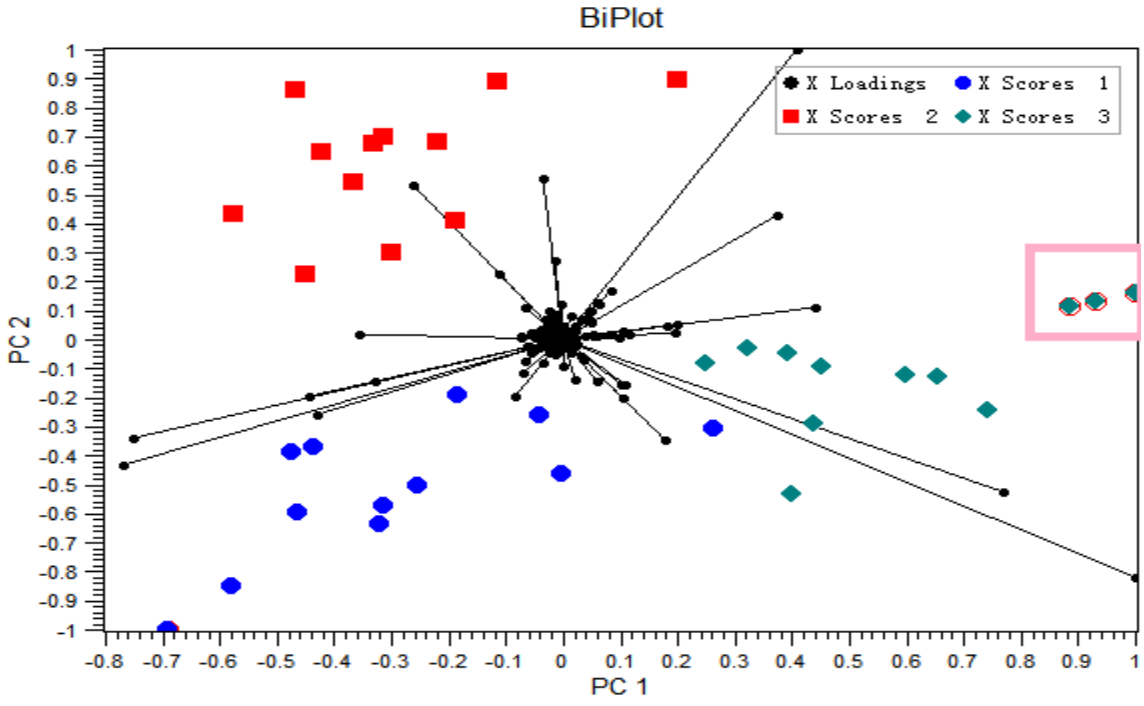
在 9.4.3.与 6.3.1.章节已经介绍到，用户可从基本数据表或图形中选择样本或者变量产生新的子数据。

前面已经介绍了奇异值的判别方法。那么若发现奇异值后，如何快速剔除奇异样本并重新建模呢？本节主要介绍这部分的内容。

如下图所示，在建模所得到的图形结果中，右键点击任一图形节点，均可产生如下图所示的菜单功能，其中包括用被选数据点产生新子数据的功能。



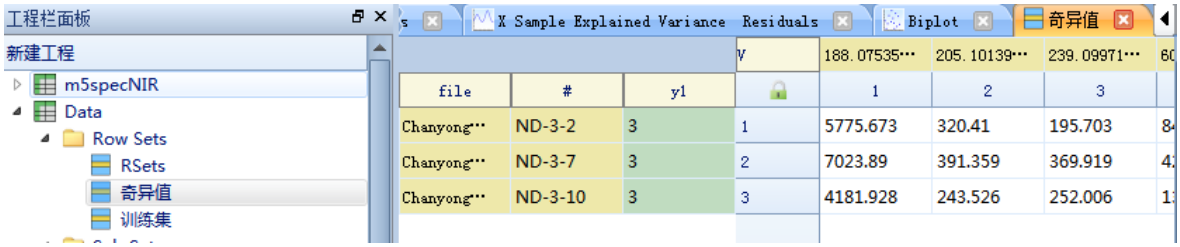
通过矩形标记功能选择目标数据点(图中以方框标记)，如下图所示。



点击上述产生新数据功能，即可出现如下图所示的界面。该界面完整列出工程导航栏中的数据，以便用户选择上图中的所选择的数据点的作用范围，即用在哪些数据上。

上图中左侧为完整数据列表，右侧上方为添加或删除数据按钮，可同时添加多个数据，而右侧下方则是选择子数据的产生方式，即是将被选数据点作为奇异值剔除，还是将未被标记的样本作为需要被保留的数据，用于重新计算和建模。

如上图中的数据，若想同时将被选的样本作为被剔除的样本，产生奇异值数据，并将未被标记的样本作为训练集，则同时勾选二复选框，得到如下图所示的结果，在工程导航栏中同时产生二个目标数据矩阵，以方便用户再次分析。



需要注意的是，程序对所有工程导航栏中出现的数据均建立了唯一的标识符，每个样本的

标记符是不同的，当用户从工程导航栏选择用于产生子数据的母数据时，若该数据不是用于建模的数据，将自动使添加数据按钮失效，从而不可添加该数据，如下图所示。

实因当前被分析的数据为 Data 下的子数据，选择 m5specNIR 下的任何数据，显然不会被处理。然而，若用户选择 Data 数据则是可行的，尽管当前被处理的数据是该数据的子数据，显然亦应包含在该数据中，因而可用于产生奇异值或训练集数据。当然，若用户选择的是变量，则同样可产生子数据，只是此时所得到的奇异值或训练集数据是对变量方向划分得到的列划分数据而已。

通过此功能可极大地方便子数据的产生，并重新建模，得到更理想的结果。

12.3. HCA 法

本软件同时提供非分层聚类和分层聚类二类方法，分别包括 K-means 和 HCA 等方法。HCA 法是受到广泛应用的探索性分析方法，其结果主要是以系统分层聚类图的形式表示，挖掘数据信息。此类聚类分方法通常基于样本间的相似性量测，可以无监督方式将样本聚成不同类别。比较而言，该法较 K-means 和 K-medians 消耗内存，在分析大样本数据时没有优

势。

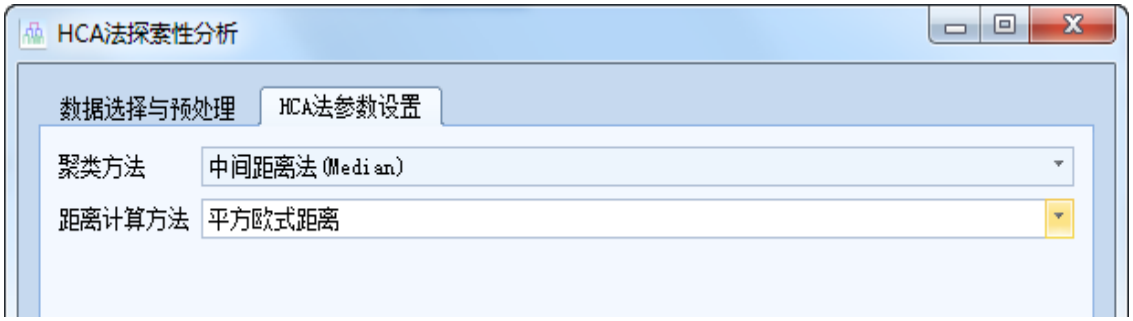
HCA 法采用不同的连接方法，以及距离计算方法产生样本聚类。需要注意的是，并不是所有的距离量度方法均满足三角不等式规则，即三角形任意二边之和大于第三边，此时 HCA 系统层次聚类图可能产生变形。

在 HCA 分析前，比较好的做法是先进行 PCA 等分析，初步了解数据结构与分布趋势。

12.3.1. 操作说明

基本操作步骤可参考 12.1.2.部分。

如上所述，HCA 法有二个特殊的地方，一是样本连接方法，另一个是距离计算方法，如下图所示。打开 HCA 建模界面后，有其参数设置的选项。



如下表详细介绍各样本连接方法的涵义。

序号	方法名称	详细说明
1	Single-linkage	最短距离法，使用类间最短距离定义样本聚类，即类与类之间的距离是两类间两两样品间的最短距离。该法趋向于产生较多聚类类别，分类效果有时不是特别理想。
2	Complete-linkage	最长距离法，使用类间最长距离定义样本聚类，即类与类之间的距离

		是两类间两两样品间的最长距离。产生更紧凑的聚类结果。
3	Average-linkage	类平均法，该法是上述二各方法的折中，以类间距离加权的方式计算。
4	Median-linkage	中间距离法，与 Average-linkage 法类似，但计算某类与其他类加权重心间的几何距离，实因最长距离夸大类间距离，而最短距离低估类间距离，从而使用介于两者间的距离计算。
5	Ward's method	Ward 离差平方和法，两类合并后，离差平方和值将增加，每次计算时选择合并使离差平方和增加最小两类，直至所有样品均归类为止。

如下表则介绍距离计算方法的涵义。

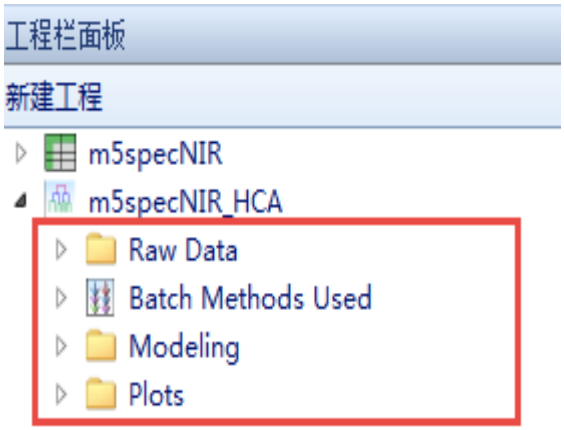
序号	**方法名称**	**详细说明**
1	欧氏距离	Euclidean Distance，使用最广泛的距离计算方法，其计算公式如下。 $$d = \sqrt{\sum_{i=1}^{N}(x_{i,1} - x_{i,2})^2}$$
2	欧氏距离平方	Squared Euclidean Distance，当某些变量占显著优势时，该法比较有效，其计算公式如下。 $$d^2 = \sum_{i=1}^{N}(x_{i,1} - x_{i,2})^2$$
3	标准化欧氏距离	Standardized Euclidean Distance，将数据进行标准化处理后计算所得到的距离。标准化方法如下式。 $$\mathbf{x}_{new} = \frac{(\mathbf{x} - \bar{\mathbf{x}})}{std(\mathbf{x})}$$ 可得如下距离计算公式。 $$d = \sqrt{\sum_{i=1}^{N}(\frac{x_{i,1} - x_{i,2}}{s_{1,2}})^2}$$

4	马氏距离	Mahalanobis Distance，实因计算欧氏距离时，需先对数据进行标准化，基于标准化后数据欧式距离的计算产生。 $$d = (\mathbf{x} - \bar{\mathbf{x}})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}})$$ 其中$\bar{\mathbf{x}}$和 **S** 分别为均值向量和协方差矩阵。若协方差矩阵为单位矩阵，则变为欧氏距离。 该变量与量纲无关，从而排除变量间相关性的干扰。
5	曼哈顿距离	City-Block Distance，又称城市街区距离，其计算公式如下。 $$d = \sum_{i=1}^{N} \lvert x_{i,1} - x_{i,2} \rvert$$ 以二维平面上的二点为例，该式变为。 $$d = \lvert x_1 - x_2 \rvert + \lvert y_1 - y_2 \rvert$$
6	明科夫斯基距离	Kowalski Distance，是对广义的欧氏距离，概括性描述多个距离量测，如下式所示。 $$d = (\sum_{i=1}^{N} \lvert x_{i,1} - x_{i,2} \rvert^{p})^{1/p}$$ 当 p = 1 为曼哈顿距离; p = 2 为欧氏距离; $p \to \infty$则为切比雪夫距离。 该距离量度的主要缺点是将个分量的量纲相同对待，其次是没有考虑到各分量分布的不同。
7	切比雪夫距离	Chebyshev Distance，该法起源于国际象棋中国王的走法，即国王每次只能往周围的 8 格中走一步，若需从 A 格走到 B 格，最少需要走的步数为，$\max(\lvert x_2 - x_1 \rvert, \lvert y_2 - y_1 \rvert)$。若扩展到多维空间，则如下式所示。 $$d = \lim_{p \to \infty} (\sum_{i=1}^{N} \lvert x_{i,1} - x_{i,2} \rvert^{p})^{1/p} = \max \lvert x_{i,1} - x_{i,2} \rvert$$
8	余弦相似度	Cosine Similarity，夹角余弦衡量二向量方向的差异，以二维空间为例，如下式所示。

		$cos(\theta) = \frac{x_1x_2 + y_1y_2}{\sqrt{x_1^2 + y_1^2}\sqrt{x_2^2 + y_2^2}}$ *N* 维样本的夹角余弦为， $cos(\theta) = \frac{x_1x_2}{\lvert x_1\rvert\lvert x_2\rvert} = \frac{\sum_{i=1}^{N} x_{i,1}x_{i,2}}{\sqrt{\sum_{i=1}^{N} x_{i,1}x_{i,2}}^2 \sqrt{\sum_{i=1}^{N} x_{i,1}x_{i,2}}^2}$
9	相关距离	Correlation Distance，如下式所示。 $d = 1 - corr(x_{i,1}, x_{i,2})$ (相关系数)

12.3.2. 模型结果概述

HCA 法所产生的模型结果，如下图所示。



在模型结果节点文件夹下，包含如下表中的子节点文件夹。

序号	节点文件夹名称	说明
1	Raw Data	建模时所选数据的一个副本，即将建模时所用的数据重新复制一份置于结果文件夹下，以保持节点的完整性。
2	Batch Methods Used	批处理方法，即保存建模时的一系列方法，包括参数设置等。本软件将单个数据处理方法的使用亦自动以批的形式体现，以

		保证其可比性与完整性。
3	Modeling	建模所产生的系列表格结果。
4	Plots	模型结果图形。

12.3.3. Raw Data 节点

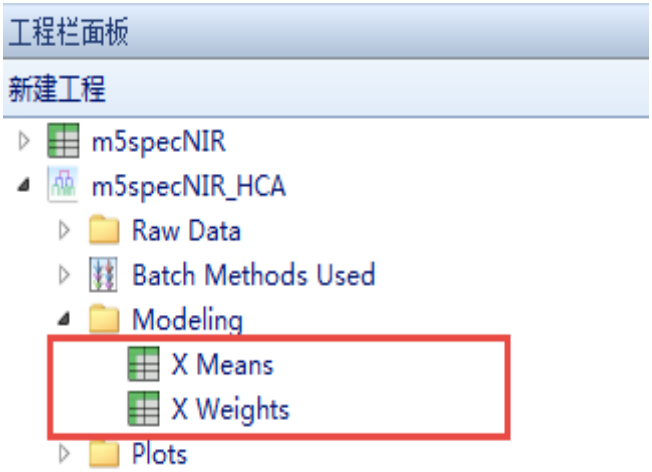
请参见 12.1.3.。

12.3.4. Batch Methods Used 节点

请参见 12.1.3.。

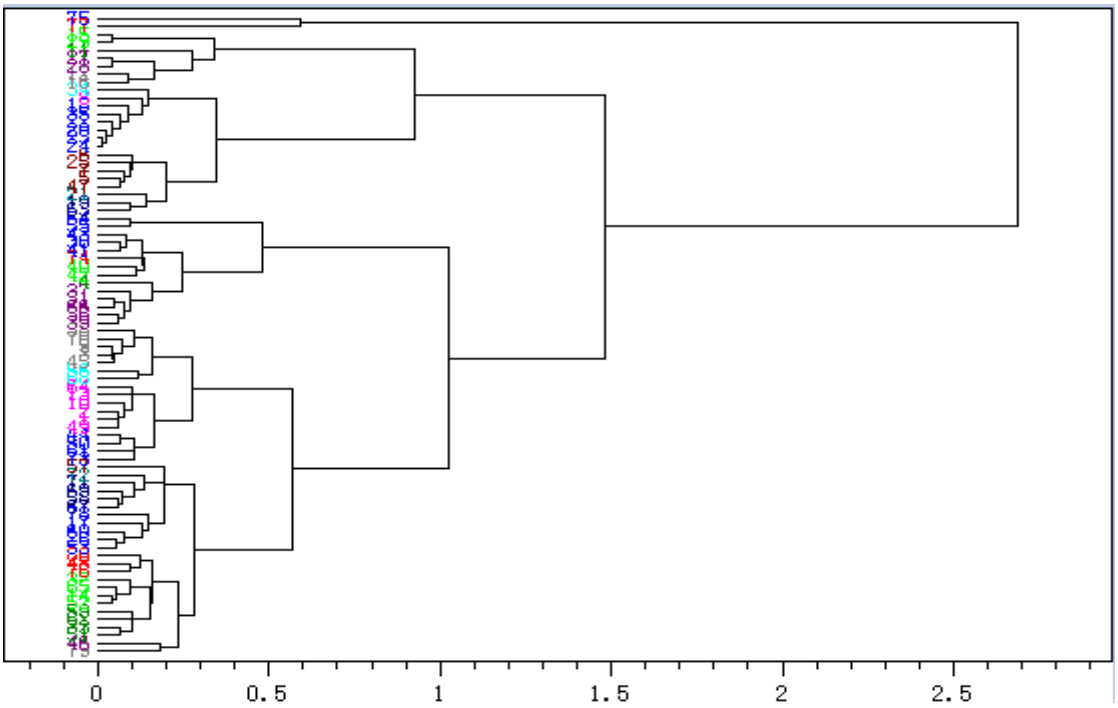
12.3.5. Modeling 节点

该节点下包含二个数据表，如下图所示，分别为变量均值及权重。。



12.3.6. Plots 节点

在 Plots 节点下即为上述系统分层聚类结果图，以图形的方式显示数据的聚类结果，如下图所示。



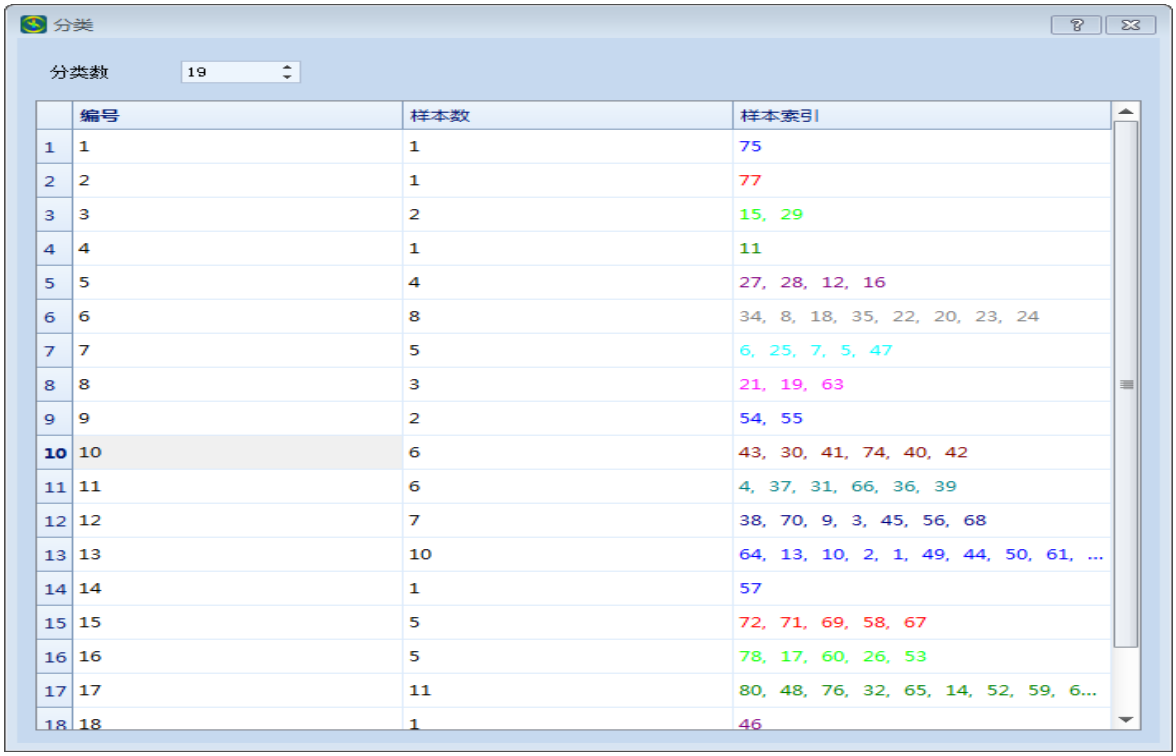
在图形中点击右键菜单，出现二个功能，一个是以 PDF 格式导出图形，另一个则以具体样本类别细节的形式得到聚类结果，如下图所示。

分类

分类数 10

	编号	样本数	样本索引
1	1	1	75
2	2	1	77
3	3	2	15, 29
4	4	5	11, 27, 28, 12, 16
5	5	8	34, 8, 18, 35, 22, 20, 23, 24
6	6	8	6, 25, 7, 5, 47, 21, 19, 63
7	7	2	54, 55
8	8	12	43, 30, 41, 74, 40, 42, 4, 37, 31, 6...
9	9	17	38, 70, 9, 3, 45, 56, 68, 64, 13, 10,...
10	10	24	57, 72, 71, 69, 58, 67, 78, 17, 60, ...

上表中不同分类样本以不同颜色标记。很显然，该表的结果与样本类别数有关，实因基于 HCA 的聚类原则，可将样本细分至单样本。若改变分类数，结果亦不断改变，如下图所示。



	编号	样本数	样本索引
1	1	1	75
2	2	1	77
3	3	2	15, 29
4	4	1	11
5	5	4	27, 28, 12, 16
6	6	8	34, 8, 18, 35, 22, 20, 23, 24
7	7	5	6, 25, 7, 5, 47
8	8	3	21, 19, 63
9	9	2	54, 55
10	10	6	43, 30, 41, 74, 40, 42
11	11	6	4, 37, 31, 66, 36, 39
12	12	7	38, 70, 9, 3, 45, 56, 68
13	13	10	64, 13, 10, 2, 1, 49, 44, 50, 61, ...
14	14	1	57
15	15	5	72, 71, 69, 58, 67
16	16	5	78, 17, 60, 26, 53
17	17	11	80, 48, 76, 32, 65, 14, 52, 59, 6...
18	18	1	46

12.4. K-means 法

如前所述，K-means 是另一类聚类方法，其中 K 是指用户需事先定义的样本类别数，而 means 则指均值，即计算新样本至各类样本的均值。概括来说，该法包括如下几个步骤：首先用户自定义样本类别 K，并随机取样本作为初始中心点，然后计算各样本至中心点的距离，划分进入类别，待所有样本均计算完成后，重新计算各类中心点(重心)，并计算各样本至新中心点的距离，重新划分样本类别，迭代直至个样本类别不再变化(收敛) 为止，输出聚类结果。

该法的结果取决于多个因素，包括 K 值的确定，初始点的选择，距离的量测方法，以及最大迭代次数等。这些因素中任一个改变都可能极大地影响聚类结果，产生不一致的结果。该法的结果可以误差平方和量度算法精确度，该值越小则表示数据点越趋近于质心，聚类效果亦越好。

针对经典 K-means 法的缺点，亦提出了不少改进方法，比如二分 K-means 法等，在此不再

赘述。

12.4.1. 操作说明

该法的操作步骤与前述方法雷同， 用户可参考 12.1.2.与 12.3.1.，差异之处在于方法参数设置，如下图所示。

上述界面中的参数解释如下表。

序号	参数	意义描述
1	类别数	数据需被划分为多少类别。
2	距离量测	距离计算方法。
3	过聚类	过聚类系数。
4	允许空类	是否允许某些类别不含样本。
5	优化初始值	是否使用程序提供的方法优化初始分类。
6	最大迭代次数	程序最大的迭代计算次数。

7	初始聚类样本	设置初始聚类样本。

上表中的距离量测方法，除 12.3.2.中所介绍的外，亦包括如下表所列举的方法。

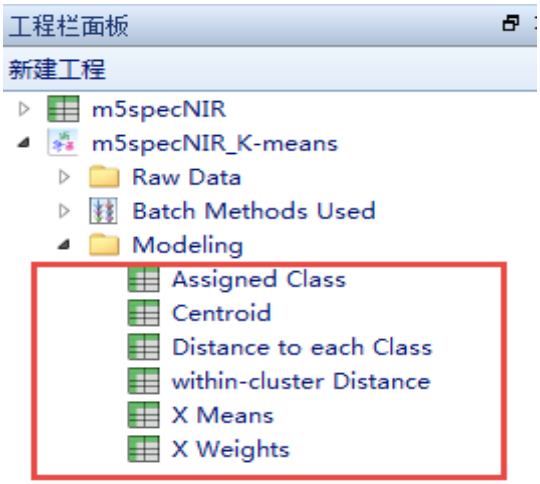
序号	方法名称	详细说明
1	绝对相关系数	Absolute Correlation Cofficient，以下式计算。 $$corr_{abs} = 1 - \lvert corr \rvert$$
2	汉明距离	Hamming Distance，二等长字符串间汉明距离即为将其中一个变为另一个所需要最小替换次数，如“1111”与“1001”间的汉明距离为 2。
3	杰卡德距离	Jaccard Distance，用于衡量集合间的差异性，是杰卡德相似系数的补集。用二个集合中不同元素占所有元素的比例表示，如下式所示。 $$d = 1 - sim_{\text{Jaccard}} = \frac{\lvert \mathbf{x}_1 \cup \mathbf{x}_2 \rvert - \lvert \mathbf{x}_1 \cap \mathbf{x}_2 \rvert}{\lvert \mathbf{x}_1 \cup \mathbf{x}_2 \rvert}$$
4	秩相关系数	Spearman Rank Correlation，亦称斯皮尔曼等级相关系数。若 $\mathbf{x}_1$ 与 $\mathbf{x}_2$ 均为含有 N 各元素的集合，对 $\mathbf{x}_1$ 与 $\mathbf{x}_2$ 分别按升序或降序排列得到 $\mathbf{x}_{new,1}$ 与 $\mathbf{x}_{new,2}$，它们中的各元素则分别为其中 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 中的排行。最后由下式计算该系数。 $$corr_{Spearman} = \frac{\sum_{i=1}^{N}(x_{new,1,i} - \overline{\mathbf{x}_{new,1}})(x_{new,2,i} - \overline{x_{new,2}})}{\sqrt{\sum_{i=1}^{N}(x_{new,1,i} - \overline{\mathbf{x}_{new,1}})^2}\sqrt{\sum_{i=1}^{N}(x_{new,2,i} - \overline{\mathbf{x}_{new,2}})^2}}$$
5	肯德尔系数	Kendall’s tau Distance，又称和谐系数，可表示多列等级变量相关程度，比如若想了解不同老师对多份学生试卷的评分标准是否一致，则可使用该法。其计算式如下， $$coff_{Kendall} = 1 - 2 \times sysdis(\mathbf{x}_1, \mathbf{x}_2)/(N \times (N-1))$$ 其中$sysdis(\mathbf{x}_1, \mathbf{x}_2)$为对称距离，若 $\mathbf{x}_1$ = {a,b,c,d}，$\mathbf{x}_2$ = {a,c,b,d}，先找出它们的所有二元约束集，再比较这些约束集，不同二元约束的个数

		即为对称距离。若 $\mathbf{x}_1 = \mathbf{x}_2$，则 $coff_{kendall} = 1$；若 $\mathbf{x}_1$ 与 $\mathbf{x}_2$ 完全无关，则 $coff_{kendall} = 0$。
6	布雷-柯蒂斯距离	Bray-Curtis Distance，其计算公式如下。 $$d = \frac{\sum_{i=1}^{N} \lvert \mathbf{x}_1 - \mathbf{x}_2 \rvert}{\sum_{i=1}^{N} \lvert \mathbf{x}_1 + \mathbf{x}_2 \rvert}$$

勾选设置初始聚类样本框，用户可直接打开数据表，选择分类样本，如下图所示。

12.4.2. 模型结果概述

模型结果节点如下图所示，其他部分不再详述，用户可参考 12.2.2.和 12.3.2.。



各节点为详细意义描述如下表。

序号	节点名称	说明
1	Assigned Class	聚类结果，以表的形式标记各样本的类别。

2	Centroid	每类样本的中心(质心)。
3	Distance to each Class	各样本到每类中心的距离。
4	within-cluster Distance	各类内距离总和。
5	X Means	变量均值。
6	X Weights	变量权重。

12.5. KNN 法

多变量的分类包括聚类分析和判别分析二个方面。聚类分析方法前面已经详细介绍到，二者的显著区别是前者无需事先定义任何样本类结构，即属于无监督的方法，对于了解数据的整体结构和先期研究非常有用；而后者则是有监督的分类方法，以系列已知信息的样本建立样本的分类规则或模型，并将其应用于新的或未知样本，判别其类别等信息，以及解释样本组别间的差异。KNN 法则是典型的有监督判别分析方法，使用广泛。

前面已详细介绍了分类方法。回归分析与分类不同，后者的响应变量 y 为样本类别信息(如是与否，好与坏，疾病与健康等二类或多类问题)，而前者则是一个或多个连续变化的量化变量(如化合物含量与浓度等)。

该法的主要思路是选择待预测未知样本在一定范围的 K 个样本，若该 K 个样本中的绝大部分属于其中的某一类别，则将未知样本判定该类别样本。具体来说，该算法包括如下几个步骤，首先设定初始化距离，计算未知样本到训练样本的距离 d，获得目前 K 个最临近样本中的最大距离 dmax，若 d < dmax，则将该训练样本作为 K 最近邻样本；然后重复上述过程直至计算完成未知样本与所有训练样本间的距离为止；最后统计 K 最近邻样本中不同类别出现的次数，输出频率最大类别作为未知样本的判别结果。简言之，该法就是以迭代的方法，不断更换未知样本的 K 最近邻样本，统计与不同类别的训练样本出现在最近邻计算

中的次数获得最终结果。

该法的主要缺点是不同类别样本个数的不平衡将导致 K 个最近邻样本中含样本数更多的样本类别占有利位置。为克服这样的问题，也有人提出采用加权方法来改进，包括对小距离样本使用更大的权重值等。

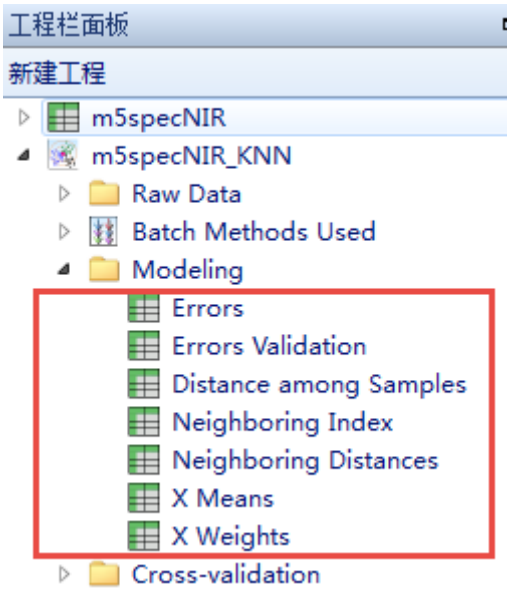
12.5.1. 操作说明

该法的操作步骤与前述方法雷同，用户可参考 12.1.2.，差异之处在于方法参数设置，如下图所示。

上述界面中的参数解释，前面的内容中已经做出介绍，其中关于距离的量测则可从 12.3.1.和 12.4.1.中获得。交互检验方法、验证集和预测集部分则可参考 12.1.2.2.和 12.1.2.3.部分。

12.5.2. 模型结果概述

模型结果的基础内容可参考 12.3.5.和 12.3.6.部分，如下图所示。

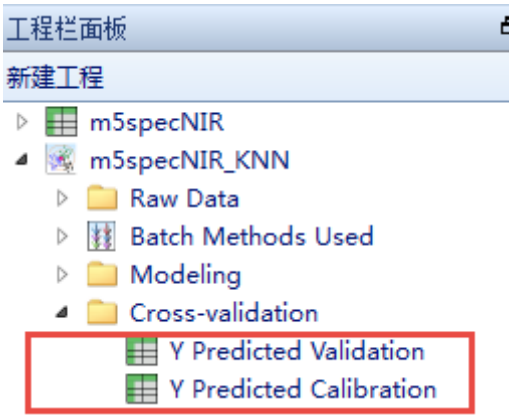


各节点的详细意义，如下表所示。

序号	节点名	说明
1	Errors	校正集的预测错误率(百分比)。
2	Errors Validation	验证集的预测错误率(百分比)。
3	Distance among Samples	各样本俩俩间的距离。
4	Neighbors Index	不同 K 值得下各样本的最近邻样本序号。
5	Neighbors Distances	不同 K 值得下各样本与最近邻样本的距离。
6	X Means	数据矩阵变量均值。
7	X Weights	变量权重。

12.5.3. Cross-validation 节点

Cross-validation 节点包含二个表格结果，如下图所示。



节点的具体涵义描述如下。

序号	节点名	说明
1	y Predicted Calibration	校正集样本聚类结果。
2	y Predicted Validation	验证集样本聚类结果。

12.6. PCA-MD 法

该法综合 PCA 分析和马氏距离计算，构建对未知样本进行分类的方法，思路简单。即对每个已知类别的样本构造 PCA 模型，然后计算未知新样本到各类的马氏距离，并以此确定样本类别。完成数据预处理后，先对各类别样本进行 PCA 分析，如下式所示:

$$\mathbf{X} = \mathbf{USV}^{T} + \mathbf{E} = \mathbf{TP}^{T} + \mathbf{E}$$

然后保留合适数量的主成分得分数，计算各样本的马氏距离，如下式所示:

$$d = \sqrt{t_i S_j^{-1} t_i^T}, \quad 其中 S_j = {T_j^T T_j}/{(N-1)}$$

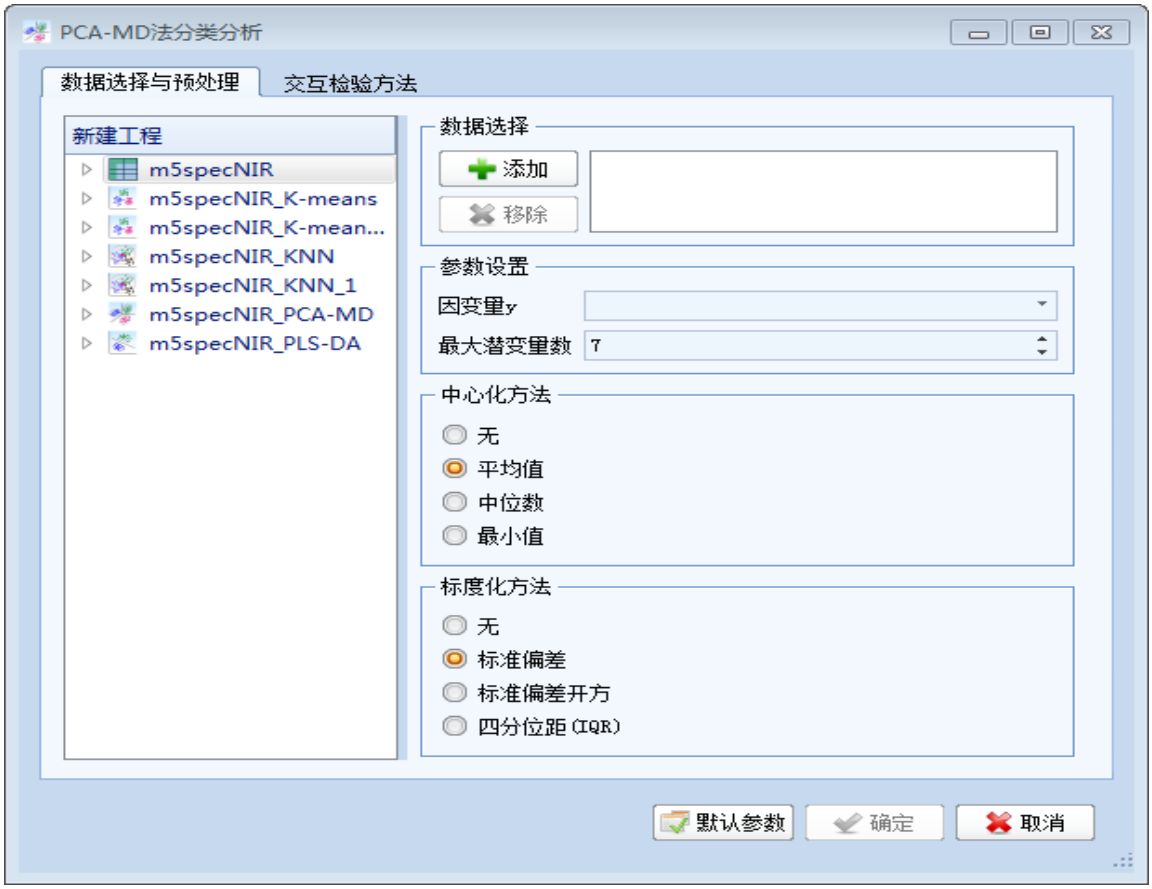
d 即为上述第 *i* 个样本到第 *j* 类的马氏距离；*N* 则为样本总数。

新样本的预测则先对未知样本做预处理，方法与建立 PCA 模型时完全相同。如前所述保留已知样本 PCA 分析时的载荷 **P**，以下式计算未知样本的得分 $\mathbf{T}_{unknown}$，并最后基于上述距离计算式得到未知样本到各类的距离，获得判别结果。

$$\mathbf{T}_{\text{unknown}} = \mathbf{X}_{\text{new}}\mathbf{P}$$

12.6.1. 操作说明

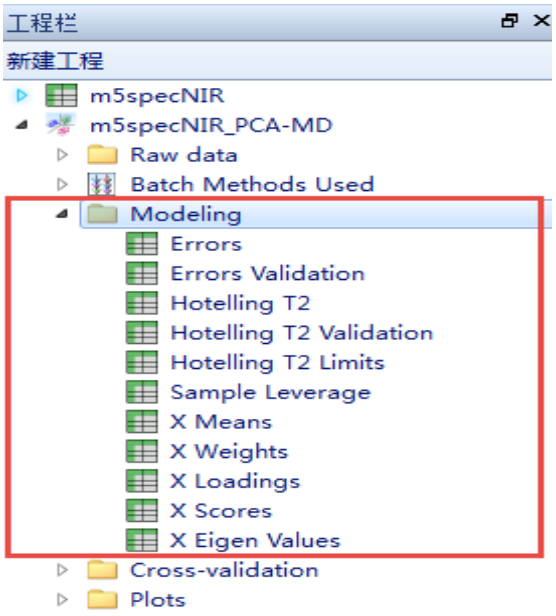
该法的操作步骤与前述方法雷同，用户可参考 12.1.2.，差异之处在于方法参数设置，如下图所示。



上述界面中的参数解释，前面的内容中已经做出介绍，其中关于距离的量测则可从 12.3.1.和 12.4.1.中获得。交互检验方法、验证集和预测集部分则可参考 12.1.2.2.和 12.1.2.3.部分。

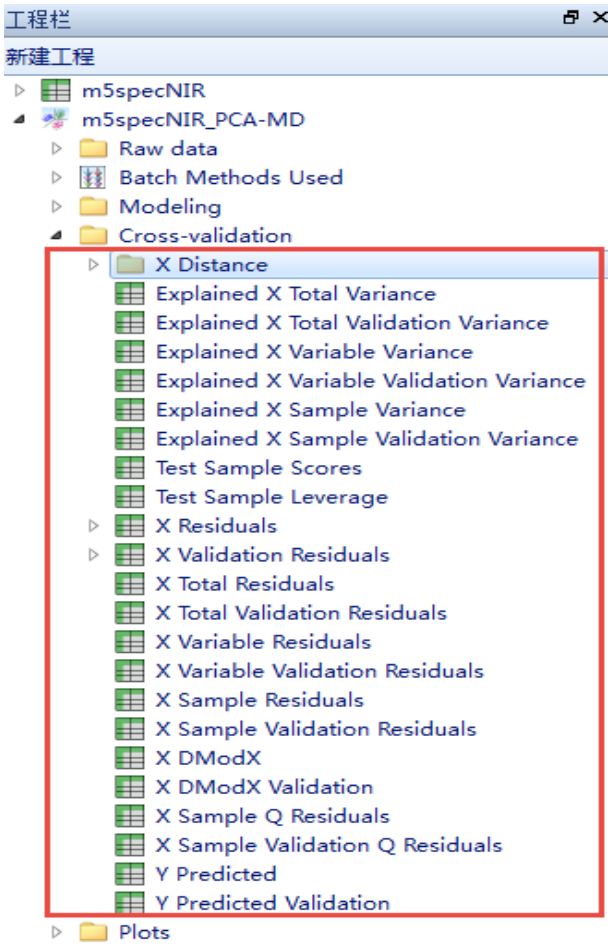
12.6.2. 模型结果概述

模型结果的基础内容可参考 12.3.5.和 12.3.6.部分，如下图所示，各节点的具体涵义可参考 12.2.5.和 12.5.2.。



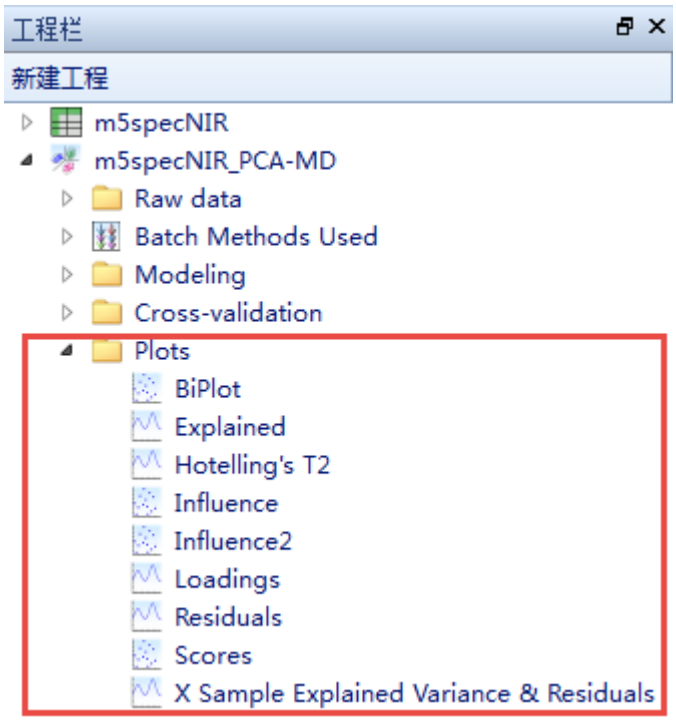
12.6.3. Cross-validation 节点

Cross-validation 节点如下图所示，其具体涵义可参考 12.2.6.。



12.6.4. Plots 节点

Plots 节点如下图所示，其具体涵义可参考 12.2.7.。



-SIMCA

暂略。

12.7. PLS-DA 法

PLS-DA 因其很好的可视化特性，已成为有监督判别分析的常用方法。PLS-DA 完全建立在 PLS 回归的基础上，用户可参考 12.13.中的部分内容，其差异是针对二类问题，即是与否或好与坏，**y** 响应变量中的值以 1 与 0 (或-1)来表示，并将预测值与 0.5 (或 0)值比较以判断类别，当然对多类问题亦可基于 PLS2 法延伸，具体请参见 12.8.。

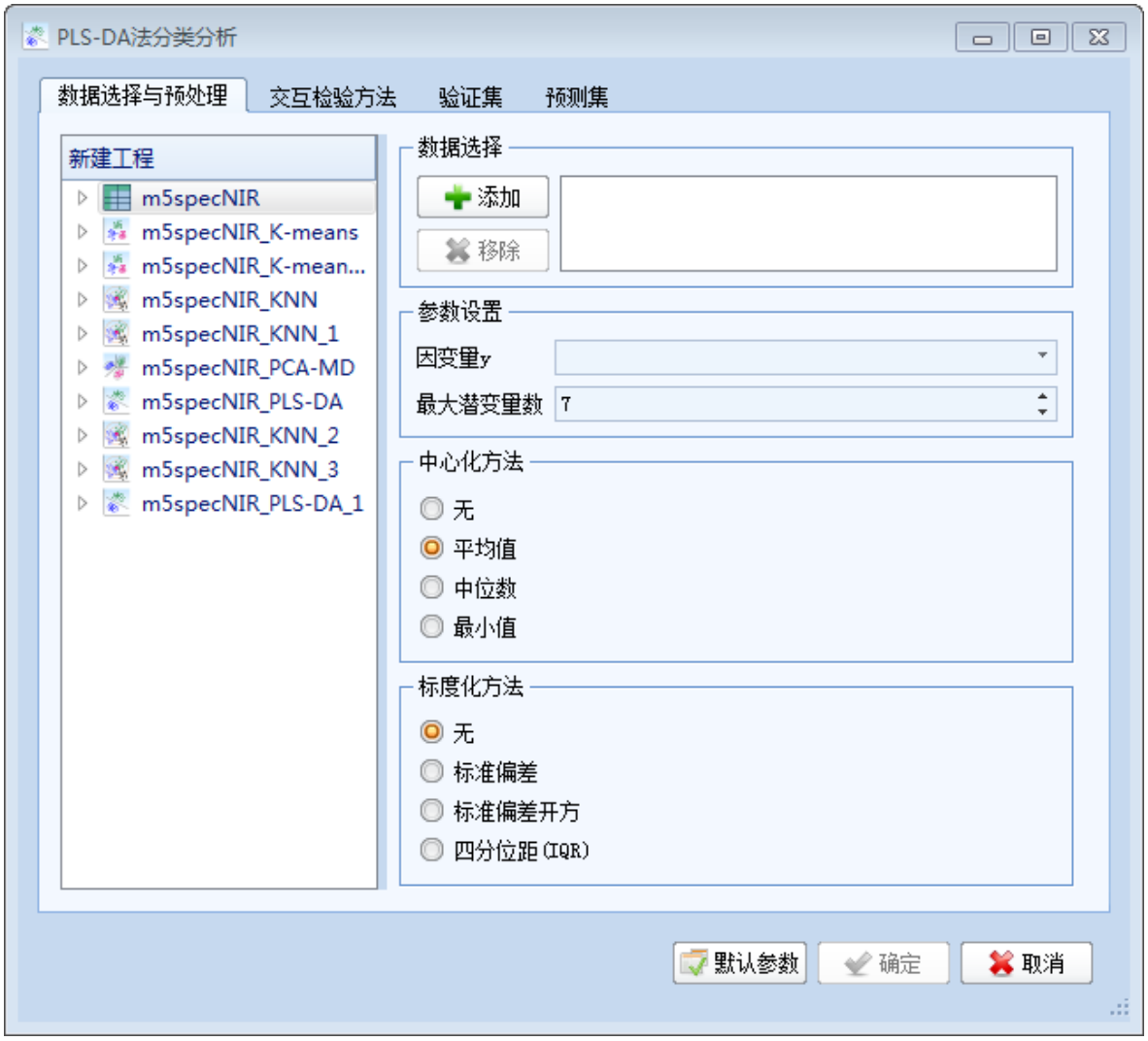
PLS-DA 可广泛用于色谱、质谱、光谱和图像等数据的分析，只要数据特征足够描述不同(类别)样本间的差异。但需要注意的是，这些数据特征通常数量很大，且往往远大于样本量，需要先进行特征选择，较少参与建模的变量，以满足构建稳健模型的样本数/特征

数条件。

PLS-DA 的结果亦可很方便地解释样本间和变量间各自的关系，以及变量对样本的解释贡献等，可参考 12.2.部分。

12.7.1. 操作说明

该法的操作步骤与前述方法雷同，用户可参考 12.1.2.，差异之处在于方法参数设置，如下图所示。

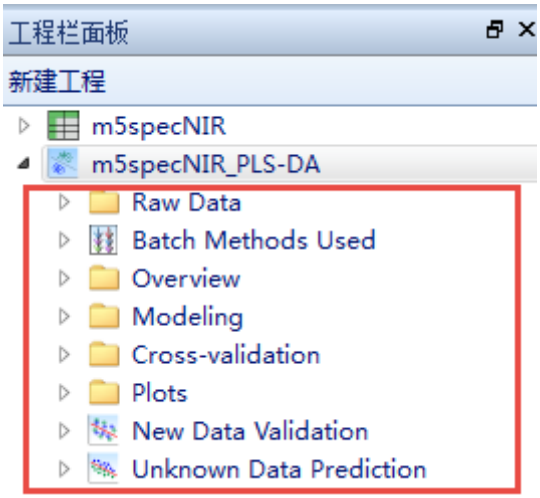


上述界面中的参数解释，前面的内容中已经做出介绍，其中关于距离的量测则可从 12.3.1.

和 12.4.1.中获得。交互检验方法、验证集和预测集部分则可参考 12.1.2.2.和 12.1.2.3.部分。

12.7.2. 模型结果概述

若同时选择验证和预测数据，则建模后可得到如下图所示的节点文件夹结果。



各节点文件夹的详细描述不再赘述，请参考 12.1.3.和 12.2.。

12.7.3. Overview 节点

该节点概括 PLS-DA 的关键数据结果，用户查看该节点可得到关于模型的主要信息，并计算不同潜变量数下的结果，如下图所示。

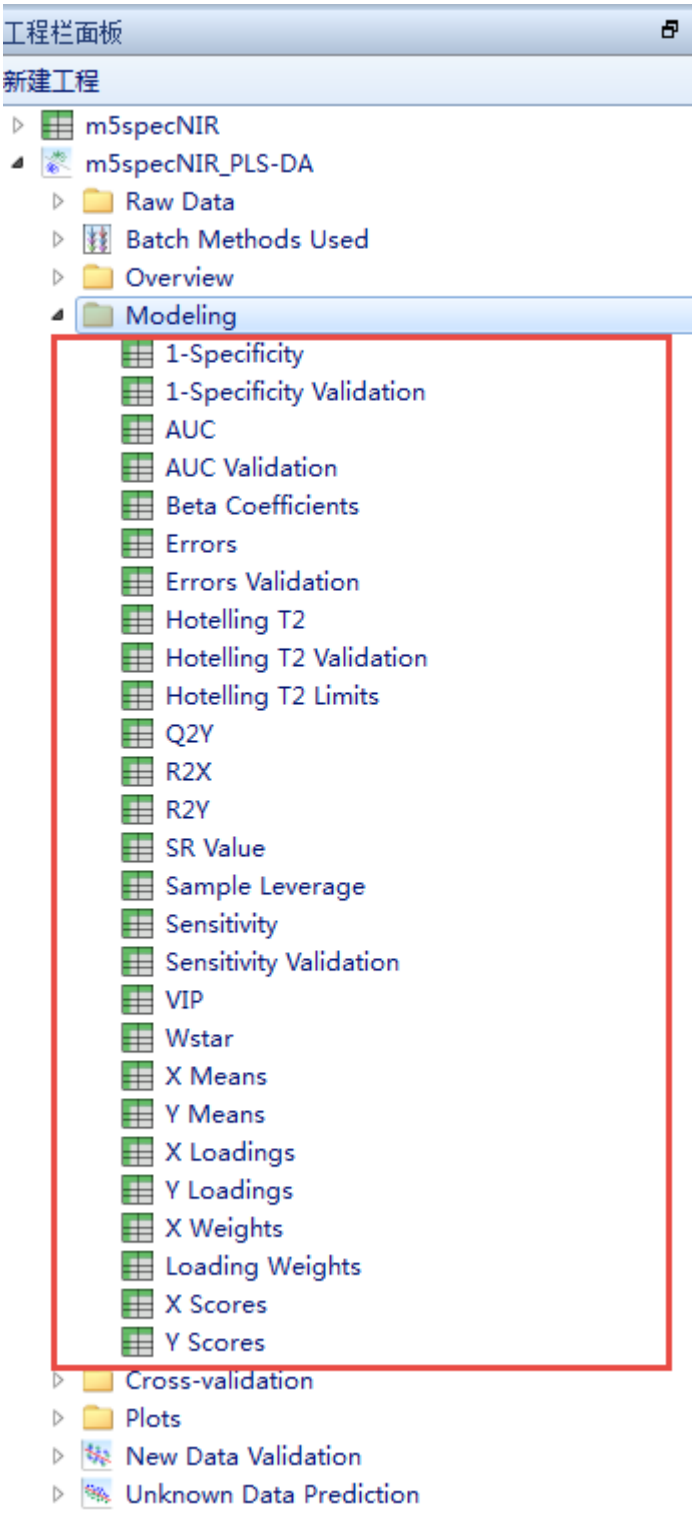
	Predictio	R2	Q2	1-Specificity	1-Specificity(Validation)	AUC	AUC(Validation)	Errors	Errors(Validation)	Sensitivity	Sensitivity(Validation)
PCs		1	2	3	4	5	6	7	8	9	10
PC 1	1	0.3904911...	0.3822943...	0.7368420...	0.717948734760284	0.8199245...	0.8145820168...	0.2375000...	0.25	0.7857142...	0.780487775802612
PC 2	2	0.5830605...	0.5706328...	0.6999999...	0.699999988079071	0.8397234...	0.8350094093...	0.2624999...	0.2625	0.7749999...	0.774999976158142
PC 3	3	0.8179340...	0.7776447...	0.7619047...	0.756097555160522	0.9160905...	0.9003771143...	0.1875	0.2	0.8684210...	0.846153855323792
PC 4	4	0.9604139...	0.9406134...	0.9024389...	0.899999976158142	0.9978001...	0.9890006252...	0.0500000...	0.0625	1	0.975000023841858
PC 5	5	0.9827973...	0.9787447...	0.9487179...	0.925000011920929	0.9993714...	0.9984286607...	0.0249999...	0.0375	1	1
PC 6	6	0.9928280...	0.9909974...	1	0.973684191703796	1	1	0	0.0125	1	1
PC 7	7	0.9958888...	0.9946629...	1	1	1	1	0	0	1	1

各结果的具体含义如下表所示。

序号	节点名	说明
1	R^2	模型相关系数(校正集样本)，以下式计算。$$R^2 = \sqrt{1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}}$$
2	Q^2	预测集样本相关系数。
3	1-Specifity	特异性，即判别的假阳性率，是真阴性与真阴性+假阳性的比值。
4	1-Specifity(Validation)	验证集特异性。
5	AUC	Area under Curve，即 ROC 曲线下面积，绘制特异性~灵敏度曲线得到。
6	AUC(Validation)	验证集曲线下面积。
7	Errors	校正集错误率(百分比)。
8	Errors(Validation)	验证集错误率(百分比)。
9	Sensitivity	灵敏度，为真阳性与真阳性+假阴性的比值。
10	Sensitivity(Validation)	验证集灵敏度。

12.7.4. Modeling 节点

该节点下的结果如下图所示，得到各模型详细结果。

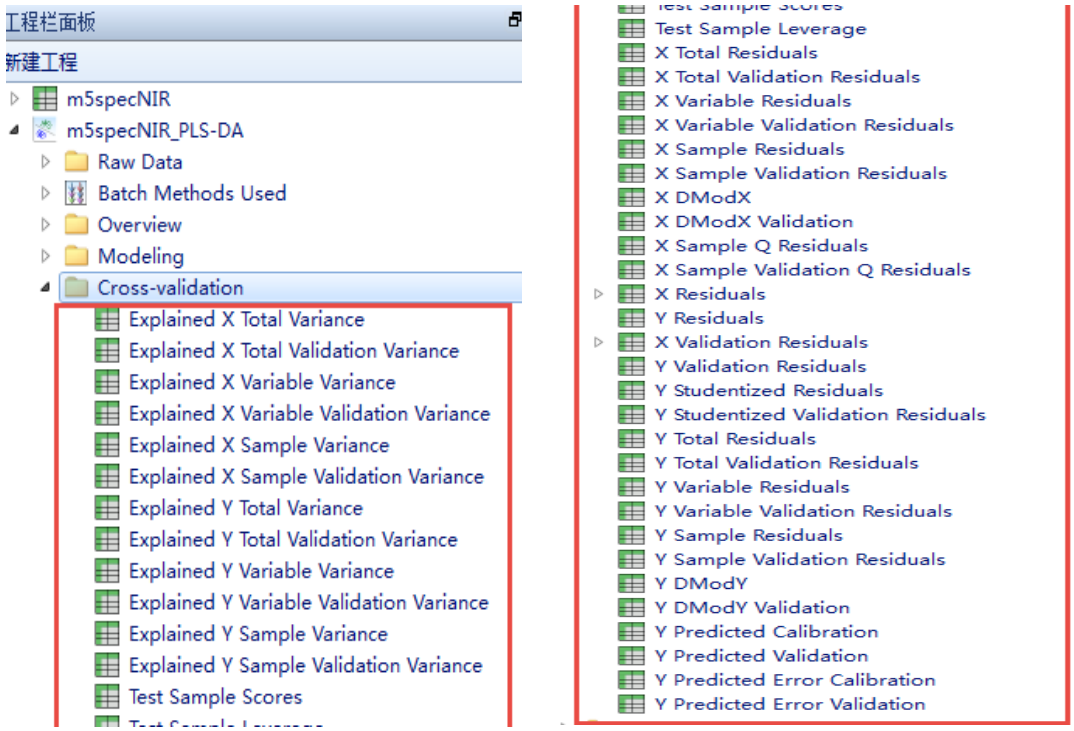


上图中绝大部分节点的信息已在 12.2.5.和 12.7.3.中描述，用户可直接参考。没有介绍到的部分，介绍在如下表中。

序号	节点名	说明
1	Hotelling T2 Limits	不同 p 值和主成分数下的 Hotelling T2 临界值，大小为 6×*N*(主成份数)。
2	SR Value	选择性比，请参见 11.7.。
3	VIP	VIP 值，请参见 11.6.。
4	Wstar	PLS 计算中，得分与原始数据转化矩阵权重，即 W^*。

12.7.5. Cross-validation 节点

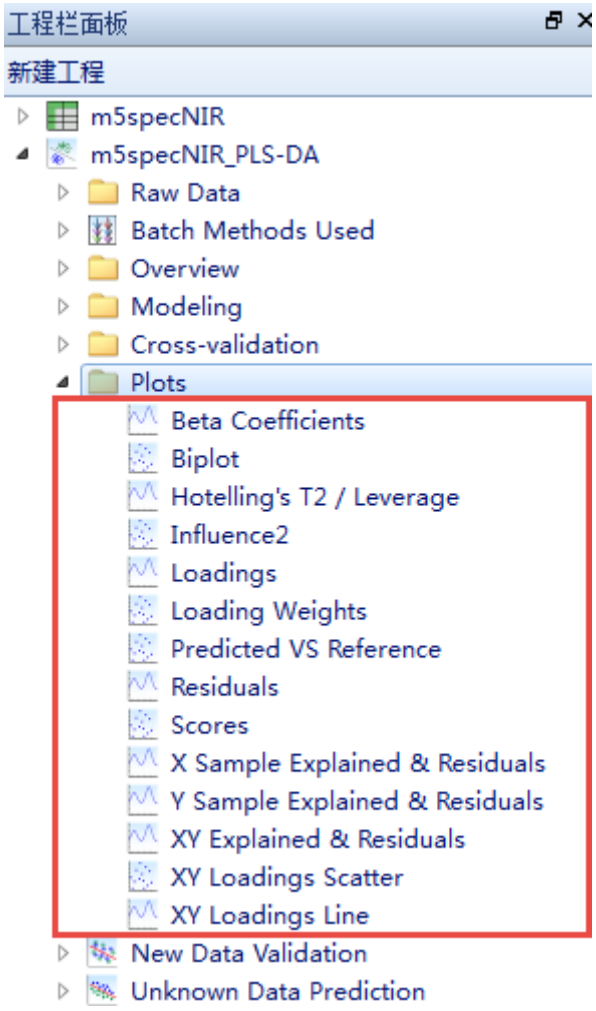
该节点集合交互检验的详细结果，具体的意义解释可参考 12.2.6.，实因 PLS 与 PCA 的差异仅在于前者同时对数据矩阵 **X** 和响应变量 **Y** 进行正交分解，分解时考虑矩阵 **T** 与 **R** 间的线性关系($\mathbf{X} = \mathbf{TP}^T$, $\mathbf{Y} = \mathbf{RQ}^T$)，即分解 **Y** 时考虑 **X** 的因素，分解 **X** 时亦考虑 **Y** 的因素，交互效验相互影响，迭代时交换迭代矢量以使对二个数据矩阵的分解过程合二为一。有人亦证明 PLS 本质上是对 $\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$ 或 $\mathbf{Y}^T\mathbf{X}\mathbf{X}^T\mathbf{Y}$ 的分解。



Cross-validation 下节点的内容，同时包含对数据矩阵 **X** 和响应变量 **Y** 评价得到的结果。

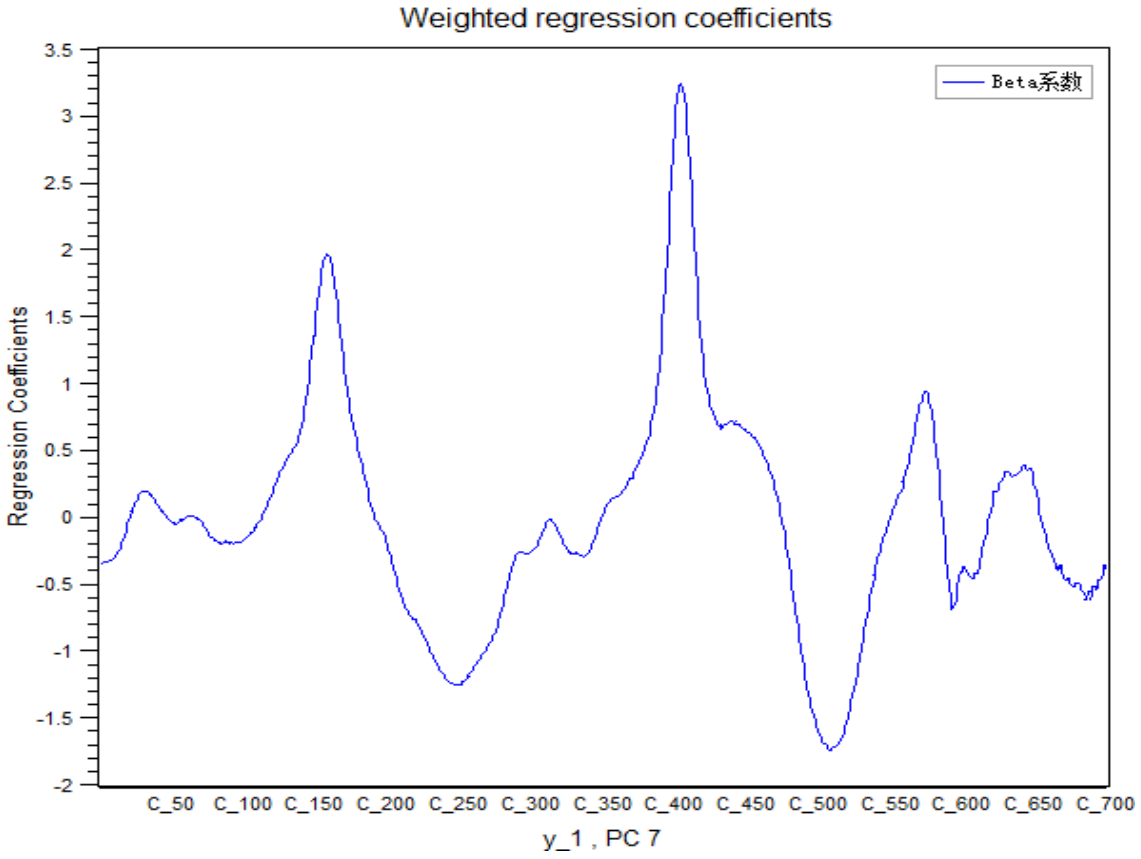
12.7.6. Plots 节点

该节点所包含的结果，如下图所示。

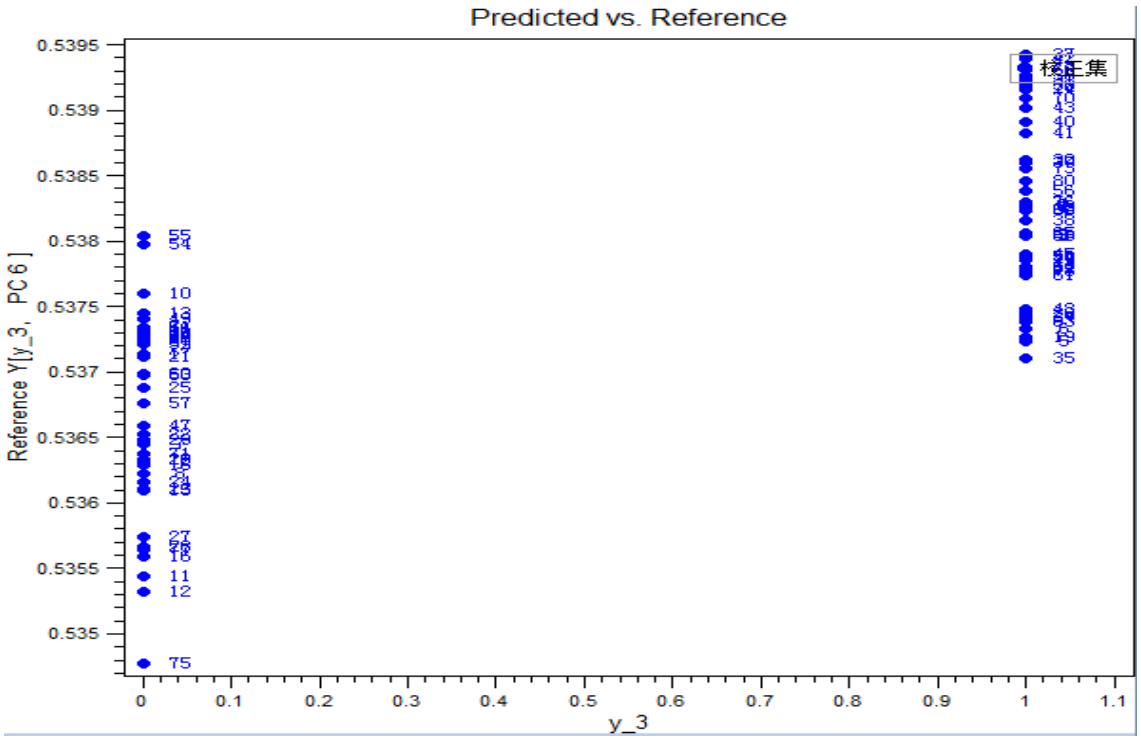


上述主要图形结果的详细操作步骤与结果解释，已在介绍 PCA 方法时说明，请参考 12.2.7.。以前没有涉及到的结果，一一介绍如下。

1) Beta Coefficients，即回归系数，可表征各变量在构建分类模型时的重要性，如下图所示。其属性修改等亦可参考 12.2.7.，在此不再赘述。



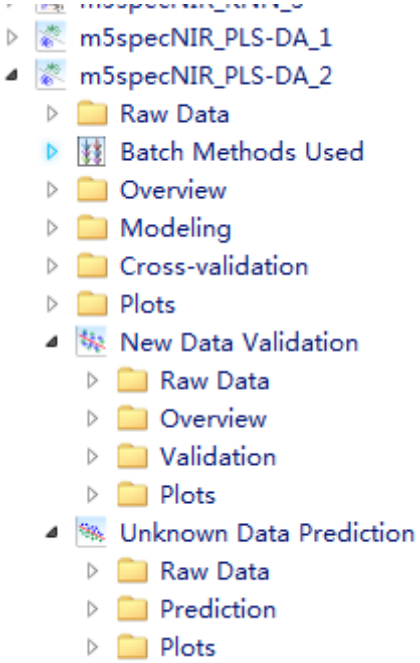
2) Predicted vs Reference：即在最优潜变量数下，预测值对实际类别值的结果，如下图所示。



3) 上图的工具栏中，亦有校正与验证集，以及不同响应变量 y 和潜变量数等结果可修改，在此不再赘述。

12.7.7. 预测与验证

若在构建模型时同时选择验证与预测集数据，则建模完成后，将得到 New Data Validation 与 Unknown Data Prediction 二个节点文件夹，如下图所示。这二个节点文件夹下的结果，分别对应验证和预测集的结果。其表格和图形结果与校正集结果雷同，不再赘述。

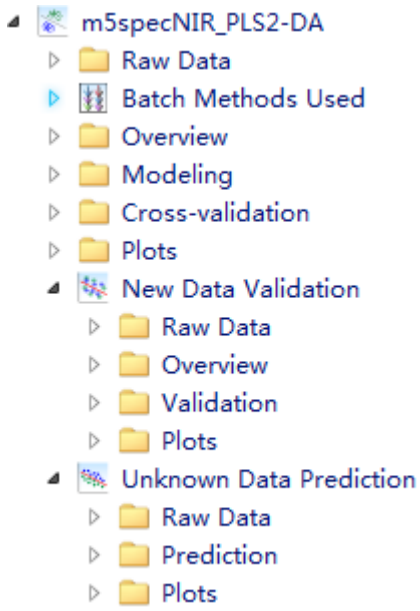


12.8. PLS2-DA 法

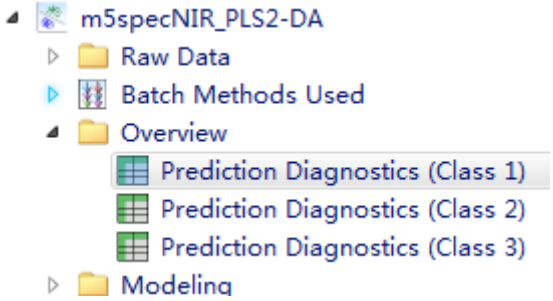
PLS2-DA 与 PLS1-DA 对应，前者用于处理含有 3 类及以上样本的数据，而后者则处理仅含 2 类样本的情形，实因 PLS2 方法可处理含有多个响应变量 **Y** 的数据，比如 Y 中第一列为化合物 **A** 的含量，第二列为化合物 **B** 的含量，而第三列则为化合物 **C** 的含量等等。此时采用 PLS2 法可同时构建不同化合物间模型。

PLS2 用于解决分类问题时，则将原始数据中的不同类别 1，2，3 等转化为仅 0 和 1(或-1 和 1)的矩阵，比如[0 0 1; 0 1 0; 1 0 0]，从而将多类样本的分类问题转换为 PLS2 可处理的回归

问题。其他分析过程与结果与 12.7.中 PLS1-DA 雷同，如下图所示，亦不再赘述。



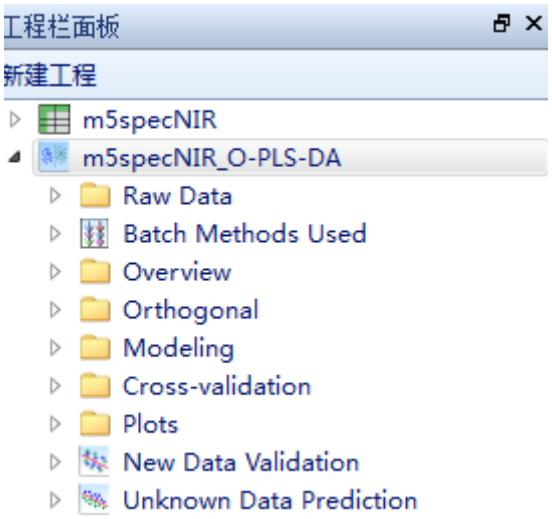
需要注意的的是，与 PLS1-DA 相比，其差异在于样本类别更多，因此如 12.7.3.中 Overview 的结果，将同时显示不同类别的结果，如下图所示，其它部分，如 Validation 节点文件夹中的结果同样如此。



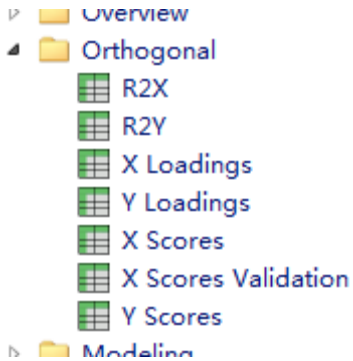
12.9. O-PLS-DA 法

O-PLS-DA 雷同于 PLS-DA，其差异在于该法基于 O-PLS 分析，而后者则基于 PLS 回归分析。关于 O-PLS 分析的更多信息，请参考 12.15.。

该法所得到的结果，如下图所示(同时选择验证集和预测集)。



与 12.7.相比，该法仅多了一个节点，即 Orthogonal，如下图所示。该部分的详细内容请参考 12.15.。



其他各节点文件夹及其节点的详细信息，包括 Raw Data, Overview, Modeling, Cross-validation,Plot,以及 New Data Validation 和 Unknown Data Prediction，其主要内容均在 12.2.和 12.7.介绍 PCA 和 PLS-DA 方法时已经详述，不再赘述，用户可参考相关内容。

这些节点文件夹中增加的少量结果，如 Modeling 下的 CovTX 和 CorrTX，以及 Cross-validation 下的 Y Studentized (Validation) Residuals 等，亦请参考 12.5.。

12.10. SVC 法

支持向量机是一种基于统计学习理论的机器学习方法，建立 VC 维理论与结构风险最小原

理的基础之上，以有限样本在模型复杂性间寻求最佳折衷，获得最好的泛化能力。支持向量是指支撑平面上把不同类别区分开的超平面上的向量点。该法最早用于解决分类问题，现已拓展用于回归分析中。当线性函数不足以解决数据结构复杂的问题时，该法引入 kernel 函数将原始数据空间映射到高维特征空间中，以处理非线性分类的问题。

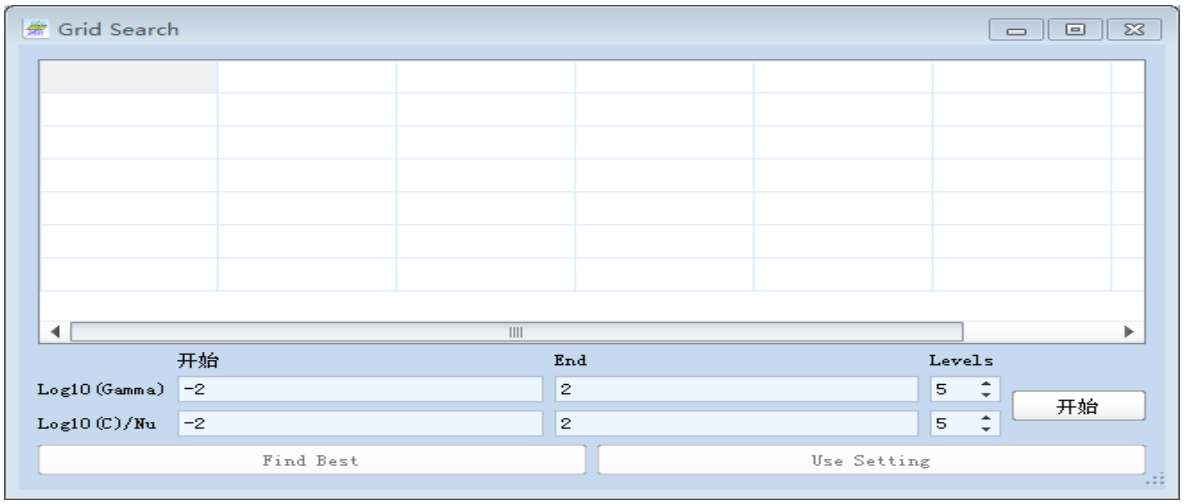
该法的数学原理请参见 18.8.，不再赘述。

12.10.1. 操作说明

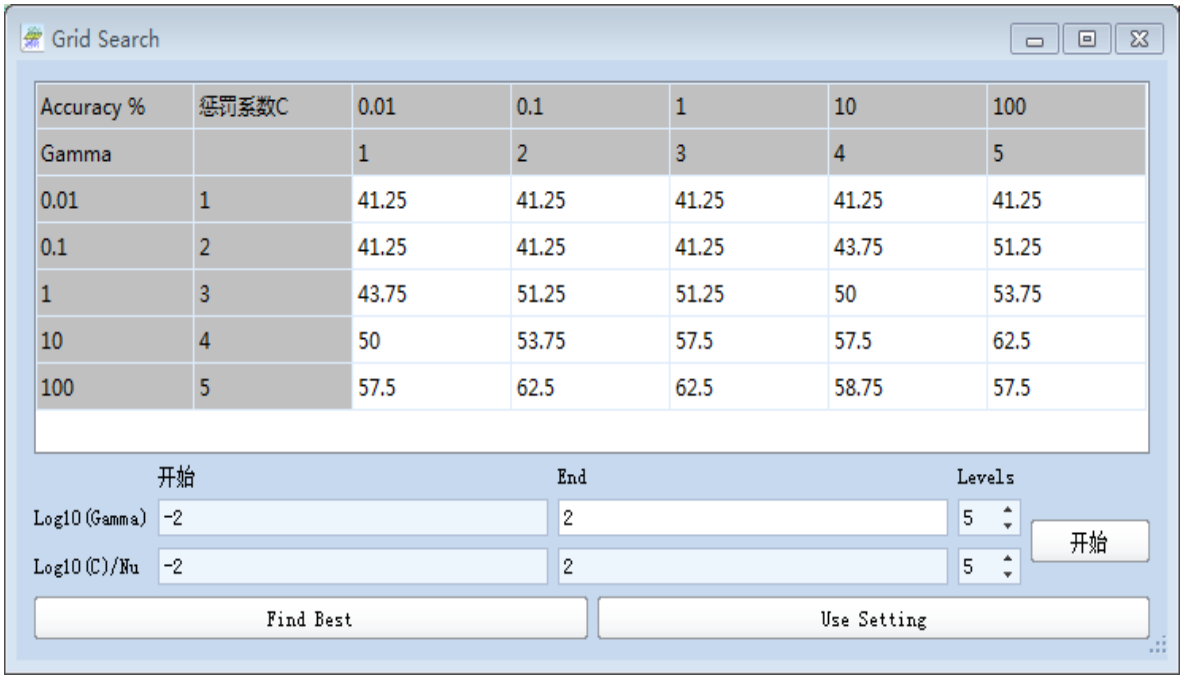
该法的操作步骤与前述方法雷同，用户可参考 12.1.2.，差异之处在于方法参数设置，如下图所示。

上述界面中的参数说明，详情请参见 18.8.。

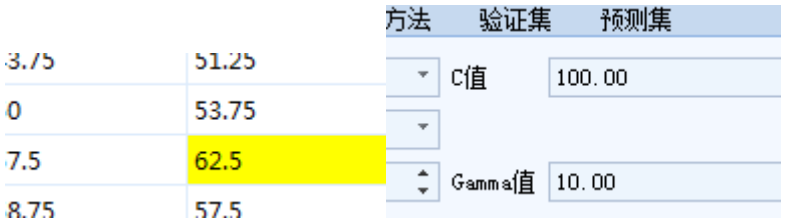
除一般参数的设置外，界面提供基于网格搜索的参数优化功能，如下图所示。



设置参数优化的边界后，点击开始按钮，即可实现参数 Gamma 和 C(或 nu)的优化，如下图所示。



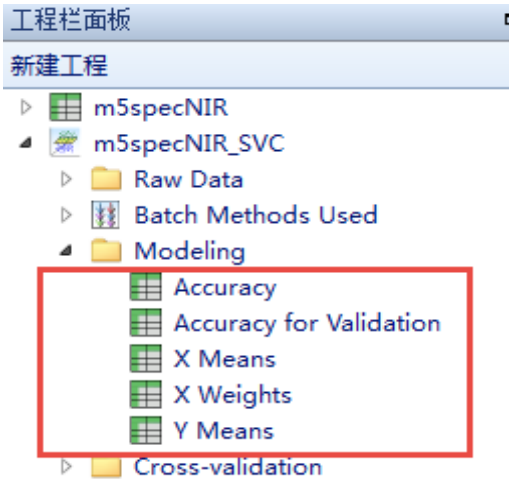
上述界面的最下端有二个功能按钮，分别找到最优计算结果，以及将该结果所对应的参数输入到程序中，以实现程序的运行，分别如下图所示。



交互检验方法、验证集和预测集部分则可参考 12.1.2.2.和 12.1.2.3.部分。

12.10.2. 模型结果概述

模型结果的基础内容可参考 12.3.5.和 12.3.6.部分，如下图所示。

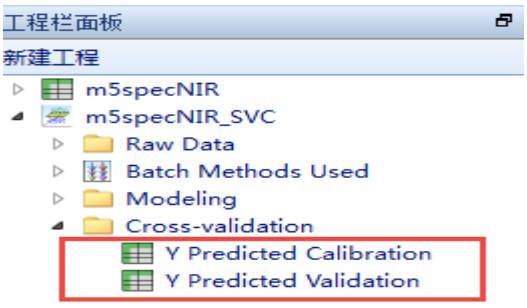


各节点的详细意义，如下表所示。

序号	节点名	说明
1	Accuracy	校正集的预测正确率(百分比)。
2	Accuracy for Validation	验证集的预测正确率(百分比)。
3	X Means	数据矩阵变量均值。
4	X Weights	数据矩阵变量权重。
5	y Means	响应变量变量均值。

12.5.3. Cross-validation 节点

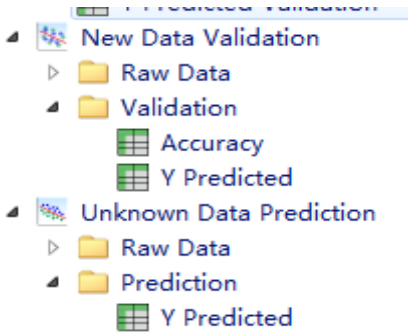
Cross-validation 节点包含二个表格结果，如下图所示。



节点的具体涵义描述如下。

序号	节点名	说明
1	y Predicted Calibration	校正集样本分类结果。
2	y Predicted Validation	验证集样本分类结果。

若 12.10.1.中同时选择验证集和预测集，则建模后同时产生如下图所示的结果。对验证集，同时获得预测正确率(百分比)，以及各样本实际预测类别值；而对预测集，则得到样本类别预测结果。



12.11. PCR 法

下面开始依次介绍本软件所涵盖的回归方法。但用户需要牢记的是，好的模型来自好的数据，若数据质量低劣，显然模型的有效性和可用性也同样是存疑的。

12.11.1. 回归分析基础

回归即指构建自变量(数据矩阵 **X**)与因变量 **y** 间定量关系的统计分析过程，所得到到的模型 **y** = *f*(**X**)可解释数据 X 与一之间的相互关系，并预测新的样本(数据)。单变量与多变量线性回归的基本模型如下二式所示。通常地，后者显然可得到更准确的结果。

$$y = b_0 + b_1 X$$

$$y = b_0 + \sum_{i=1}^{n} b_k x_k$$

构建回归模型前，需要收集数据矩阵 **X** 和因变量 **y**，并基于某规则优化获得上述模型中参数，如 **y** 实际量测值与预测值间偏差平方和最小为目标函数；模型一旦确定，即可用于预测新的数据集 $\mathbf{X}_{unknown}$。显然当 $\mathbf{X}_{unknown}$ 容易得到(如复杂体系光谱等)，而响应变量 **y** 难于得到时(如活性化合物或生物标志含量)，基于模型计算与预测的方法比传统直接量测的方法省时省力得多。

ⓘ 构建稳健模型，一方面需提取训练集中的相关信息，计算上述回归参数，另一方面则需使用独立测试集验证模型的有效性，比较训练集与测试集所得到的结果评判模型。一个好的回归模型需尽可能关注信息变量，即极大提高他们的权重，而尽可能低地降低不相关无效变量的权重(对模型解释没有作用或作用极低的特征)，并防止过拟合现象，即可有效区分信息与随机噪声对模型的贡献或影响。

ⓘ PCR 是一种投影方法，与 PCA 法雷同。该法包括二个步骤，即首先对数据矩阵 X 进行 PCA 分解，然后使用一定数量的主成分与 y 进行 MLR 建模。因而该法与 MLR 的本质区别就在于使用 PCA 分解所得到的主成分，而非原始数据 X 构建模型。

关于 PCA 的介绍，可参考 12.2.；关于 MLR 的介绍，可参考 12.12. PCR 法详情，可参考 18.4.。

回归模型主要包括如下表所示的结果。

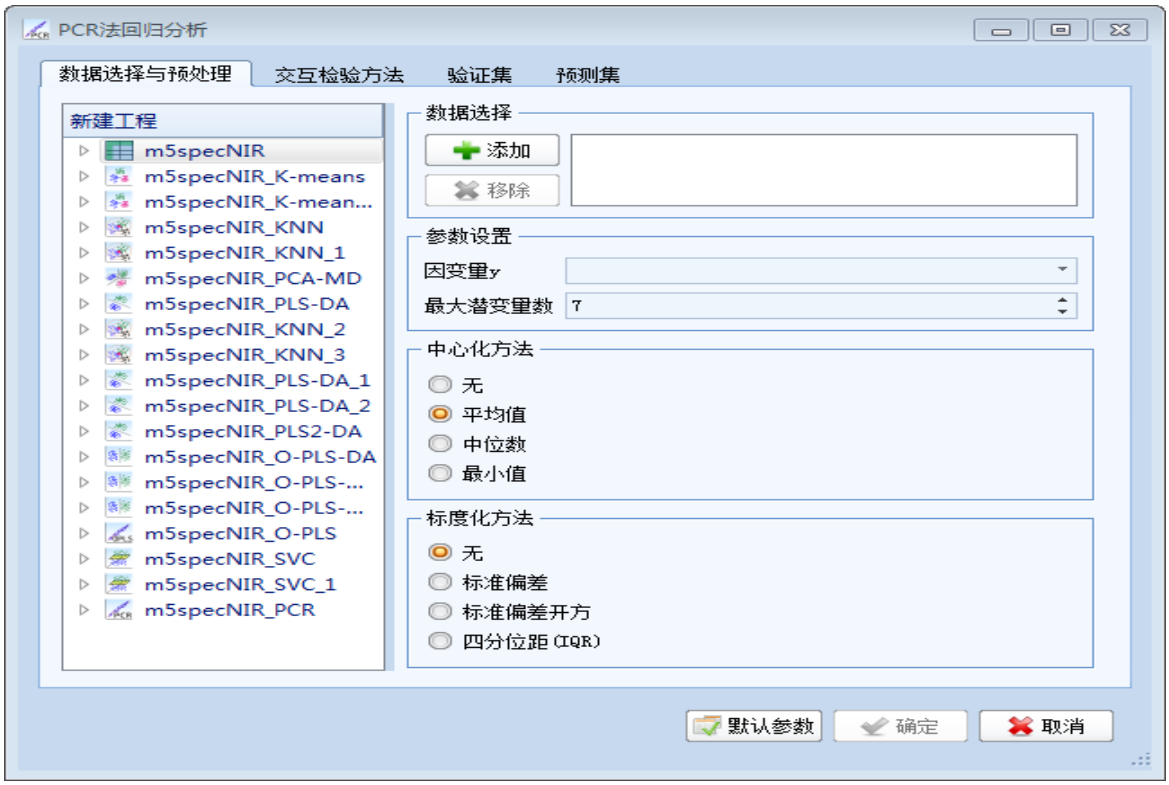
序号	结果名称	说明
1	回归系数	B-coefficients，即回归模型中各系数 $b_0,b_1,\ldots,b_n$。
2	y 预测值	基于实际量测 **X** 矩阵与已知模型所得到的响应变量 **y** 值。
3	残差	响应变量 **y** 实际值与预测值间的差值。
4	得分	请参考 12.2.5.。

5	载荷	请参考 12.2.5.。

本软件包括 MLR、PCR、PLS1、PLS2、OPLS 和 SVR 等多种回归分析方法。MLR 通常用于数据矩阵 **X** 所含变量数少于 20 的情形，且这些变量间的相关性较小。若响应变量 **y** 中包括多个属性值，**y** 实为矩阵 **Y**，且需使用所有变量构建模型，则 PLS 法通常是最好的选择。通常对强非线性数据，PCR 和 PLS 是首选的回归方法，且对含多个响应变量值的数据，对它们构建独立的模型往往比同时构建多个 **y** 值的模型结果更优。特别地，若多个 y 变量含有一定程度噪声，但不同 y 间又有某些相关性，则 PLS 法是最好选择。此外，对新数据集，PCR 与 PLS 法往往得到相近的预测结果，但 PLS 法可使用更少的主成分数。SVR 法则可用于构建非线性回归模型。

12.11.2. 操作说明

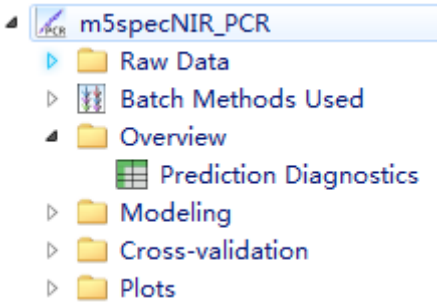
该法的操作步骤与前述方法雷同，用户可参考 12.1.2.，其初始界面如下图所示。



交互检验方法、验证集和预测集部分则可参考 12.1.2.2.和 12.1.2.3.部分。

12.11.3. 模型结果概述

模型结果的内容如下图所示。



上图中的主要内容已经在前面介绍到，用户可参考 12.1.3.，回归分析所新增加的内容，包括各节点文件夹和具体节点结果，以及验证集和预测集结果，均可参考 12.13.2，在此不再赘述。

12.12. MLR 法

MLR 法是基于经典最小二乘的回归分析方法，以获得单响应变量 **y** 与数据矩阵 **X** 间的线性关系，是单变量线性回归分析的逻辑延伸。回归模型系数可由下式得到。

$$b = (X^T X)^{-1} X^T y$$

从上式可以看出，对共线性数据，矩阵逆运算可能导致模型计算及结果的不稳定。因此数据变量间的非线性独立性是获得可靠 MLR 结果的基础。另一方面，MLR 也同样要求数据样本数大于变量数，否则无法得到数据唯一解。

MLR 回归结果的方差分析(ANOVA)，可由如下表概括。

方差来源		平方和	自由度	均方	F 值
1	回归	$SS_{Regression}$	*k*	$MS_{Regression}$	$MS_{Regression}/MS_{Error}$
2	残差	SS_{Error}	*N*-*k*-1	MS_{Error}	
3	总计	SS_{Total}	*N*-1		

上表中 SSRegression，SSError，SSTotal 分别指总回归离差平方和，误差离差平方和，以及总离差平方和，其中 SSRegression 表征模型能所描述的信息量，由模型形式，数据质量，以及模型所用数据矩阵等决定。

此外需要注意的是，对噪声数据该法可能得到过拟合的结果。若在构建模型时同时选择验证与预测集数据，则建模完成后，将得到 New Data Validation 与 Unknown Data Prediction 二个节点文件夹，该部分内容不再赘述，用户可参考 12.2.8.。

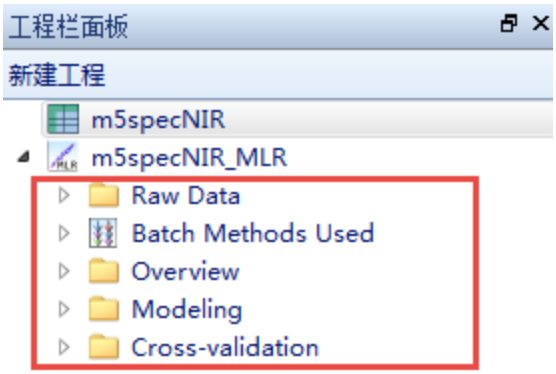
更多内容则不再赘述，实因本法的基础性，一般用户比较了解，绝大多数多元统计通过书籍，亦进行了大量叙述。

12.12.1. 操作说明

该法的操作步骤与前述方法雷同，用户可参考 12.1.2.，其初始界面亦与 12.11.1.中图形雷同，不再赘述。交互检验方法、验证集和预测集部分同样可参考 12.1.2.2.和 12.1.2.3.部分。

12.12.2. 模型结果概述

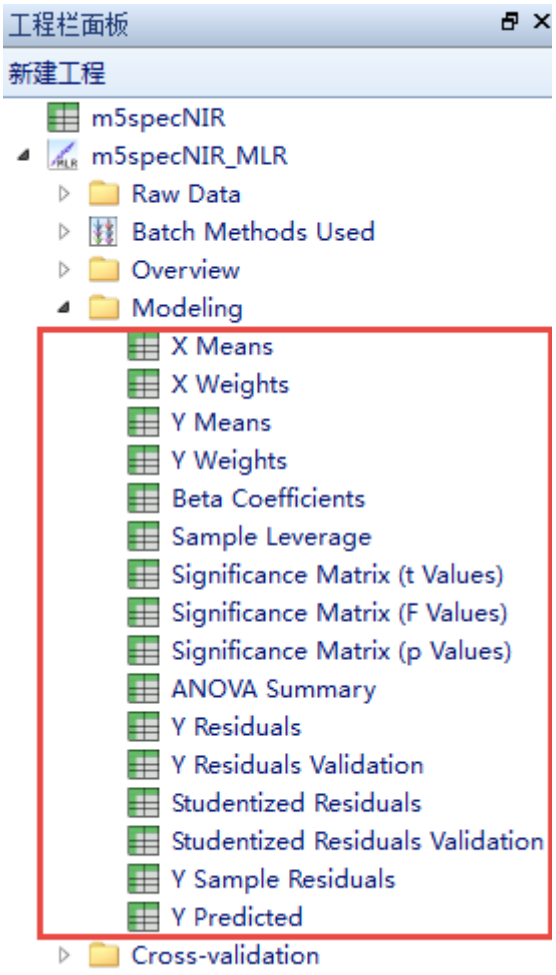
模型结果的内容如下图所示。



上图中的主要内容已经在前面介绍到，用户可参考 12.1.3，在此不再赘述。

12.12.3. Modeling 节点

该节点下的结果如下图所示，得到各模型的详细结果。



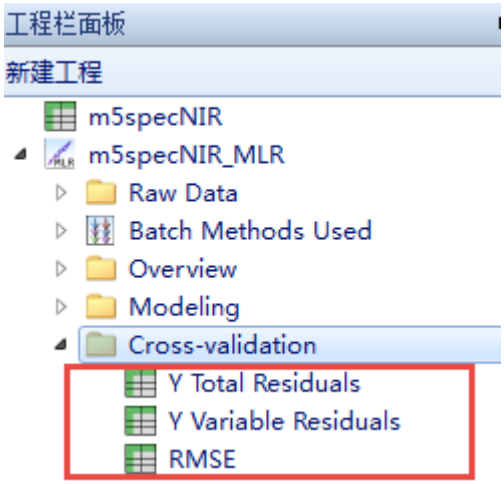
上图中各节点的详细介绍如下表。

序号	节点名	说明
1	X Means	数据矩阵 **X** 的变量均值，其大小为 *1*×*n*。
2	X Weights	数据矩阵 **X** 的变量权重，其大小为 *1*×*n*。
3	y Means	响应变量 **y** 的均值，其大小为 *1*×*1*。
4	y Weights	响应变量 **y** 的权重，其大小为 *1*×*1*。
5	Beta Coefficients	回归系数 B 值，其大小为 *1*×*n*。若在建模参数中勾选考虑截距选项，其大小变为 *1*×(*n* + 1)。

6	Sample Leverage	样本杠杆值，其大小为 *1×n*。
7	Significance Matrix (t Values)	回归系数假设检验的值(t 检验)，可判断模型显著性因素，其大小为 *1×n*。t 值越小则意味着该变量对结果影响越小(不明显)，应予以剔除。
8	Significance Matrix (F Values)	F 检验值。
9	Significance Matrix (p Values)	拒绝原假设的最小显著性水平，其大小为 *1×n*。获得 P 值后，将其与给定显著性水平比较以决定是否接受原检验。通常 t 值越大则 p 值越小。
10	ANOVA Summary	上述 12.12.2.中所述 ANOVA 分析的整体结果，即回归方程的显著性检验结果，表示模型中被解释变量与所有解释变量间线性关系总体上是否显著。
11	y Residuals	**y** 残差值，其大小为 *m×1*。
12	y Residuals Validation	验证集的 **y** 残差值，其大小为 *m×1*。
13	Studentized residuals	学生化残差，可诊断样本奇异值，其绝对值可发现强影响点，大小为 *m×1*。其计算如下式所示。 $$Residual_{\text{student}} = \frac{e_i}{\hat{\delta}\sqrt{1 - h_{ii}}}$$
14	Studentized residuals Validation	学生化残差(验证集)。
15	y Sample Residuals	**y** 样本残差，其大小为 *m×2*，同时包含校正集和验证集结果。
16	y Predicted	**y** 预测残差，其大小为 *m×2*，同时包含校正集和验证集结果。

12.12.4. Cross-validation 节点

该节点集合交互检验的详细结果，具体的意义解释可参考 12.2.6.。具体所包含的结果，如下表所示。



各节点的详细介绍，如下表所示。

序号	节点名	说明
1	y Total Residuals	**y** 总残差。
2	y Variable Residuals	**y** 变量残差。
3	RMSE	均方根误差，由下式计算: $RMSE = \sqrt{\frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{m}}$

12.13. PLS 法

PLS 回归亦称潜空间投影法或简单 PLS 法，其作用在于最大化数据矩阵 **X** 与响应变量 **Y** 的协方差，构建它们间的定量模型关系，并优化寻找 **X** 中的潜变量，以更好地预测 y 潜变量，与 PCA 法所得主成分类似。

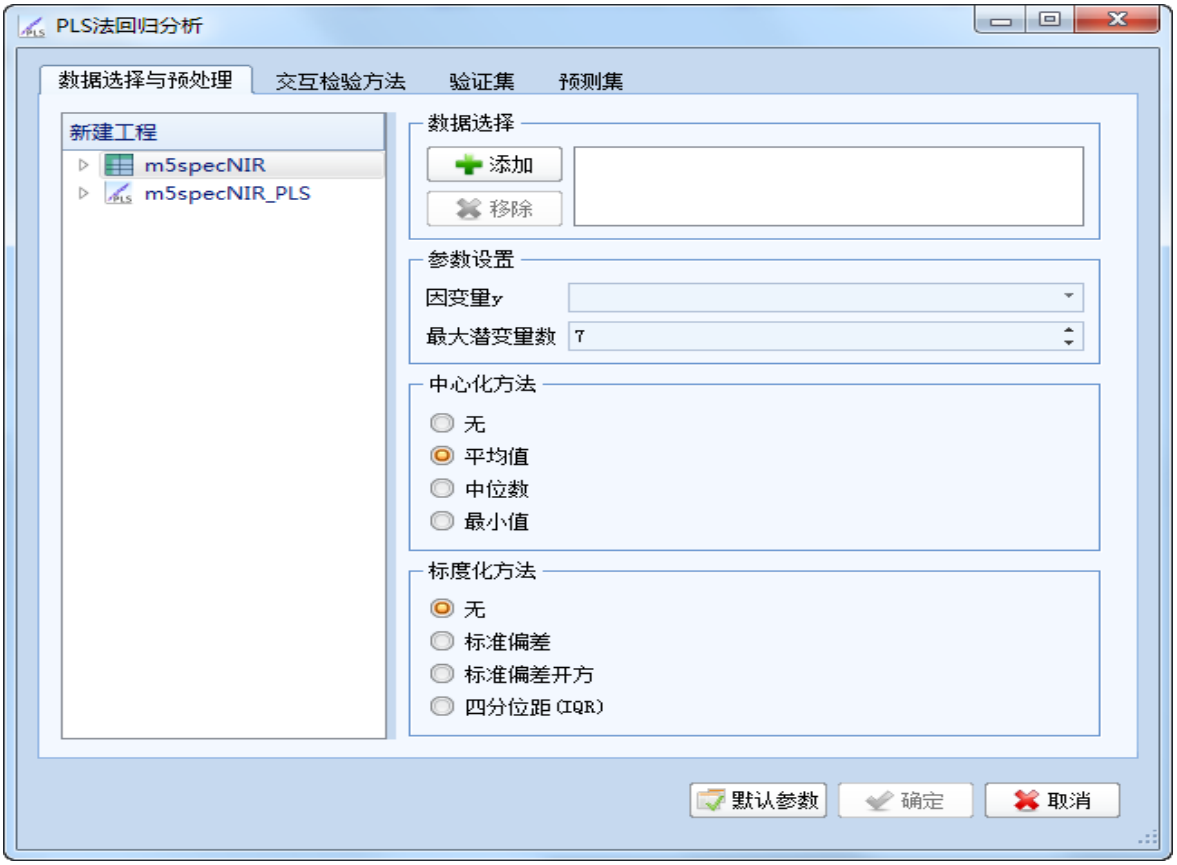
PLS 分析的总体思路可概括为：首先提取 X 中与 Y 相关性最大的得分 t，并基于 t 产生 Y 载荷 q，进而计算得到 Y 得分 u，获得 X 与 Y 得分 u 与 t 间的模型，即 u = f(t)以最大化原始数据间的定量关系。

与 PCA 比较，其差异在于该法同时综合数据 X 与 Y 的信息拟合模型，如前所述在数据间迭代计算寻找相关因子，该法专注于响应变量 Y 的预测，以更少的因子数便可达到所需的优化结果，即模型收敛的最小化残差。

关于 PLS 的更详细信息，可参考 18.5.；与 PCA 类似部分，亦可参考 12.2.。

12.13.1. 操作说明

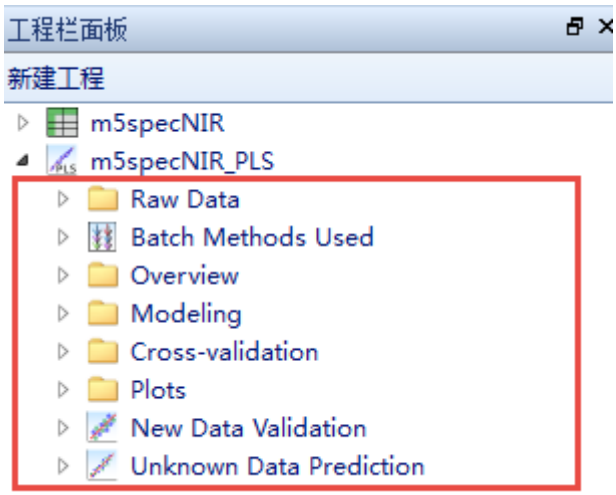
该法的操作步骤与前述方法雷同，用户可参考 12.1.2.，如下图所示。



交互检验方法、验证集和预测集部分则可参考 12.1.2.2.和 12.1.2.3.部分。

12.13.2. 模型结果概述

模型结果的基础内容可参考 12.3.5.和 12.3.6.部分，如下图所示。



各节点的详细意义，如下表所示。

序号	节点名称	说明
1	Raw Data	目录下为建模时所选数据的副本。
2	Batch Methods Used	目录下的节点可以用来建批，它相当于保存了 PLS 建模时的一系列命令，包括参数设置。
3	Overview	模型主要结果概述。
4	Modeling	模型结果。
5	Cross-validation	建模时的交互检验结果。
6	Plot	目录下面是常用图形节点，对 Modeling 和 Cross-validation 目录下数据的绘图。
7	New Data Validation	新数据的验证，可在建模时选择数据验证，亦可在建模

		后独立验证。
8	Unknown Data Prediction	未知数据的预测，可在建模时选择数据预测，亦可在建模后独立预测。

下面介绍如上节点文件夹中的主要结果。

12.13.3. Overview 节点

该节点概括 PLS 分析的关键数据结果，用户查看该节点可得到关于模型的主要信息，并计算不同潜变量数下的结果，如下图所示。

	Prediction	R2Y	Q2Y	PRESS	RMSEP	Bias	SEP	Slope	Offset	Correlation
PCs		1	2	3	4	5	6	7	8	9
PC 1	1	0.3904911...	0.3822943...	7.2835634...	0.3017358...	-0.005107...	0.3035960...	0.3826083...	6.3129923...	0.6032304...
PC 2	2	0.5830605...	0.5706328...	5.0628045...	0.2515652...	-0.001608...	0.2531472...	0.5733019...	4.3650218...	0.7466639...
PC 3	3	0.8179340...	0.7776447...	2.6218606...	0.1810338...	0.0001428...	0.1821759...	0.7822238...	2.2287636...	0.8779387...
PC 4	4	0.9604139...	0.9406134...	0.7002456...	0.0935578...	-0.002895...	0.0941030...	0.9190349...	0.8256634...	0.9691369...
PC 5	5	0.9827973...	0.9787447...	0.2506273...	0.0559717...	-0.001256...	0.0563107...	0.9671154...	0.3352688...	0.9890443...
PC 6	6	0.9928280...	0.9909974...	0.1061514...	0.0364265...	3.6865062...	0.0366563...	0.9913211...	0.0888522...	0.9953456...
PC 7	7	0.9958888...	0.9946629...	0.0629304...	0.0280469...	0.0001981...	0.0282231...	0.9978731...	0.0219634...	0.9972490...

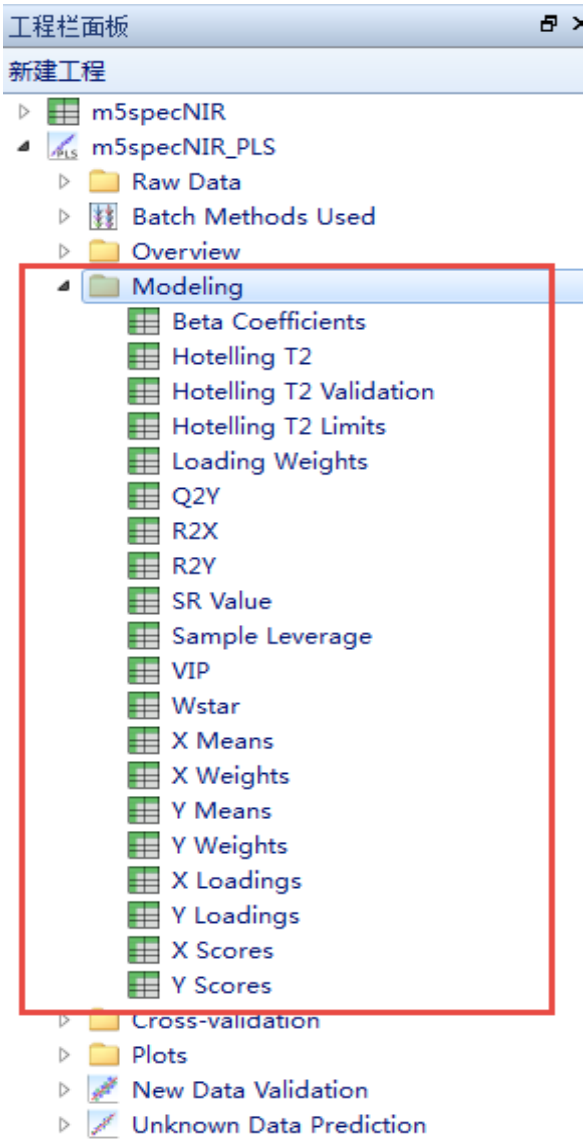
各结果的具体含义如下表所示。

序号	名称	说明
1	R2y	模型相关系数(校正集样本)，以下式计算。 $$R^2 = \sqrt{1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}}$$
2	Q2y	预测集样本相关系数，好的模型不仅在于 R2y 值较大，且 R2y 与 Q2y 的差值较小。

3	PRESS	预测残差平方和，以下式计算。 $$\mathrm{PRESS} = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$
4	RMSEP	预测均方根偏差，以下式计算。 $$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{N}}$$
5	Bias	$$\mathrm{Bias} = \frac{\sum_{i=1}^{N} \hat{y}_i - y_i}{N}$$
6	SEP	预测标准偏差，以下式计算。 $$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i - y_i - Bias)^2}{N}}$$
7	Slope	模型斜率，以下式计算。 $$\mathrm{Slope} = \frac{N\sum_{i=1}^{N} \hat{y}_i y_i - \sum_{i=1}^{N} \hat{y}_i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} {y_i}^2 - (\sum_{i=1}^{N} y_i)^2}$$
8	Offset	模型截距，以下式计算。 $$\mathrm{Offset} = \frac{1}{N}(\sum_{i=1}^{N} \hat{y}_i - slope \sum_{i=1}^{N} y_i)$$
9	Correlation	相关系数，以原始量测与模型预测计算。

12.13.4. Modeling 节点

PLS 分析所得到的模型节点结果，如下图所示。



PLS 分析所得到的主要结果，其具体意义如下表所示。

序号	名称	说明
1	得分	Scores，与 PCA 得分类似，差异仅在于 PLS 考虑 X 与 Y 间的关联性。
2	T 得分	T Scores 数据矩阵 X 得分，提取 X 中最大化信息以预测 Y 值。
3	U 得分	U Scores 响应变量 Y 得分，提取被数据 X 给定因子数解释的响应值 Y 结构信息。

4	载荷	Loading，与 PCA 载荷类似，其差异性如上所述。
5	P 载荷	P Loadings，数据矩阵 X 载荷。
6	Q 载荷	Q Loadings，响应变量 Y 载荷。需要注意的是与 PCA 不同，PLS 载荷未标准化，P 与 Q 载荷标度不一致。
7	载荷权重	Loading Weights，表达 X 中各变量与 Y 得分 u 的关系。

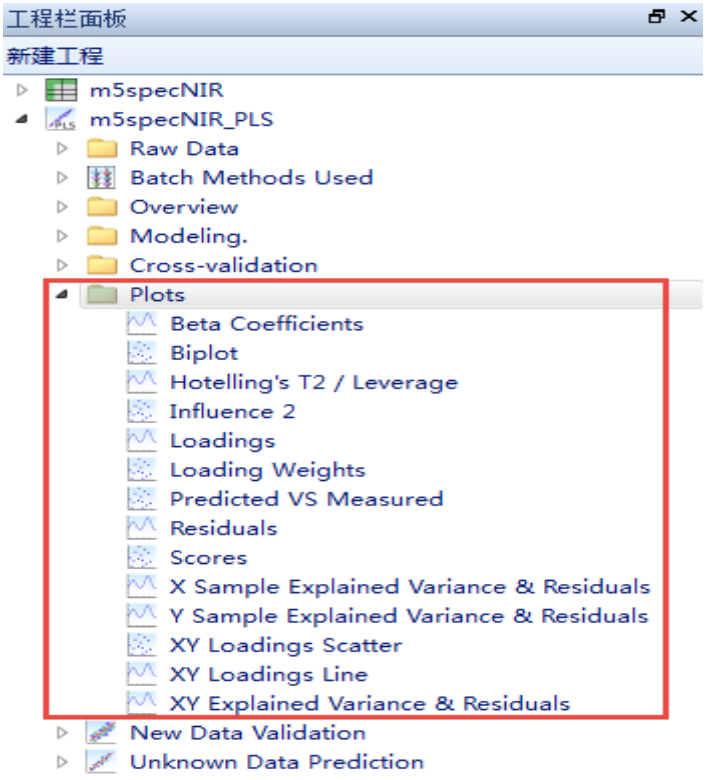
模型节点中各部分的具体介绍，请参见 12.7.4.，不再赘述。

12.13.5. Cross-validation 节点

本软件提供非常丰富的模型分析及评价结果，请参见 12.7.5.，不再赘述。

12.13.6. Plots 节点

该节点集中 PLS 分析所得到的图形结果，如下图所示。



12.13.6.1. 图形数据来源

序号	图形	说明
1	Beta Coefficients	数据取自 Modeling 节点文件夹下的 Beta Coefficients。
2	Bi-plot	由二部分构成，X Scores 取自 X Scores，而 X Loadings 则取自 X Loadings。
3	Explained	对校正集，选中被解释信息时，表达 Explained X Total Variance，而选中残差时，则表达 X Total Residuals。
		对验证集，选中被解释信息时，表达 Explained X Total Validation Variance，而选中残差时，则表达 Explained y Total Validation Variance。
4	Hotelling's T2 / Leverage	当选中 T2 值时，表达 Hotelling's T2，而选中杠杆值时，则表达 Sample Leverge。 当选中 T2 值时，红色水平线表达 Hotelling's T2 Limits。
5	Influence2	对校正集，X 轴选中 T2 值时为 Hotelling's T2，而选中杠杆值时则为 Sample Leverage；Y 轴则为 X Sample Q-Residuals。
		对验证集，X 轴选中 T2 值时为 Hotelling's T2 Validation，而选中杠杆值时则为 Test Sample Leverage；Y 轴则为 X Sample Q-Residuals。
6	Loadings	即为 X Loadings。
7	Loading Weights	选中 X 时为 Modeling 节点文件夹下的 Loading Weights，而选中 y 时，则为 Modeling 节点文件夹下的 y Loadings。
8	Predicted VS	对校正集，X 轴取自 Raw Data y，而 y 轴则取自 Cross-validation 节点

	Measured	文件夹下的 y Predicted Calibration。
		对校正集，X 轴取自 Raw Data y，而 y 轴则取自 Cross-validation 节点文件夹下的 y Predicted Validation。
9	Residuals	即为 X Sample Q-Residuals。
10	Scores	对校正集为 X Scores，对验证集则为 Test Sample Scores。
11	X Sample Explained Variance & Residuals	对校正集，选中被解释信息时，表达 Explained X Sample Variance，而选中残差时，则为 X Sample Residuals。
		对验证集，选中被解释信息时，表达 Explained X Sample Validation Variance，而选中残差时，则为 X Sample Validation Residuals。
12	Xy Loadings Scatter	X 选择时，为 X Loadings；y 选择时，则为 y Loadings。
13	Xy Loadings Line	X 选择时，为 X Loadings；y 选择时，则为 y Loadings。
14	Xy Explained Variance & Residuals	对校正集，选中 X，且被解释信息选中时，表达为 Explained X Total Variance，而残差选中时，表达为 X Total Residuals。 选中 y，且被解释信息选中时，表达为 Explained y Total Variance，而残差选中时，则表达为 y Total Residuals。
		对验证集，选中 X，且被解释信息选中时，表达为 Explained X Total Validation Variance，而残差选中时，表达为 X Total Validation Residuals。 选中 y，且被解释信息选中时，表达为 Explained y Total Validation Variance，而残差选中时，则表达为 y Total Validation Residuals。

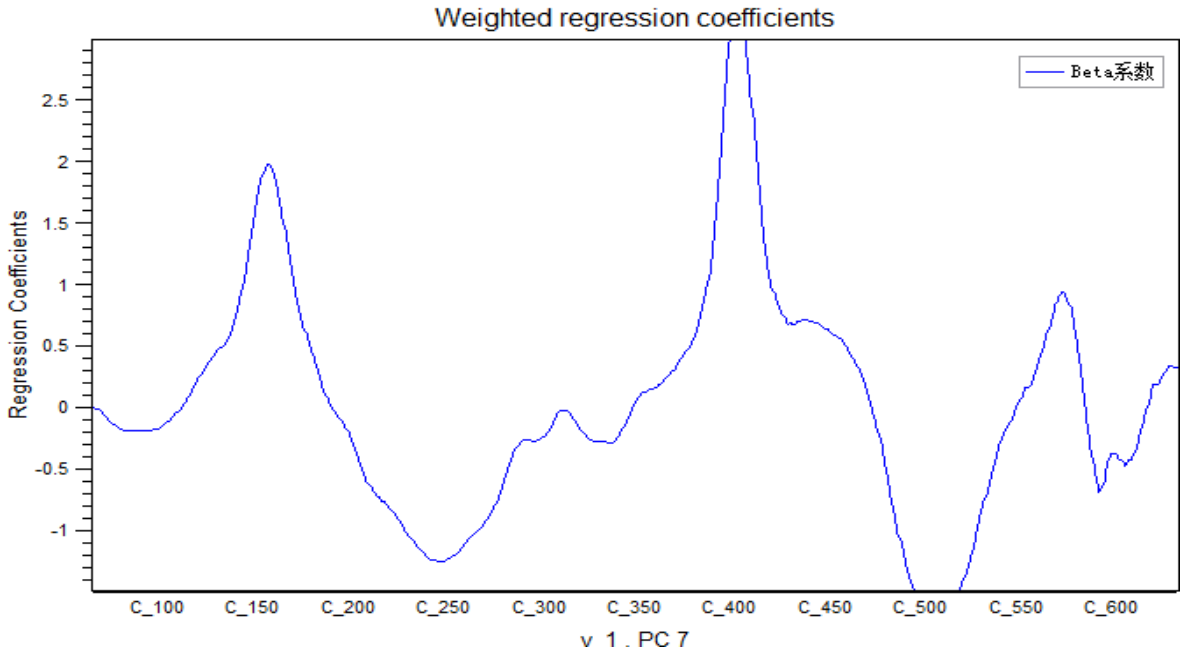
上述主要图形的操作使用及意义解释与 PCA 雷同，请参见 12.2.7.。PLS 分析增加的图形结

果，一一介绍如下。

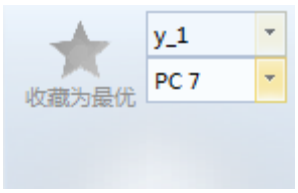
12.13.6.2. Beta Coefficients

具体操作步骤等内容不再赘述，用户可参考 9.2.1.1.，以及 12.2.7.2.部分。

该图的初始状态如下图所示。



除图形的基本工具外，在图形工具栏中同时增加如下图所示的功能。

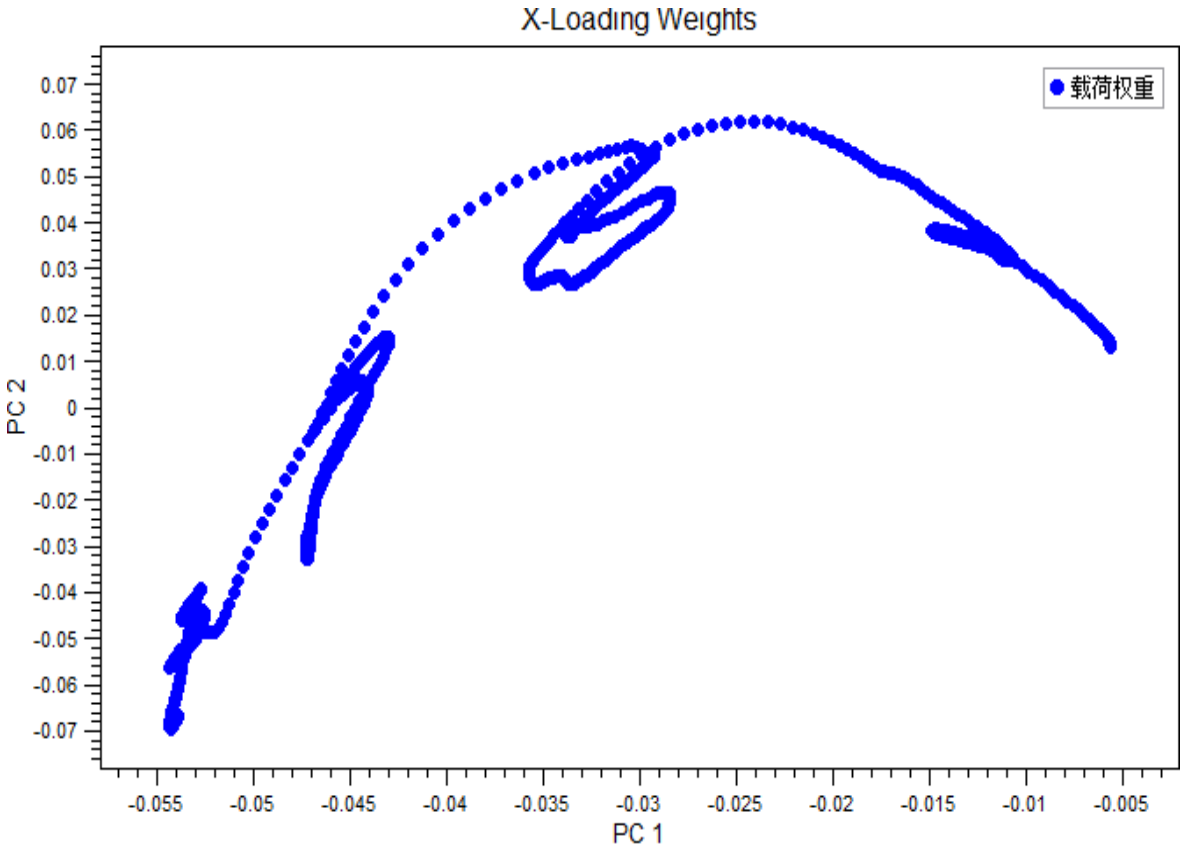


具体使用方法，以及图形属性修改，则不再赘述，请参考 12.2.7.。

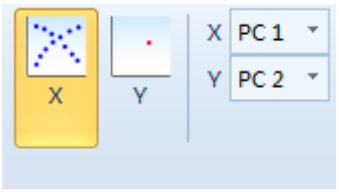
12.13.6.3. Loading Weights

具体操作步骤等内容不再赘述，用户可参考 9.2.1.1.，以及 12.2.7.2.部分。

该图的初始状态如下图所示。



除图形的基本工具外，在图形工具栏中同时增加如下图所示的功能。

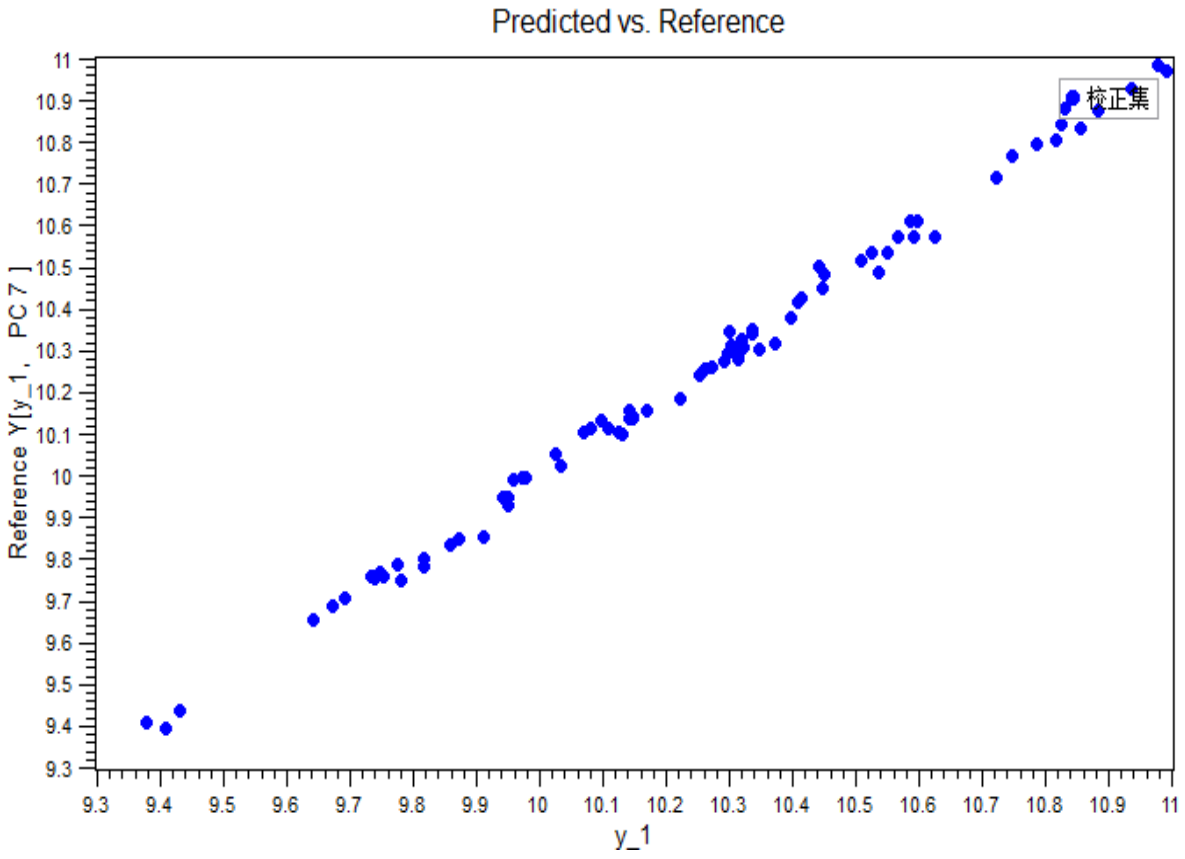


具体使用方法，以及图形属性修改，则不再赘述，请参考 12.2.7.。

12.13.6.4. Predicted VS Measured

以原始数据的 **y** 值作横坐标，而模型预测值作纵坐标绘图，是模型所得结果最直观的表达形式。

具体操作步骤等内容不再赘述，用户可参考 9.2.1.1.，以及 12.2.7.2.部分。该图的初始状态如下图所示。



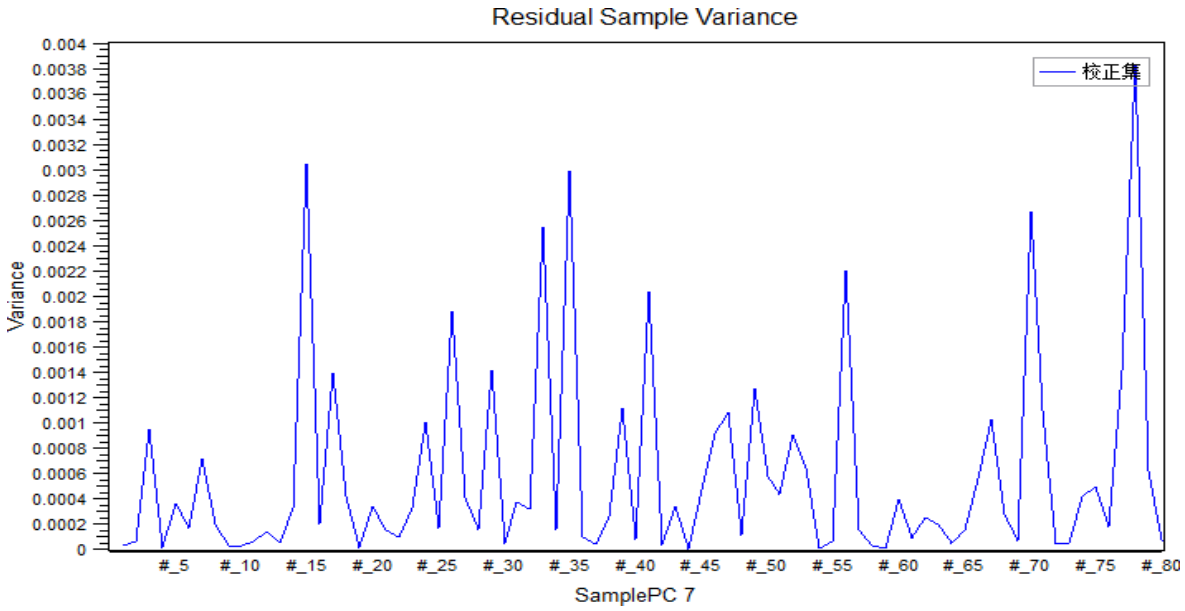
图中所显示的主成分数，与参数设置有关。除图形的基本工具外，在图形工具栏中同时增加如下图所示的功能。

具体使用方法，以及图形属性修改，则不再赘述，请参考 12.2.7.。

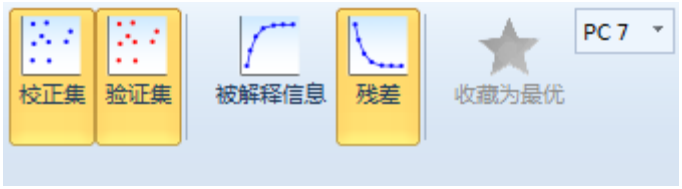
12.13.6.5. y Sample Explained Variance & Residuals

具体操作步骤等内容不再赘述，用户可参考 9.2.1.1.，以及 12.2.7.2.部分。

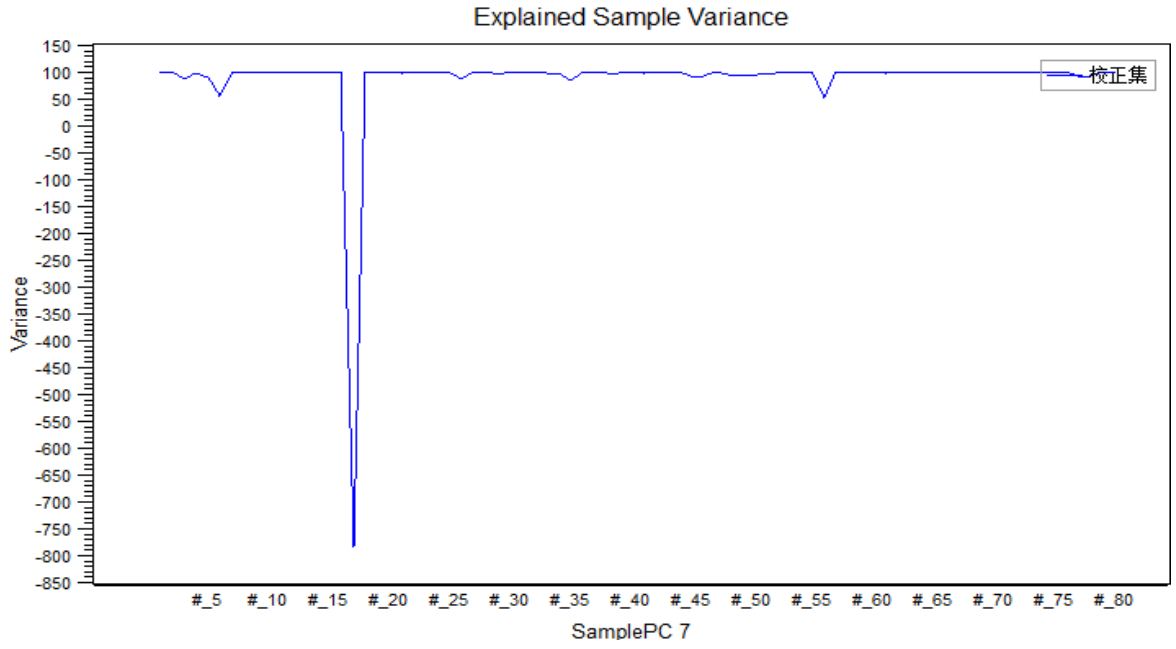
该图的初始状态如下图所示。



除图形的基本工具外，在图形工具栏中同时增加如下图所示的功能。



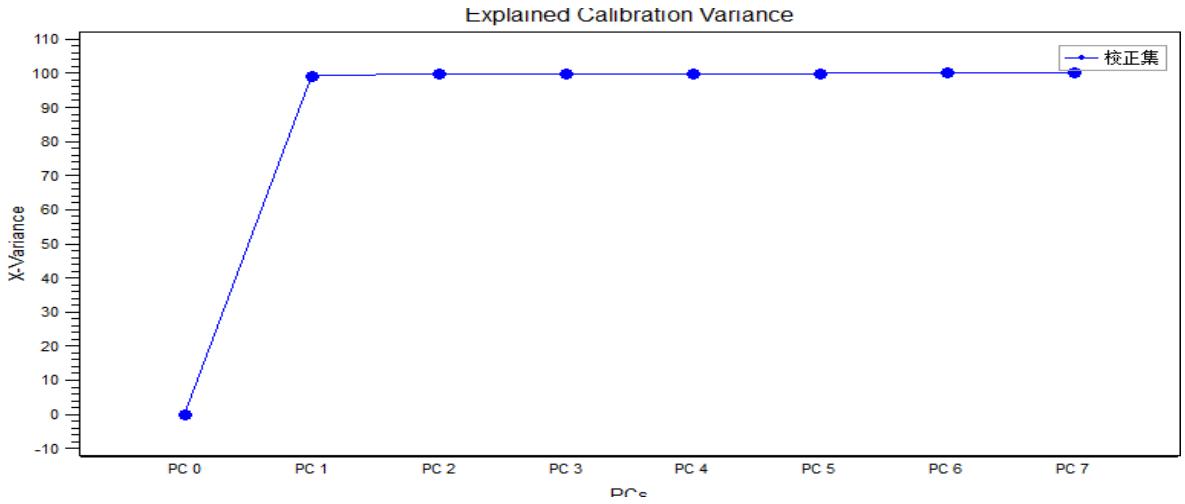
具体使用方法，以及图形属性修改，则不再赘述，请参考 12.2.7.。若表达信息由残差改为被解释信息，则得到如下图所示的结果。



12.13.6.6. Xy Explained Variance & Residuals

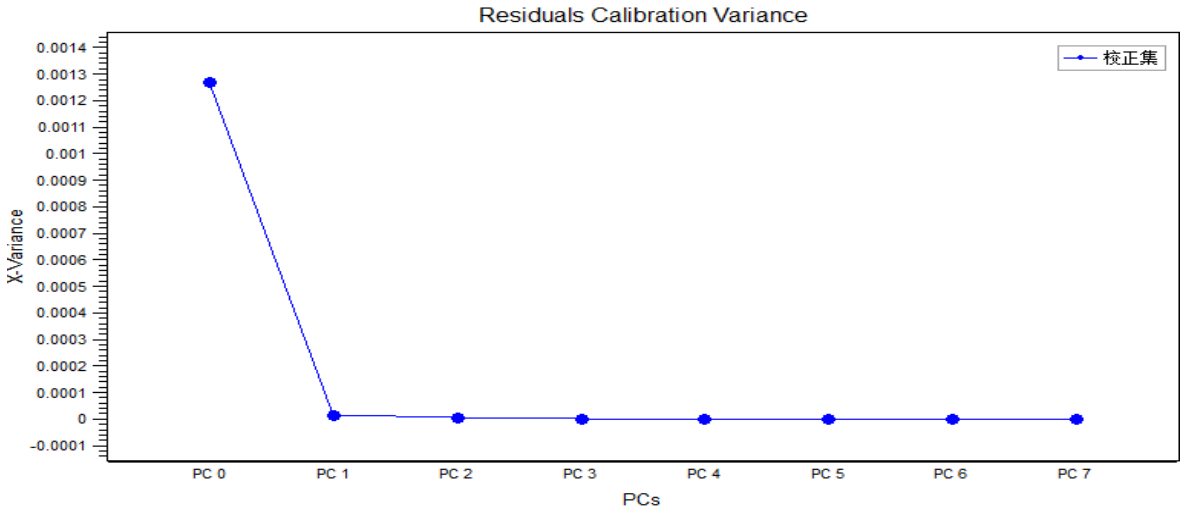
具体操作步骤等内容不再赘述，用户可参考 9.2.1.1.，以及 12.2.7.2.部分。

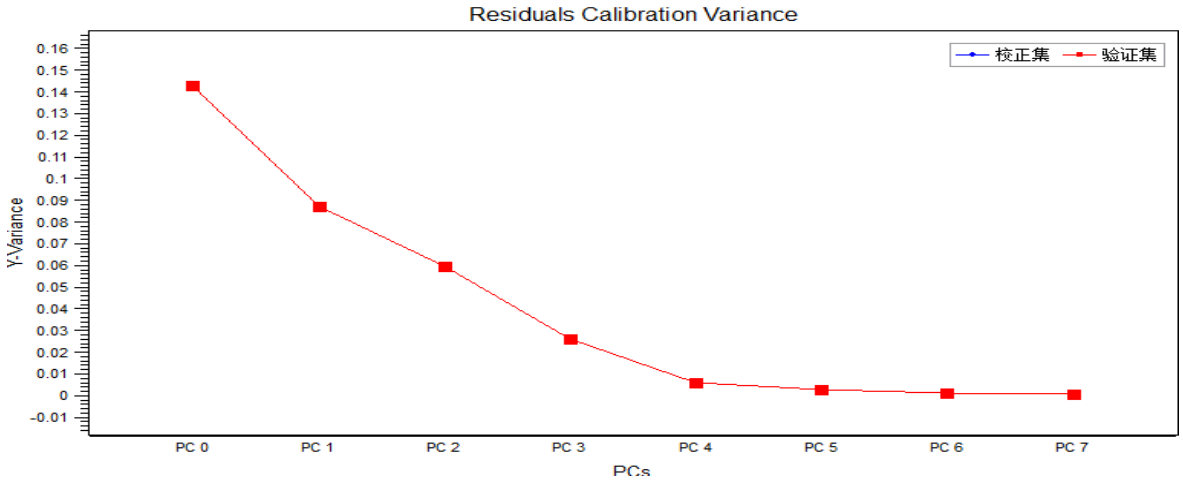
该图的初始状态如下图所示。



除图形的基本工具外，在图形工具栏中同时增加如下图所示的功能。

具体使用方法，以及图形属性修改，则不再赘述，请参考 12.2.7.。若表达信息变为残差，则得到如下图所示的结果。

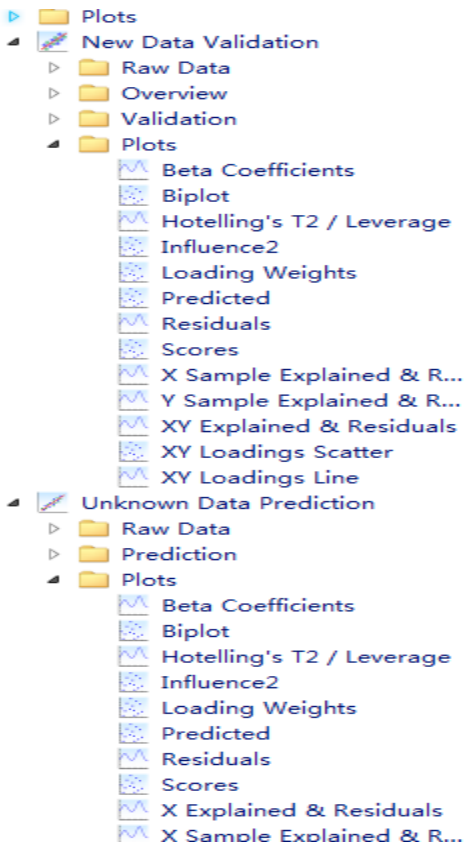




校正集分析结果，不再赘述，亦请参见 12.2.7.。

12.13.7. 预测与验证

若在构建模型时同时选择验证与预测集数据，则建模完成后，将得到 New Data Validation 与 Unknown Data Prediction 二个节点文件夹，如下图所示。这二个节点文件夹下的结果，分别对应验证和预测集结果。其表格和图形结果与校正集结果雷同，不再赘述。



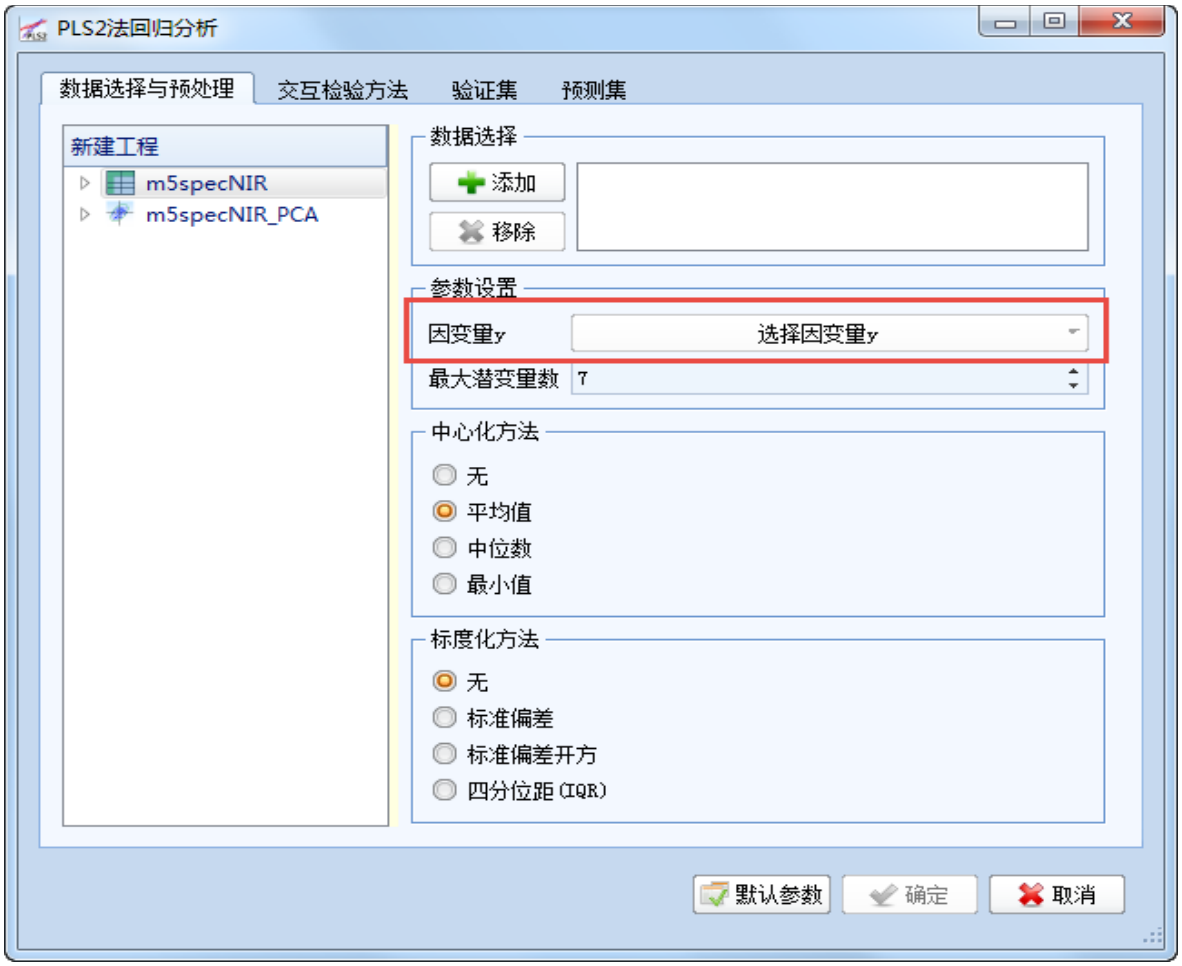
12.14. PLS2 法

PLS2 法与 PLS1 对应，用户可直接参考上一节中对 PLS1 法的介绍。该法的差异在于可统一构建数据矩阵 **X** 与多个响应变量 **y** 的模型，即响应值为矩阵 **Y**，其优越性在于同时考虑不同响应变量间的相互影响。因而若不同 **y** 值间非独立存在，即存在某种关联性，而分析的目的在于从整体上研究他们与数据矩阵的定量关系，则 PLS2 是不错的选择。

需要注意的是，对多个响应变量 **y** 单独建模，有可能得到的结果。

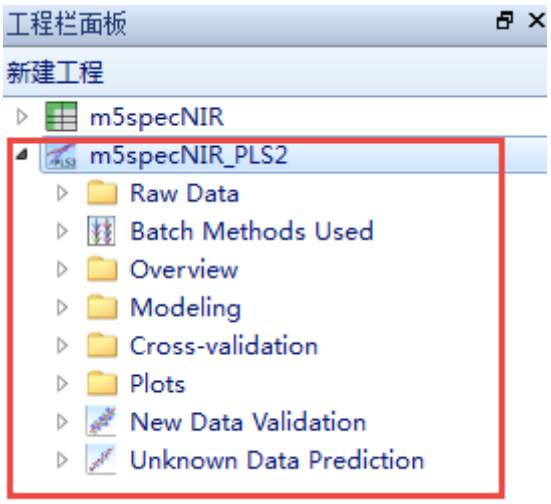
12.14.1. 操作说明

PLS2 法的使用步骤，与 PLS1 类似，可参考 12.1.2.以及上一章中的内容，初始界面如下图所示。其差异仅在于选择响应变量 **y** 时，需同时选择二个及以上 **y** 值。

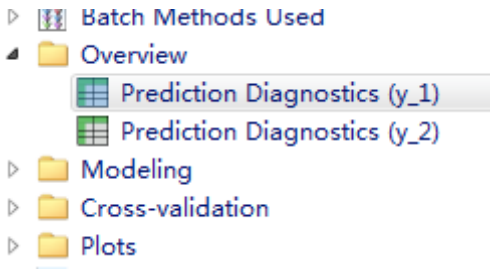


12.14.2. 模型结果概述

PLS2 分析所得到的模型结果节点文件夹如下图所示，具体介绍亦请参考上一章中的内容。



各节点文件夹所对应的结果，除 Overview 外，与 PLS 雷同。Overview 下则包括被选的多个响应变量 y 的结果，如下图所示。



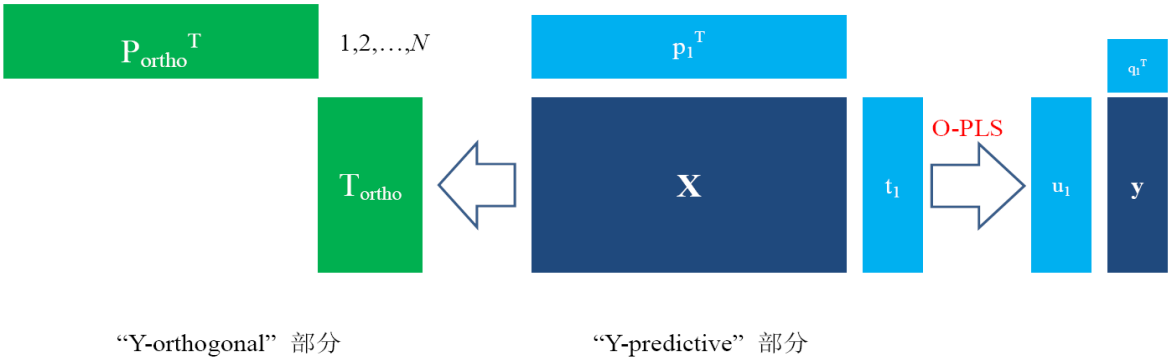
在本例中则同时包括 y_1 和 y_2 的二个结果，与 PLS 法雷同，具体结果如下图所示。

	Prediction	R2Y	Q2Y	PRESS	RMSEP	Bias	SEP	Slope	Offset	Correlation
PCs		1	2	3	4	5	6	7	8	9
PC 1	1	0.3084944...	0.2970776...	7.6191642...	0.3086090...	-0.005335...	0.3085629...	0.3541245...	6.6042552...	0.5785366...
PC 2	2	0.4342708...	0.4135269...	5.9841456...	0.2734992...	-0.002914...	0.2734836...	0.5006058...	5.1076547...	0.6910850...
PC 3	3	0.5814345...	0.5494138...	4.0485858...	0.2249607...	-0.001359...	0.2249566...	0.6666317...	3.4101771...	0.8040042...
PC 4	4	0.6772934...	0.6155950...	3.7870891...	0.2175743...	0.0099922...	0.2173448...	0.6761926...	3.3236864...	0.8181836...
PC 5	5	0.8969543...	0.8563351...	0.9354817...	0.1081365...	-0.010067...	0.1076668...	0.8941933...	1.0727087...	0.9589376...
PC 6	6	0.9264323...	0.9178885...	0.1857564...	0.0481866...	-0.001373...	0.0481670...	0.9707346...	0.2981146...	0.9919365...
PC 7	7	0.9570349...	0.9449996...	0.1974419...	0.0496792...	-0.000738...	0.0496737...	0.9733312...	0.2721774...	0.9913731...

其余各部分，包括 Modeling、Cross-validation、Plots、New Data Validation 和 Unknown Data Prediction，均与 PLS 结果雷同，可参考上一章中的相关内容。

12.15. O-PLS 法

O-PLS 法是 PLS 的一个延伸，提出至今已获得广泛应用，尤其在组学数据的分析处理中，如代谢组学。该法通过修正传统 NIPALS 算法，以将数据分解为二个不同部分，即与响应变量 **y** 相互关联，以及与其正交(无关联)而 **X** 独有的二部分信息，即如下图所示的“Y-orthogonal”和“Y-predictive”二部分。



从上图可以看出，数据矩阵 **X** 与 **y** 的分解可由如下二式表示。“Y-predictive”中的第一潜变量涵括 **X** 与 **y** 间的最大变化与相关性，其方向可从 t_1 与 p_1 得到，而“Y-orthogonal” 则描述 **X** 与 **y** 中不相关的信息。

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{T}_{ortho}^T\mathbf{P}_{ortho}^T + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{t}_1\mathbf{q}_1^T + \mathbf{F}$$

O-PLS 法的优点主要包括二个方面，其一是通过上述数据分解提高模型的解释与诊断能力，其二则是提高结果的可视化表达能力，如 11.11 与 11.12 中所介绍的 S-plot 等。关于 O-PLS 的更多信息，请参考 18.7.。

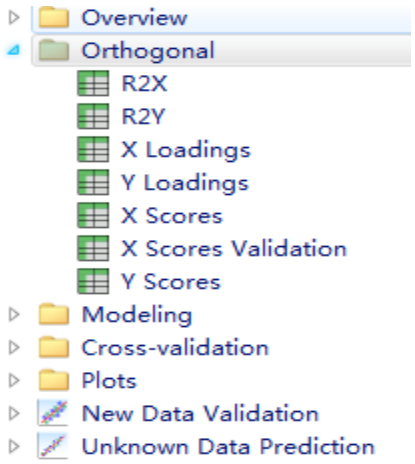
12.15.1. 操作说明

具体操作步骤等内容不再赘述，用户可参考 9.2.1.1.，以及 12.2.7.2.部分。

12.15.2. 模型结果概述

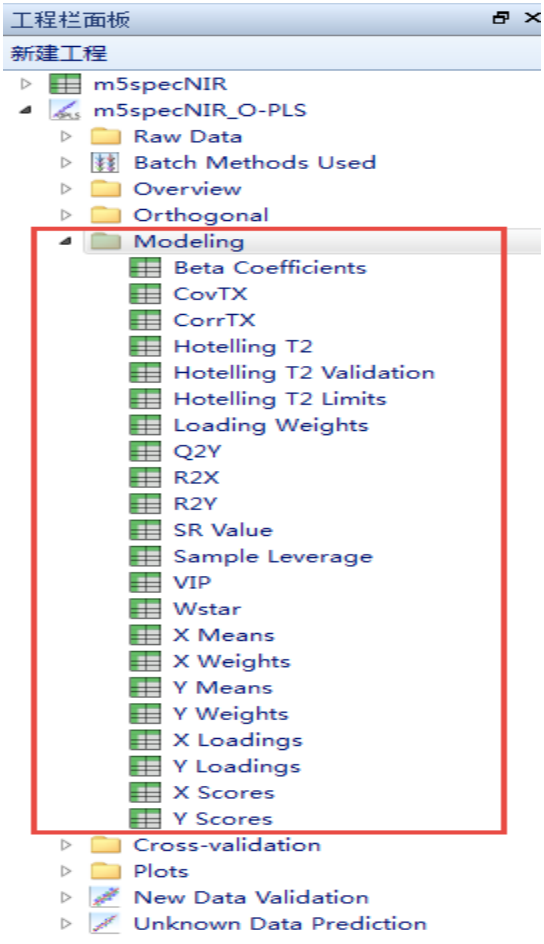
O-PLS 分析所得到的模型结果节点文件夹如下图所示。与 PLS 所得结果相比，增加了

Orthogonal 文件夹，其中所包括的具体节点结果，其意义则可从 PLS 结果找到，不再赘述。



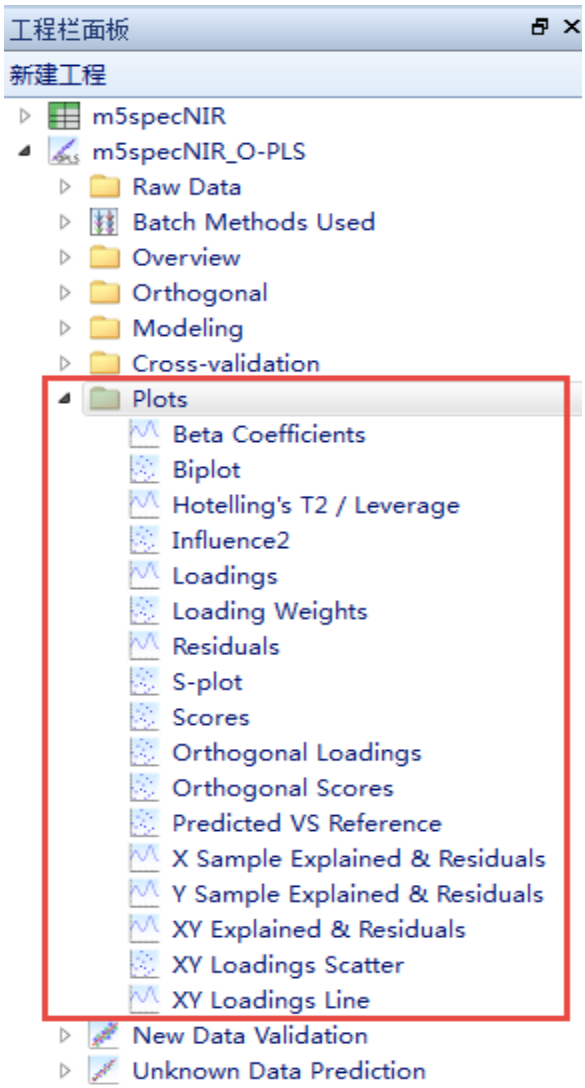
12.15.3. Modeling 节点

该节点文件夹下的结果如下图所示，从图中可看出其主要结果与 PLS 一致，增加的内容包括 CovTX 和 CorrTX，可参考 11.11.，不再赘述。



12.15.4. Plots 节点

该节点涵括 O-PLS 分析所得到的图形结果，如下图所示。



图中的绝大部分结果，已经在前面详细介绍，请参见 12.2.7.和 12.3.6.。增加的结果则包括 S-plot、Orthgonal Loadings 和 Orthogonal Scores，其数据来源概述为如下表。图形的操作与意义表达亦不再赘述。

序号	图形	说明
1	S-plot	其 X 和 Y 坐标值分别为 Modeling 节点下的 CovTX 和 CorrTX。

2	Orthogonal Loading	其 X 和 Y 坐标值分别为 Orthogonal 下的 X Loadings 与 Modeling 下的 X Loadings。
3	Orthogonal Scores	对校正集，其 X 与 Y 坐标值分别为 Orthogonal 下 X Score 与 Modeling 下的 X Scores。
		对验证集，其 X 与 Y 坐标值分别为 Orthogonal 下 X Score Validation 与 Modeling 下的 Test Sample Scores。

若选择验证集与预测集数据，则其结果同样以 New Data Validation 和 Unknown Data Prediction 节点文件夹的形式产生在工程导航栏中，所包含的具体节点信息及结果解释不再赘述，内容完全包括在上述内容中。

12.16. SVR 法

SVR 法与上述各回归方法具有本质的差别，可基于 Kernel 函数将非线性问题转换为高维空间上的线性问题。与 12.10.中所介绍的 SVC 法雷同，SVR 可由其扩展而来，算法本质一致，区别在于 SVR 中所寻找的最优超平面，是使得其与所有样本点的总偏差最小，而不是 SVC 中使不同类别样本点间的距离最大，即“分得最开”。

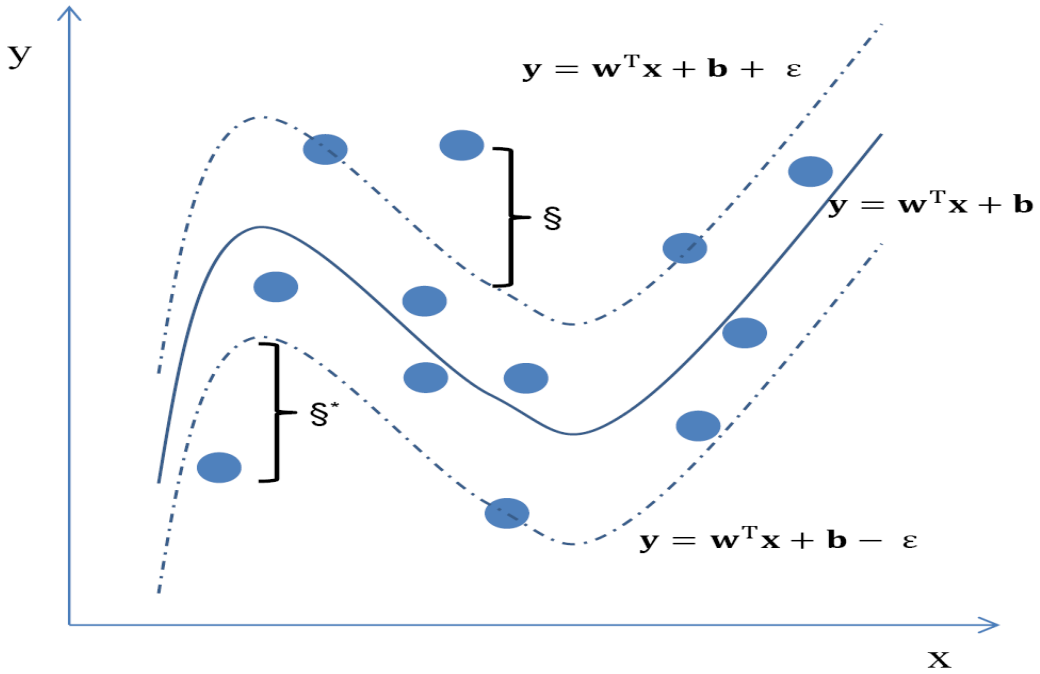
以下图为例，SVR 分析过程在于寻找最优的 **w** 和 **b** 值，以 $f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + \mathbf{b}$ 拟合数据点$<x_i, y_i>$, $i = 1, 2, \ldots, n$。与此同时引入松弛变量§和§*，构造如下式所示的优化目标。

$$\min_{\mathrm{w,b,\S}} \frac{||\mathbf{w}||^2}{2} + C\sum_{i=1}^{N}(\S_i + \S_i^*)$$

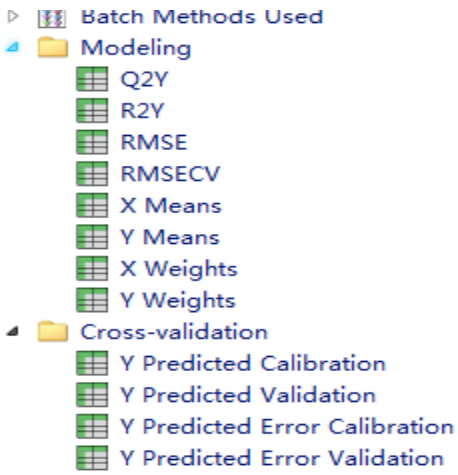
约束条件为:

$$y_i - w^{\mathrm{T}}\mathrm{x}_i - \mathrm{b} \le (\varepsilon + \S_i),\ \S_i \ge 0$$

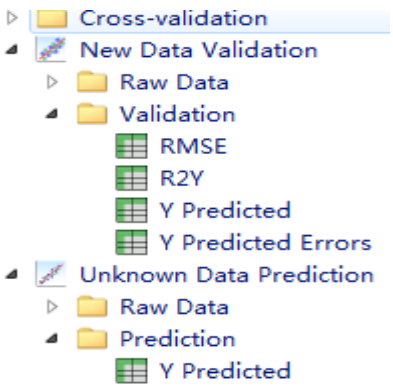
$$w^{\mathrm{T}}\mathrm{x}_i + \mathrm{b} - y_i \le (\varepsilon + \S_i^*),\ \S_i^* \ge 0,\ i = 1, 2, \ldots, N$$



接下来的优化求解请参见 18.8.，不再赘述。SVR 所得到的结果如下图所示。



图中的结果前面已经做了详细介绍，可参考 12.13，在此不再赘述。



第十三章　预测

构建好的模型，其目的在于应用，即将模型应用于新样本的验证或预测，并更进一步评价模型的表现。如前所述，本软件基于算法流实现智慧型的数据处理，在往算法流中“注入”需要处理的数据时，可同时选择训练集，验证集和预测集，从而完成验证或预测功能。本章所述预测，是独立于算法流之外实现新样本的验证或预测。

新数据验证或预测前，需使用构建模型时相同的预处理方法处理，然后将模型作用于所得到的数据即得到结果。本软件提供丰富的新样本验证或预测结果评价方法。

13.1. 新样本验证

13.1.1. 分类

基于已知模型对验证集数据进行分类分析。

操作步骤:

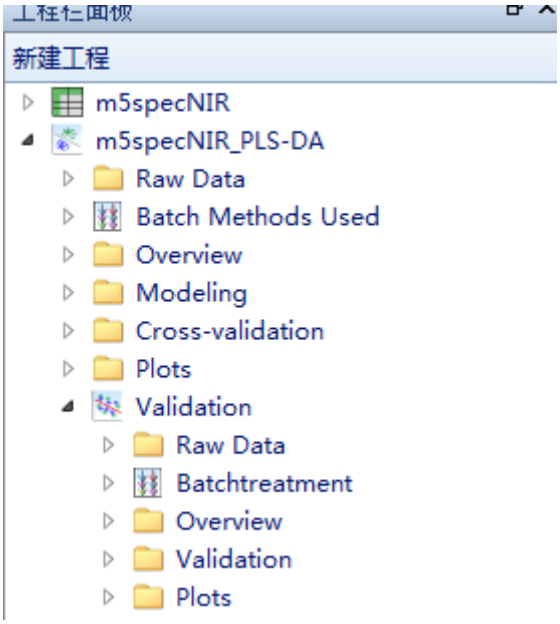
步骤 1: 点击**预测** -> **分类分析(新样本验证)**，弹出如下对话框:

步骤 2: 从左侧模型节点中选择已知模型(分类模型)，然后选择被验证的数据完成验证。

若该模型中包含预处理方法，则可点击界面中的预处理方法选择按钮，弹出如下对话框以完成选择。

步骤 3：点击**确定**或**应用**即可开始运算，点击**应用**则继续停留在此界面，可继续操作。点击**取消**，则取消操作并关闭对话框。

完成新样本的验证后，在工程导航栏中产生如下图所示的结果。详细信息可参考上一章中对各建模方法的介绍。



13.1.2. 回归分析

回归分析中的预测与分类雷同，其仅差异在于选择模型时，需添加已知回归模型。新样本验证完成后，得到如下图所示的结果，详细结果亦可参考上一章中的介绍。

13.2. 预测

本部分与上一节雷同，其差异仅在于验证是针对已知响应变量 **y** 的情形，而预测则是针对响应变量 **y** 未知的情形。

ⓘ 但需要注意的是，实因验证分析中的响应变量 **y** 值已知，因而其评价结果与预测完全不同，可提供的信息亦更加丰富。

探索性分析，以及分类与回归分析的使用雷同，如下图所示。不再赘述。结果的介绍亦可参考上一章中的相关内容，亦不再赘述。

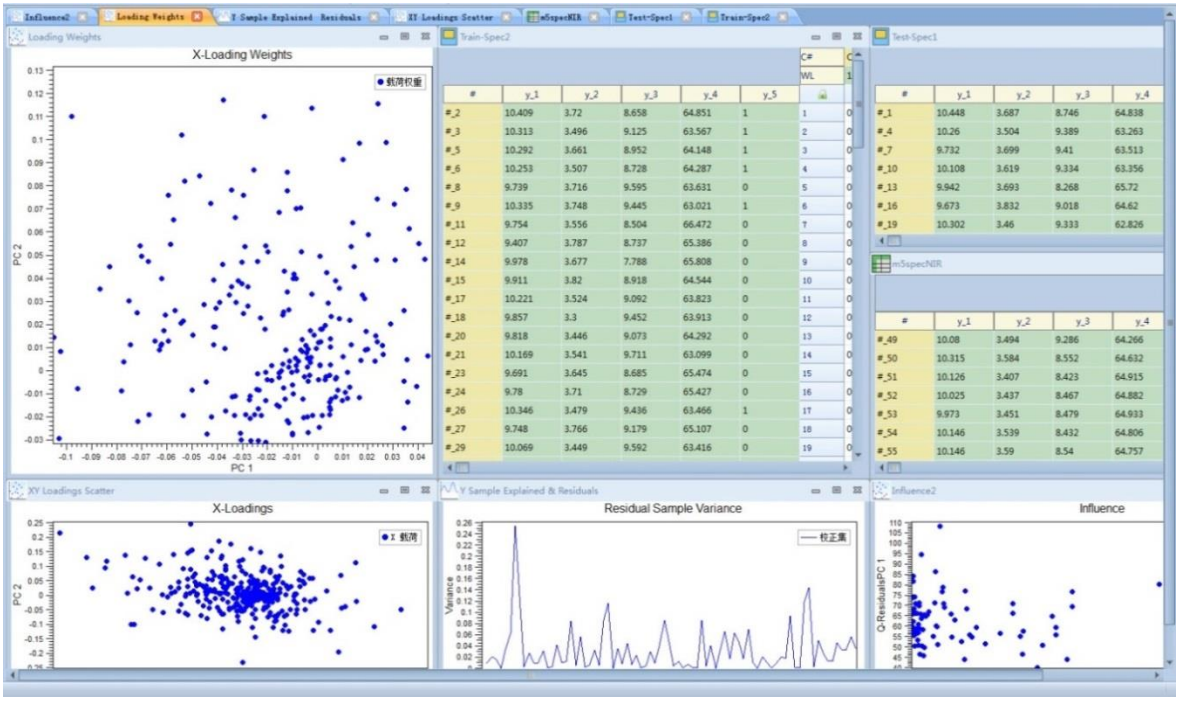
第十四章 窗口

本软件窗口功能强大，具体如下图所示。

14.1. 平铺窗口

将工作区内已打开的窗口平铺显示。

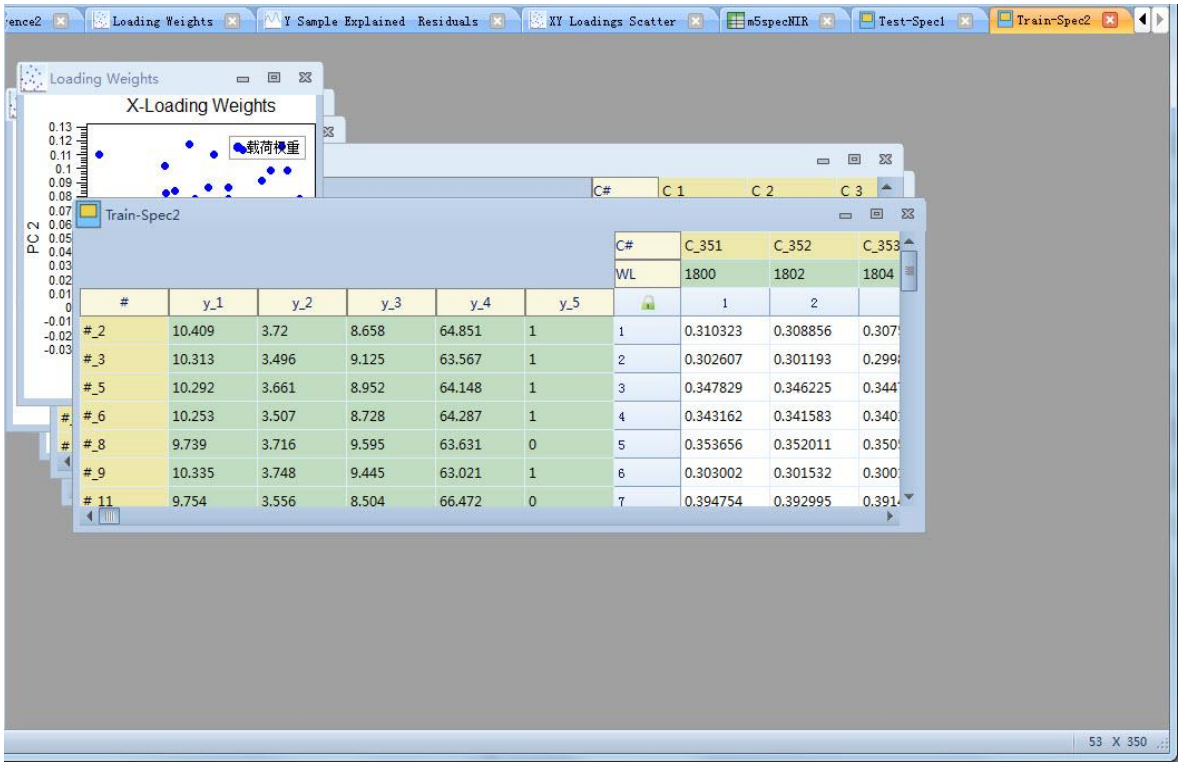
操作步骤为：点击**窗口** -> **平铺窗口**即可，效果如下图:



14.2. 层叠窗口

将工作区内已打开的窗口层叠显示。

操作步骤为：点击**窗口** -> **层叠窗口**即可，效果如下图:



14.3. 上一个活动窗口

显示当前窗口的上一个活动窗口。

操作步骤为：点击**窗口** -> **上一个活动窗口**即可。

14.4. 下一个活动窗口

显示当前窗口的下一个活动窗口。

操作步骤为：点击**窗口** -> **下一个活动窗口**即可。

14.5. 关闭所有窗口

关闭工作区内所有已打开的窗口。

操作步骤为：点击**窗口** -> **关闭所有窗口**即可。

14.6. 关闭当前窗口

关闭工作区内当前已显示窗口。

操作步骤为：点击**窗口** -> **关闭当前窗口**即可。

14.7. 关闭其它窗口

关闭工作区内除当前被显示窗口外的所有其他窗口。

操作步骤为：点击**窗口** -> **关闭其它窗口**即可。

14.8. 关闭左侧窗口

关闭工作区内当前已显示窗口的左侧窗口。

操作步骤为：点击**窗口** -> **关闭左侧窗口**即可。

14.9. 关闭右侧窗口

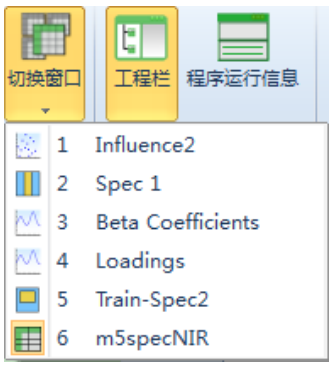
关闭工作区内当前已显示窗口的右侧窗口。

操作步骤为：点击**窗口** -> **关闭右侧窗口**即可。

14.10. 切换窗口

实现工作区内已打开窗口间的快速切换。

操作步骤为：点击**窗口** -> **切换窗口**即可看到当前被打开的所有窗口，如下图所示。



选择窗口列表中的任一窗口，即可切换到该窗口。

14.11. 工程栏

显示/隐藏用户界面左侧的工程导航栏。

操作步骤为：点击**窗口** -> **工程栏**即可。默认状态下工程导航栏处于打开状态，方便用户浏览当前操作。

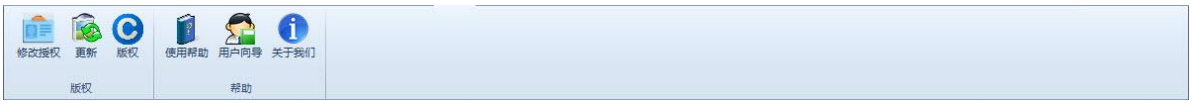
14.12. 程序运行信息

显示/隐藏程序运行信息窗口。

操作步骤为：点击**窗口** -> **程序运行信息**即可。默认状态下该窗口处于打开状态，以方便用户查看程序运行信息。

第十五章 帮助

软件中帮助菜单提供如下图所示的功能。



15.1. 修改版权

本部分介绍暂略。

15.2. 更新

本部分介绍暂略。

15.3. 版权

本部分介绍暂略。

15.4. 使用帮助

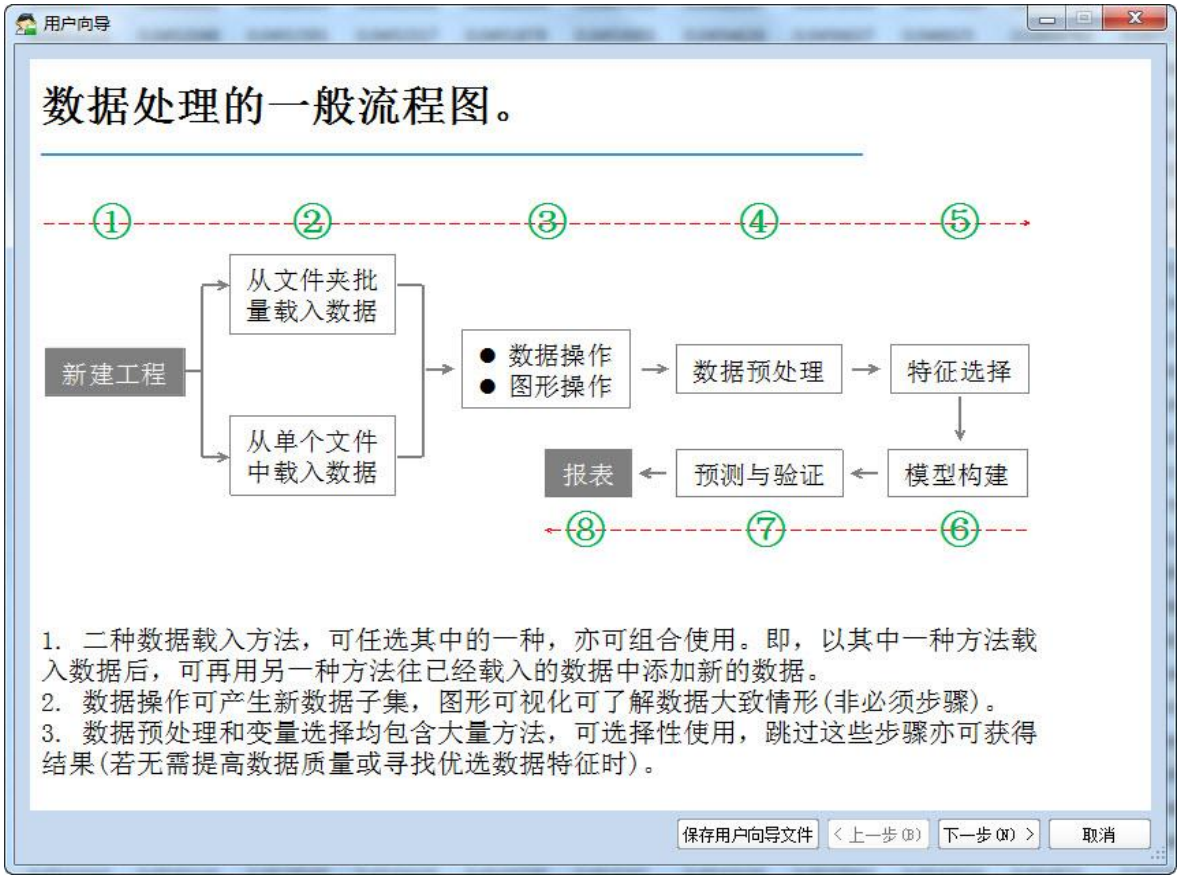
本部分介绍暂略。

15.5. 用户向导

用户向导是本软件具体使用方法的快速入门。与第二章内容相比，用户向导更简洁，且以图形形式表示，尤其是其中的关键操作功能与软件的实际功能关联，用户可以从用户向导部分直接使用软件功能。

操作步骤:

步骤 1: 点击**帮助** -> **用户向导**，弹出如下初始对话框:

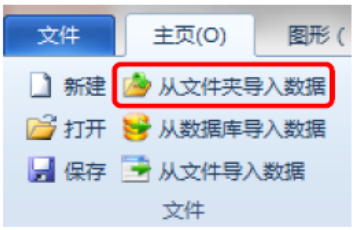


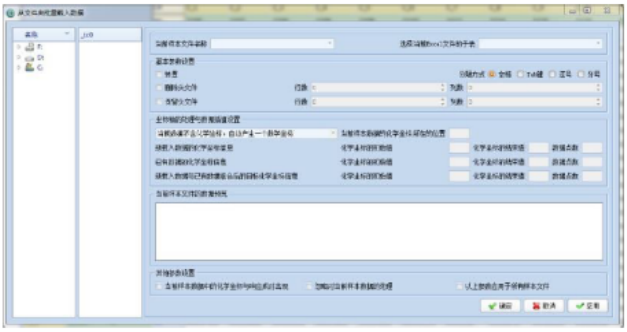
步骤 2：依次点击**下一步**可查看更多内容，依次排列如下。点击**保存用户向导文件**则可将整个向导文件另存为 PDF 格式文档。

用户向导

第二步：往新建工程中导入待处理的数据

方法1：从文件夹批量载入数据。





单个载入数据的界面。

● 单击主菜单“主页”中“从文件夹导入数据”项，则出现右侧批量载入数据的窗口。

● 在数据载入窗口，用户可选择数据所在的文件夹路径，并选择数据格式，目前程序支持txt，csv，xls，xlsx，spc等数据的批量载入。

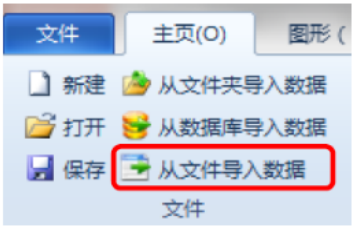
● 被选择格式的所有数据将出现在下拉式列表中，用户可设置载入参数，可单个载入数据，亦可将所设参数应用于所有文件，批量载入。

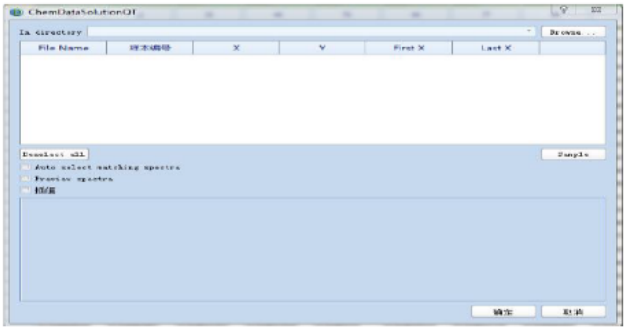
保存用户向导文件 | < 上一步 (B) | 下一步 (N) > | 取消

用户向导

第二步：往新建工程中导入待处理的数据

方法2：从单个文件中载入数据。





单个载入数据的界面。

● 单击主菜单“主页”中“从文件导入数据”项，则出现右侧所示数据载入的窗口。

● 根据被载入数据的不同格式，右侧数据载入窗口将有所不同。

● 在载入单个文件中的数据时，亦需要根据数据的实际情况设置有关参数，比如包含多个数据的文件载入，因数据长度的不同，系统设置了载入规则，这些规则将显示在载入的信息栏中。

保存用户向导文件 | < 上一步 (B) | 下一步 (N) > | 取消

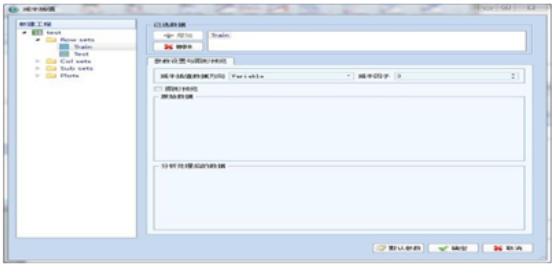

用户向导
第三步（1）：数据操作。
起始界面
task
C#
WL
C_1
C_2
C_3
C_4
1100
1102
1104
1106
#
y_1
y_5
1
2
3
4
#_1 10.448 1 1 0.0444948 0.0443834 0.0442581 0.0442124
#_2 10.409 1 2 0.0465041 0.0463485 0.0462297 0.0462051
#_3 10.313 1 3 0.0469579 0.046817 0.0466632 0.0466015
#_4 10.26 1 4 0.0454611 0.0453212 0.0452048 0.0451591
#_5 10.292 1 5 0.0539477 0.0537859 0.0536497 0.0536129
#_6 10.253 1 6 0.052083 0.0518756 0.0517733 0.0517475
#_7 9.732 0 7 0.0567156 0.0565167 0.0564035 0.0563486
#_8 9.739 0 8 0.056241 0.0560315 0.055933 0.055881
#_9 10.335 1 9 0.0487862 0.0485873 0.0484845 0.0484452
#_10 10.108 0 10 0.0492719 0.0490503 0.0489668 0.048934
#_11 9.754 0 11 0.0544335 0.0542774 0.0541613 0.0540967
#_12 9.407 0 12 0.0546683 0.0545415 0.0544006 0.0543259
数据导入后的基本数据表。
● 数据导入到数据表后，可对数据本身（X）或对应的Y值（属性值或类别），以及行或列方向的属性标注进行添加、修改、删除等操作。
● 行或列方向的属性标注的名称将在绘图和建模时使用到。
● 程序以样本作为数据矩阵的行，特征作为列。
产生子数据
转换成分类
剪切
复制
复制（带表头）
粘贴
插入（I）
添加（A）
删除（D）
查找
Range Find
Check Data
替换
升序排列
降序排列
跳到某一行
跳到某一列
Add to favorites
输出到数据库
数据表的右键菜单。
保存用户向导文件
< 上一步(B)
下一步(N) >
取消


用户向导
第三步（2）：图形。
1 2 3 4
0.0444948 0.0443834 0.0442581 0.0442124
0.0465041 0.0463485 0.0462297 0.0462051
0.0469579 0.046817 0.0466632 0.0466015
①从数据表中选择目标数据区域。
文件
主页(O)
图形（L）
曲线图
散点图
条形堆积图
填充图
②从主菜单的“图形”中选择所需绘制的图形类型。
Matrix
Select
Legend
Zoom Range
Create Range
Export
标注
Property
③所绘图形及图形中的右键菜单。
文件
主页(O)
图形（L）
预处理(T)
特征选择(V)
线条
箭头
矩形框
椭圆
文字
选择
框选放大
框选还
④将图形作为当前窗口，此时“图形”菜单中显示可对图形进行的操作功能。
保存用户向导文件
< 上一步(B)
下一步(N) >
取消

用户向导

第四步：数据预处理。

主菜单中的“预处理”栏。



点击“预处理” 栏中的方法后所出现的界面（以减半插值为例）。

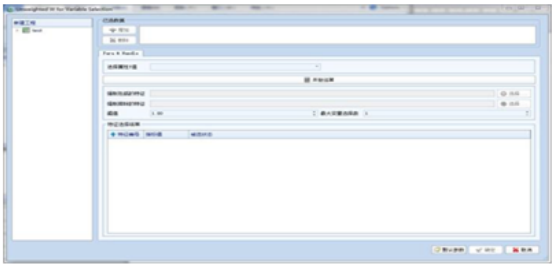
- 不同预处理方法中的左侧界面有所不同，但框架类似。
- 在左侧的界面中，先从其左边的数据列表中选择目标数据，再设置参数，可预览原始数据和预处理后的数据。
- 计算完成后将在主界面的左侧导航栏中产生新的节点，完整记载数据预处理的结果。

保存用户向导文件 | < 上一步(B) | 下一步(N) > | 取消

用户向导

第五步：特征选择。

主菜单中的“特征选择”栏。



点击“特征选择” 栏中的方法后所出现的界面（以不加权法为例）。

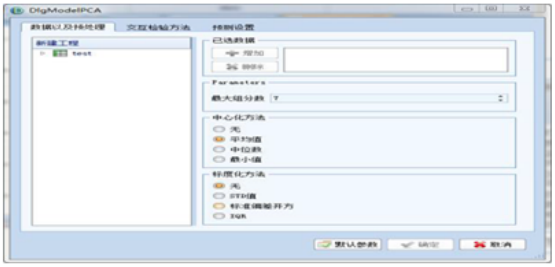
- 不同特征选择方法中的左侧界面有所不同，但框架类似。
- 操作时先从其左边的数据列表中选择目标数据，再设置参数，特征选择后可获得图形结果和被选特征的信息。
- 计算完成后将在主界面的左侧导航栏中产生新的节点，完整记载特征选择的结果。

保存用户向导文件 | < 上一步(B) | 下一步(N) > | 取消

用户向导

第六步：模型构建。

主菜单中的“建模”栏。



点击“建模” 栏中的方法后所出现的界面（以PCA法为例）。

- 不同建模方法中的左侧界面有所不同，但框架类似。

- 在左侧的界面中，先从其左边的数据列表中选择目标数据，再设置参数进行建模，计算完成后将在主界面的左侧导航栏中产生新的节点，完整记载所有建模结果，包括数据，图形，结果，模型。

保存用户向导文件 | < 上一步(B) | 下一步(N) > | 取消

用户向导

第七步：预测与验证。

主菜单中的“预测”栏。

点击“预测” 栏中的方法后所出现的界面（以分类问题为例）。

- 在左侧的界面中，先从其左边的已有模型中选择用于预测或验证的模型，再从左边的数据列表中选择目标数据，即可开始计算，计算完成后将在主界面的左侧导航栏中产生新的节点，完整记载所有预测和验证的结果。

保存用户向导文件 | < 上一步(B) | 下一步(N) > | 取消

用户向导

第八步：报表。

主菜单“主页”栏中的报表。

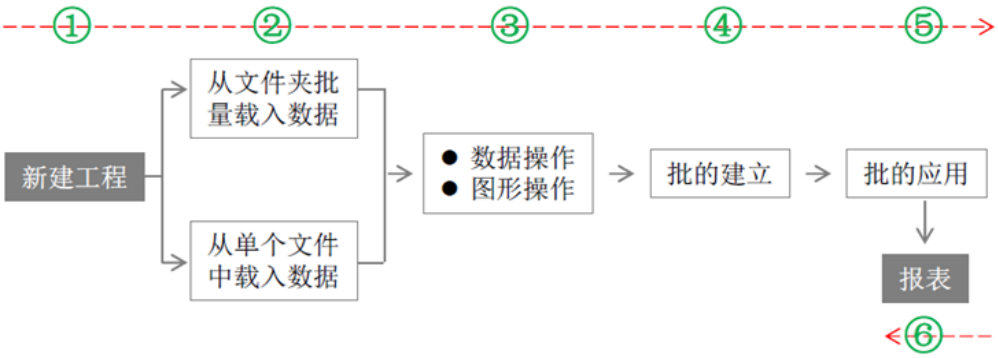
单击“新建报表”后的界面。

- 用户可从界面左边选择所需添加到报表中的内容，并可调整顺序或添加标题等，对于大的数据，可转置后再显示。
- 报表设置内容将作为报表的表头部分。
- 修改报表则先载入已有报表，再进行新的修改。

保存用户向导文件 | < 上一步 | 下一步 > | 取消

用户向导

基于“批”的数据处理流程图。



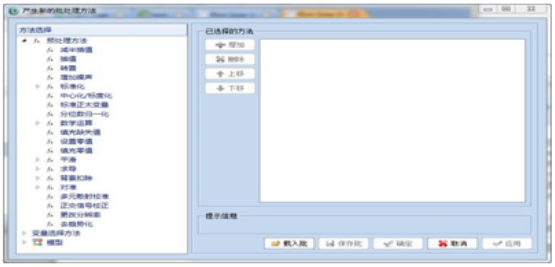
1. “批”是指用户先建立一个任意的数据处理流程，包括数据预处理、特征选择、模型构建、预测与验证等步骤，并选择流程中每个环节所需采用的数据处理方法；通过“批”的应用选择目标数据，实现快速的数据分析处理。
2. “批”是程序的亮点和特色，实现了数据处理的“一键化”和“多模型化” 。
3. 接下来仅介绍与前述内容中不同的部分。

保存用户向导文件 | < 上一步 | 下一步 > | 取消

用户向导

1，批的建立。

主菜单“主页”栏中的“新建批”。



单击“新建批”后的界面。

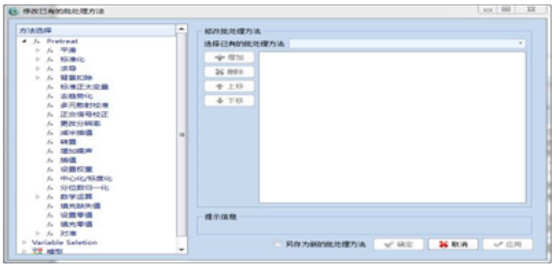
- 用户可从界面的左边选择需要添加到“批”中的方法，调整分析顺序。
- 用户可保存当前“批”，或载入已保存的“批”。
- 仅能选择一个相同子类的方法，但可选择多个建模方法。
- 新建“批”将在主界面导航栏中产生新的节点。

保存用户向导文件 | < 上一步 | 下一步 > | 取消

用户向导

2，批的修改。

主菜单“主页”栏中的“修改批”。



单击“修改批”后的界面。

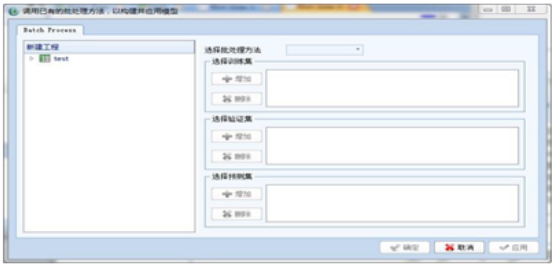
- 实现对已经建立好的“批”的修改，即从已有的“批”列表中选择一个“批”，重新进行其数据处理方法的增加、删除、以及顺序调整等。
- 修改后的“批”可覆盖被修改的“批”，亦可产生新的“批”。

保存用户向导文件 | < 上一步 | 下一步 > | 取消

用户向导

3，批的应用。

主菜单“主页”栏中的“应用批”。



单击“应用批”后的界面。

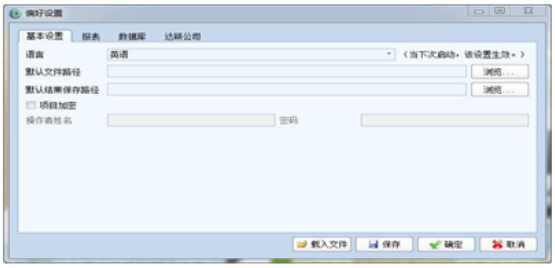
- 选择一个已经建立好的“批”，以及训练、验证和预测数据集即可。
- 训练、验证和预测数据集可不全选，即可选择其中的一个或多个。
- 交互检验和模型参数的设置根据“批”的不同，程序将自动动态调整。

保存用户向导文件 | 〈上一步 | 下一步〉 | 取消

用户向导

4，偏好设置。

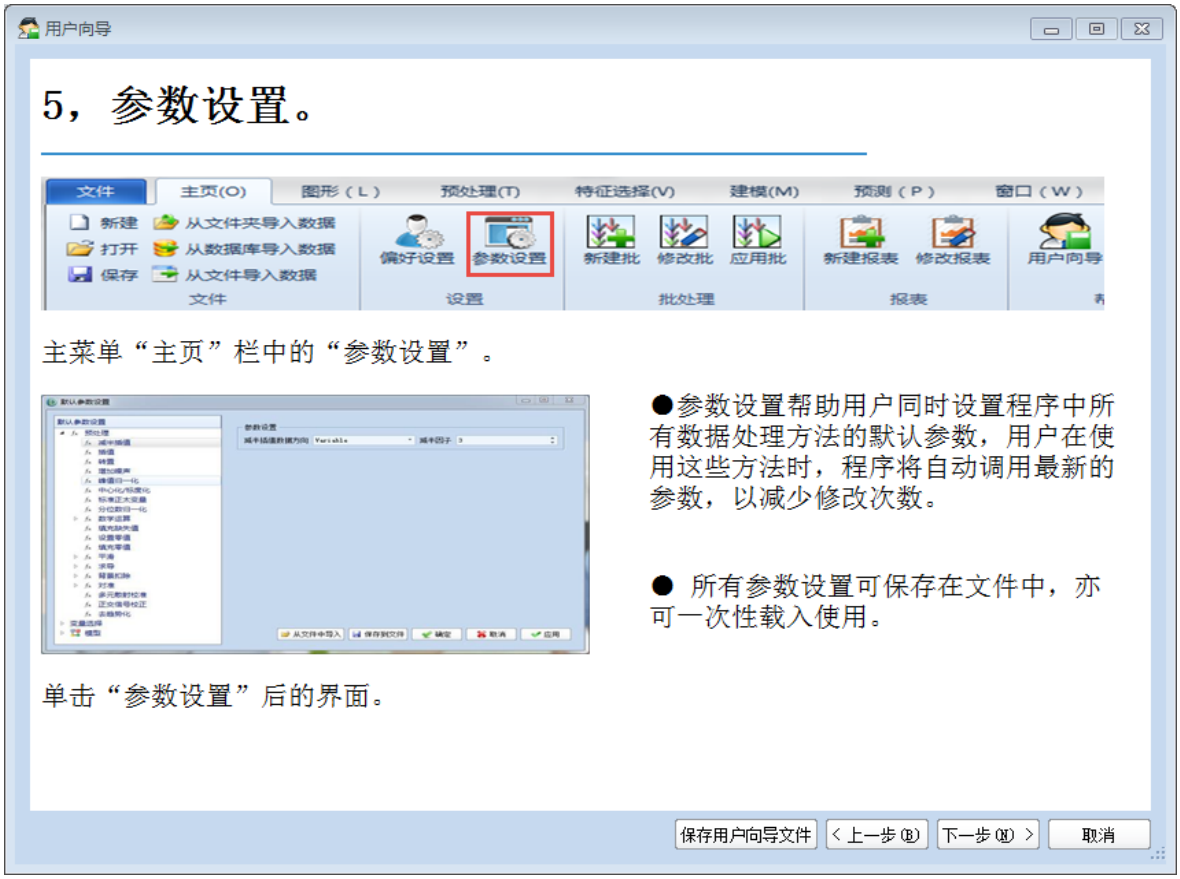
主菜单“主页”栏中的“偏好设置”。

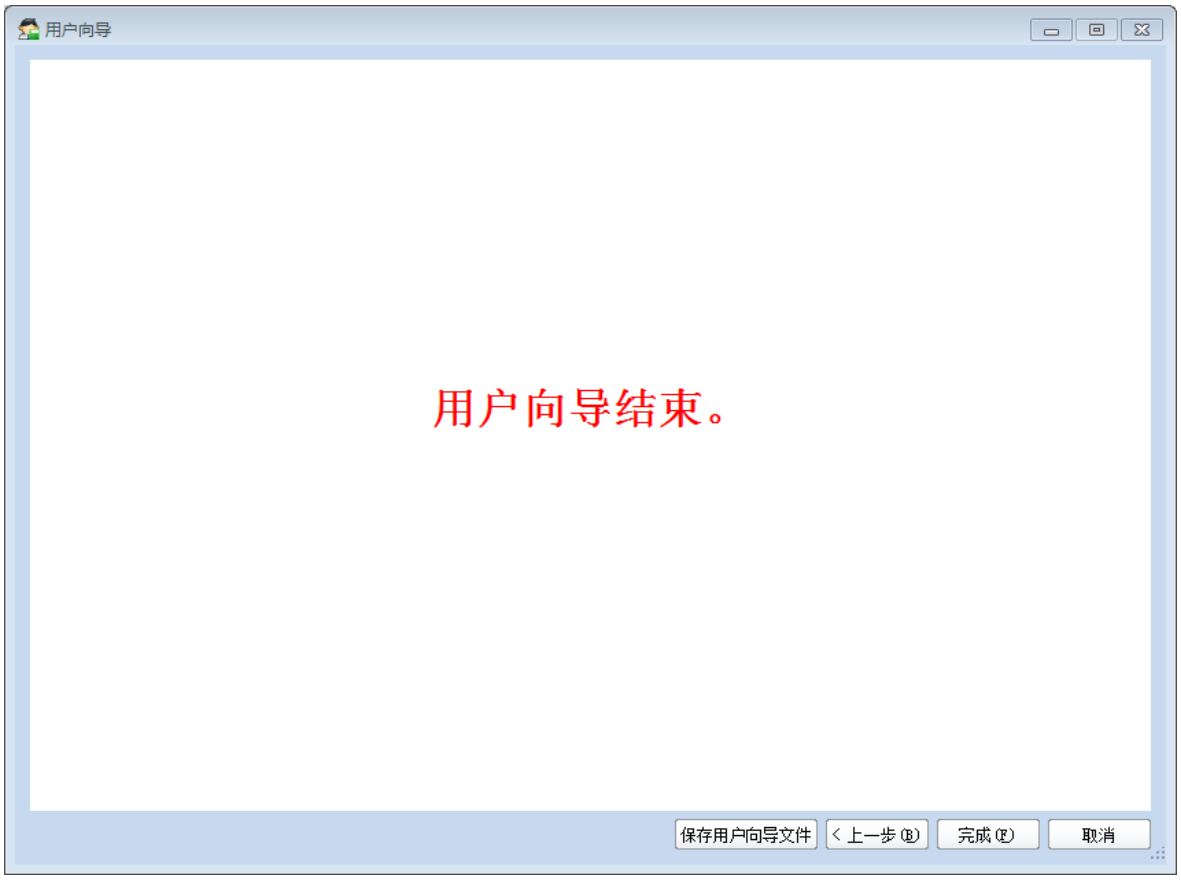


单击“偏好设置”后的界面。

- 偏好设置用于设置用户在使用程序的过程中经常遇到的各种设置（数据处理方法的参数设置除外）。

保存用户向导文件 | 〈上一步 | 下一步〉 | 取消


用户向导
5，参数设置。
文件
主页(O)
图形（L）
预处理(T)
特征选择(V)
建模(M)
预测（P）
窗口（W）
新建
打开
保存
从文件夹导入数据
从数据库导入数据
从文件导入数据
文件
偏好设置
参数设置
设置
新建批
修改批
应用批
批处理
新建报表
修改报表
报表
用户向导
主菜单“主页”栏中的“参数设置”。
●参数设置帮助用户同时设置程序中所有数据处理方法的默认参数，用户在使用这些方法时，程序将自动调用最新的参数，以减少修改次数。
● 所有参数设置可保存在文件中，亦可一次性载入使用。
单击“参数设置”后的界面。
保存用户向导文件
< 上一步 (B)
下一步 (N) >
取消


用户向导
用户向导结束。
保存用户向导文件
< 上一步 (B)
完成 (F)
取消

15.6. 关于我们

关于大连达硕信息技术有限公司和本软件的具体信息。

操作步骤:

步骤 1: 点击**帮助** -> **关于我们**，弹出如下对话框:

步骤 2: 点击公司网址链接，可在浏览器中打开。点击**确定**，则关闭本对话窗口。

第十六章　应用案例

本章暂略，相关内容将在其他文件中详细介绍。

第十七章　结果的准确性

本软件中数据处理方法所得到的结果，均与目前受到广泛国际认可的软件进行了严格的第三方验证与审查，达到甚至优于这些软件所得到的结果，包括 Unscrambler、SIMCA-P、SPSS，以及部分由 Matlab 所得到的结果。

具体对照结果暂略，有兴趣的用户可自行比较。

第十七章　数据处理方法概述

本章详尽介绍几个重要数据处理方法的数学原则。

17.1. 概述

本软件涉及的数据处理方法非常多，其中部分方法已在上述章节中有所说明，用户在阅读本章内容时，可对照这部分的内容，以求更好地理解这些方法和结果，以及图形解释。

17.2. 符号说明

本章中所所涉及数学符号的意义，如下表所示。

序号	符号	说明
1	x	标量。
2	$\mathbf{x} = [x_1\ x_2\ \dots\ x_n]$	多个标量所构成的向量。
3	**X**	多个长度相等的向量所构成的矩阵。
4	$\mathbf{x}^T$, $\mathbf{X}^T$	向量或矩阵的转置。
5	$x_{ij} = \boldsymbol{X}(i,j)$	矩阵中的任一元素(本处指第 i 行，第 j 列元素)。
6	**X**(i,·), **X**(·,j)	矩阵的某一行或某一列(本处指某 i 行或某 j 列)。
7	*y*, **y**, **Y**	响应值，本软件所指因变量(分别为标量，向量或矩阵)。
8	β	理论回归系数。
9	*a*, *A*	主成分分析中的主成分数(分别总数和当前数)。

10	b_0, $\mathbf{b}_0$	截距(分别为标量或向量)。
11	b, **b**	回归系数(分别为标量或向量)。
12	C	模型中心化符号，若为中心化，则该值为 0，否则为 1。
13	d	自由度。
14	E_a	提取 a 个主成分后模型的 X 残差。
15	f, F_a	提取 a 个主成分后模型的 Y 残差。
16	h, **H**	样本杠杆值(分别为标量或矩阵)。
17	i, I	样本数。
18	j, J	Y 变量数。
19	k, K	X 变量数。
20	N	矩阵元素数。
21	**p**, **P**	X 载荷(分别为向量或矩阵)。
22	**q**, **Q**	Y 载荷(分别为向量或矩阵)。
23	**t**, **T**	得分(分别为向量或矩阵)。
24	**u**, **U**	初始得分(分别为向量或矩阵)。
25	**w**, **W**	X 载荷权重(分别为向量或矩阵)。
26	$\bar{x}$, $\bar{\mathbf{x}}$	向量 **x** 或矩阵 **X** 的均值。

27	$\bar{y}$, $\bar{\mathbf{y}}$	响应向量 **y** 或矩阵 **Y** 的均值。
28	$\hat{y}$	响应值 **y** 的预测值。

17.3. PCA 法

PCA 法基本知识在此不再赘述，用户可参考。PCA 分析的一般模型可表述为如下式子，即将原始数据分解为得分(**T**)和载荷(**P**)矩阵二个部分。

$$\mathbf{X} = \mathbf{T}\,\mathbf{P}^{T} + \mathbf{E}$$

数据经过中心化处理后，可得：

$$\mathbf{X} = \mathbf{1}\cdot\mathbf{x}_{mean} + \mathbf{T}_{(A)}\mathbf{P}_{(A)}^{T} + \mathbf{E}_{(A)}$$

或以另一种形式表达为：

$$x_{ik} = \mathrm{x}_{mean,k} + \sum_{a=1}^{A} t_{ia}p_{ka} + e_{ik(A)}$$

如前所述，本软件中采用 NIPALS 的方法完成 PCA 分解，通过多次迭代回归计算的方式每次获得一个主成分，即，将数据矩阵 **X** 与得分$\hat{\mathbf{t}}$进行回归分析，获得载荷$\hat{\mathbf{P}}$,再基于所获得$\hat{\mathbf{P}}$，计算新的$\hat{\mathbf{t}}$，如此往复计算，完成数据分解，并获得最终的得分和载荷矩阵。

具体计算过程如下：

首选从变量方向对数据进行标度化处理，以确保数据变量间可比较的噪声水平。本软件亦在使用 PCA 方法时，在界面上可快速实现方法的选择；然后对数据进行中心化处理，这通常是 PCA 分析前必须使用的步骤。

在此基础上，先预设初始得分向量$\hat{\mathbf{t}}_a$为数据中的一列(通常为平方最大的一列)，然后循环如下几步，直到程序收敛为止。

$$\hat{\mathbf{p}}'_a = (\hat{\mathbf{t}}'_a\hat{\mathbf{t}}_a)^{-1}\,\hat{\mathbf{t}}'_a\mathbf{X}_{\text{a-1}}$$

$$\hat{\mathbf{p}}_a = \hat{\mathbf{p}}_a \, (\hat{\mathbf{p}}'_a \hat{\mathbf{p}}_a)^{-0.5} \text{ (归一化处理}\hat{\mathbf{p}}_a\text{)}$$

$$\hat{\mathbf{t}}_a = \mathbf{X}_{a-1} \, \hat{\mathbf{p}}_a \, (\hat{\mathbf{p}}'_a \, \hat{\mathbf{p}}_a)^{-1}$$

$$\hat{\tau}_a = \hat{\mathbf{t}}'_a \, \hat{\mathbf{t}}_a \text{ (计算迭代终止条件)}$$

$$\mathbf{X}_a = \mathbf{X}_{a-1} - \hat{\mathbf{t}}_a \hat{\mathbf{p}}'_a \text{ (开始新的主成分计算)}$$

迭代终止条件：

$$\hat{\tau}_{a,\text{后一次计算}} - \hat{\tau}_{a,\text{前一次计算}} \leq 0.0001 \hat{\tau}_{a,\text{后一次计算}}$$

或者

$$\left| \left| t_{\text{前一次计算}} - t_{\text{后一次计算}} \right| \right| < 1.\text{e} - 12$$

并设定最大迭代次数为 50-100 次。

17.4. PCR 法

如前所述，PCR 分析是在 PCA 分析的基础上建立起来的，其模型如下二式所示。

$$\mathbf{X} = \mathbf{T}\,\mathbf{P}^{\mathrm{T}} + \mathbf{E}$$

$$\mathbf{y} = \mathbf{T}\,\mathbf{b} + \mathbf{f}$$

即先对数据矩阵进行 PCA 分解，然后将所得到的得分矩阵 **T** 与 **y** 进行回归分析，回归系数 $\hat{\mathbf{b}}$可由下式计算得到。

$$\hat{\mathbf{b}} = \hat{\mathbf{P}}\,\hat{\mathbf{q}}$$

其中，$\hat{\mathbf{P}}$为$\mathbf{X}$载荷(含 A 个主成分)，即$\hat{\mathbf{P}} = \{\hat{\mathbf{p}}_{ka}, k = 1,2,\ldots,K; a = 1,2,\ldots,A\}$；而$\hat{\mathbf{q}}$则为 Y 载荷，即$\hat{\mathbf{q}} = (\hat{\mathrm{q}}_1, \hat{\mathrm{q}}_2, \ldots, \hat{\mathrm{q}}_{\mathrm{A}})'$，由下式计算得到。

$$\hat{\mathbf{q}} = (diag(1/\hat{\tau}_a))^{-1} \, \hat{\mathbf{T}}'\mathbf{y}$$

由此可得：

$$\hat{\mathbf{b}} = \hat{\mathbf{P}} \, (diag(1/\hat{\tau}_a)) \hat{\mathbf{P}}'\mathbf{X}'\mathbf{y}$$

17.5. PLS1 法

PLS 法包括二个类型，一种是处理仅含一个因变量 **y** 的简化方法，称为 PLS1。而另一方法则是可同时处理含有多个因变量 **y** 的方法，称为 PLS2。

如下二式为 PLS 方法的基本模型。

$$\mathbf{X} = \mathbf{T}\,\mathbf{P}^{T} + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{T}\,\mathbf{Q}^{T} + \mathbf{F}$$

同样地，在进行 PLS1 分析前，先对变量进行标度化处理，并对数据 **X** 和 **y** 进行中心化处理，即：

$$\mathbf{X}_0 = \mathbf{X} - \mathbf{1}\bar{\mathbf{x}}'$$

$$\boldsymbol{y}_0 = \mathbf{y} - \mathbf{1}\bar{\mathbf{y}}$$

然后重复如下几个步骤：

$$\mathbf{X}_{a-1} = \mathbf{y}_{a-1}\mathbf{w}_a' + \mathbf{E}$$

$$\hat{\mathbf{w}}_a = c\mathbf{X}_{a-1}'\mathbf{y}_{a-1} \text{ (归一化处理}\hat{\mathbf{w}}_a\text{)}$$

其中，$c = (\mathbf{y}_{a-1}'\mathbf{X}_{a-1}\mathbf{y}_{a-1})^{-0.5}$

$$\mathbf{X}_{a-1} = \mathbf{t}_a\hat{\mathbf{w}}_a' + \mathbf{E} \text{ (估计}\mathbf{t}_a\text{)}$$

$$\mathbf{y}_{a-1} = \hat{\mathbf{t}}_a\mathbf{q}_a + \mathbf{f} \text{ (估计}\mathbf{q}_a\text{)}$$

即，$\hat{\mathbf{q}}_a = \mathbf{y}_{a-1}'\hat{\mathbf{t}}_a/\hat{\mathbf{t}}_a'\hat{\mathbf{t}}_a$

$$\hat{\mathbf{E}} = \mathbf{X}_{a-1} - \hat{\mathbf{t}}_a\hat{\mathbf{p}}_a'$$

$$\hat{\mathbf{f}} = \mathbf{y}_{a-1} - \hat{\mathbf{t}}_a\hat{\mathbf{q}}_a$$

以上二式可得各种模型的统计量，并以残差重置数据，开始新的计算，如下三式所示。

$$\mathbf{X}_a = \hat{\mathbf{E}}$$

$$\mathbf{y}_a = \hat{\mathbf{f}}$$

a = a + 1

获得因子数 *A*，即可确定模型，如下二式所示。

$$\hat{\mathbf{b}} = \widehat{\mathbf{W}}(\widehat{\mathbf{P}}'\widehat{\mathbf{W}})^{-1}\hat{\mathbf{q}}$$

$$\hat{\mathbf{b}}_0 = \bar{\mathbf{y}} - \bar{\mathbf{x}}'\hat{\mathbf{b}}$$

新样本的完整预测，则执行如下计算：

首先与构建校正模型时相同，标度化处理数据矩阵，并做中心化处理。

$$\mathbf{x}'_{i,0} = \mathbf{x}'_i - \bar{\mathbf{x}}'$$

注意，此处的$\bar{\mathbf{x}}'$为校正集的值。

对每个潜变量的计算，则执行如下几步。

$$\mathbf{t}'_{i,a} = \mathbf{x}'_{i,a-1}\widehat{\mathbf{w}}_a$$

其中$\widehat{\mathbf{w}}_a = c\mathbf{X}'_{a-1}\mathbf{y}_{a-1}$（归一化处理$\widehat{\mathbf{w}}_a$）

$$c = (\mathbf{y}'_{a-1}\mathbf{X}_{a-1}\mathbf{y}_{a-1})^{-0.5}$$

以下式计算新的残差。

$$\mathbf{x}_{i,a} = \mathbf{x}_{i,a-1} - \hat{\mathbf{t}}_{i,a}\hat{\mathbf{p}}'_a$$

若需要继续获得新的潜变量，则重复上述步骤，否则获得以下式计算预测结果。

$$\hat{y}_i = \bar{y} + \sum_{a=1}^{A}\hat{t}_{i,a}\hat{q}_a$$

简单地，亦可以下式计算预测结果。

$$\hat{y}_i = \hat{b}_0 + x'_i\hat{b}$$

需要注意的是，式中 P 和 Q 无须做标准化处理，而 T 和 W 则需要归一化到 1。

17.6. PLS2 法

先预设一个$\hat{\mathbf{u}}_a$值以计算载荷权重$\mathbf{w}'_a$，并做归一化处理。如下三式所示。

$$\mathbf{X}_{a-1} = \hat{\mathbf{u}}_a \mathbf{w}'_a + \mathbf{E}$$

$$\hat{\mathbf{w}}_a = c\mathbf{X}'_{a-1}\hat{\mathbf{u}}_a$$

$$c = (\hat{\mathbf{u}}'_a \mathbf{X}_{a-1}\mathbf{X}'_{a-1}\hat{\mathbf{u}}_a)^{-0.5}$$

在 PLS1 法估计得到$\mathbf{q}_a$后，并在残差前，先判断是否收敛(最后迭代步骤无有意义的变化即可)。若未达收敛条件，则以下式估计$\mathbf{u}_a$，

$$\mathbf{Y}_{a-1} = \mathbf{u}_a \hat{\mathbf{q}}'_a + \mathbf{F}$$

亦即：

$$\hat{\mathbf{u}}_a = \mathbf{Y}_{a-1}\hat{\mathbf{q}}_a(\hat{\mathbf{q}}'_a\hat{\mathbf{q}}_a)^{-1}$$

以此再计算循环新的$\mathbf{w}_a$值，开始新的迭代计算。若达到收敛条件，则以下式计算得到模型。

$$\hat{\mathbf{B}} = \hat{\mathbf{W}}(\hat{\mathbf{P}}'\hat{\mathbf{W}})^{-1}\hat{\mathbf{Q}}'$$

$$\mathbf{b}'_0 = \bar{\mathbf{y}}' - \bar{\mathbf{x}}'\hat{\mathbf{B}}$$

17.7. O-PLS 法

如前所述，预处理后的数据(包括转置、标度化和中心化等)，可依如下步骤完成 OPLS 计算，构建模型并获得结果。

本处以含有多个响应变量的数据 **Y** 为例，具体的计算步骤如下。

1) 对数据 **Y** 中的每一列，估计其对应的 **w** 值，并构建矩阵 **W**，即，

$\mathbf{w}^T = \mathbf{y}^T\mathbf{X} / (\mathbf{y}^T\mathbf{y})$

$\mathbf{W} = [\mathbf{W}\ \mathbf{w}]$

2) 对矩阵 **W** 进行 PCA 分解，即，

$\mathbf{W} = \mathbf{T}_W\mathbf{P}_W^T$

基于如下 3) - 8)步完成传统 PLS 计算(数据矩阵 **X** ~ 响应变量矩阵 **Y**)。

3) 以某一响应变量 **Y** 值，设置初始 **u**，并重复如下 4) - 8)步直至程序收敛。

4) $\mathbf{w}^T = \mathbf{u}^T\mathbf{X} / (\mathbf{u}^T\mathbf{u})$

5) $\mathbf{w} = \mathbf{w} / ||\mathbf{w}||$

6) $\mathbf{t} = \mathbf{X}\mathbf{w} / (\mathbf{w}^T\mathbf{w})$

7) $\mathbf{c}^T = \mathbf{t}^T\mathbf{Y} / (\mathbf{t}^T\mathbf{t})$

8) $\mathbf{u} = \mathbf{Y}\mathbf{c} / (\mathbf{c}^T\mathbf{c})$

收敛条件：$||\mathbf{u}_{\text{后一次计算}} - \mathbf{u}_{\text{前一次计算}}|| / ||\mathbf{u}_{\text{前一次计算}}|| < 10^{-10}$

若未达致终止条件，则返回到第 4 步，否则进入下一步计算。

9) $\mathbf{p}^T = \mathbf{t}^T * \mathbf{X} / (\mathbf{t}^T\mathbf{t})$

若无需计算正交潜变量，则进入步骤 18，否则进入下一步计算。

正交化处理 **P** 与 $\mathbf{T}_w$ 中的各列变量，并设置 $\mathbf{w}_{ortho} = \mathbf{p}$。

10) $\mathbf{p} = \mathbf{p} - (\mathbf{t}_w^T\mathbf{P}/(\mathbf{t}_w^T\mathbf{t}_W))\mathbf{t}_w;$

11) $\mathbf{w}_{ortho} = \mathbf{w}_{ortho} / ||\mathbf{w}_{ortho}||$

12) $\mathbf{t}_{ortho} = \mathbf{X}\mathbf{w}_{ortho} / (\mathbf{w}_{ortho}^T\mathbf{w}_{ortho})$

13) $\mathbf{P}_{ortho}^T = \mathbf{t}_{ortho}^T\mathbf{X} / (\mathbf{t}_{ortho}^T\mathbf{t}^{ortho})$

14) $\mathbf{E}_{O\text{-}PLS} = \mathbf{X} - \mathbf{t}_{ortho}\mathbf{p}_{ortho}^T$

以下三式累积计算结果，返回到步骤 3，并设置 $\mathbf{X} = \mathbf{E}_{O\text{-}PLS}$。

15) $\mathbf{T}_{ortho} = [\mathbf{T}_{ortho}\ \mathbf{t}_{ortho}]$

16) $\mathbf{P}_{ortho} = [\mathbf{P}_{ortho}\ \mathbf{p}_{ortho}]$

17) $\mathbf{W}_{ortho} = [\mathbf{W}_{ortho}\ \mathbf{w}_{ortho}]$

以如下步骤计算下一 PLS 成分的变化，从数据矩阵 X 和 Y 中去除当前 PLS 成分，并保存计算结果。

18) $\mathbf{E} = \mathbf{X} - \mathbf{t}\mathbf{p}^T$

19) $\mathbf{F} = \mathbf{Y} - \mathbf{t}\mathbf{c}^T$

20) $\mathbf{T}_{PLS} = [\mathbf{T}_{PLS}\ \mathbf{t}]$

21) $\mathbf{W}_{PLS} = [\mathbf{W}_{PLS}\ \mathbf{w}]$

22) $\mathbf{P}_{PLS} = [\mathbf{P}_{PLS}\ \mathbf{p}]$

若需计算下一 PLS 成分，则返回到步骤 1，并设置 **X** = **E**，**Y** = **F**；否则进入下一步，即，

23) $\mathbf{X}_{ortho} = \mathbf{T}_{ortho}\mathbf{P}_{ortho}^T$

基于 PCA 分析分解 $\mathbf{X}_{ortho}$。

24) $\mathbf{X}_{ortho} = \mathbf{T}_{PCA\text{-}ortho}\mathbf{P}_{PCA\text{-}ortho}^T + \mathbf{E}_{PCA\text{-}ortho}$

将被去除的 PLS 成分加入到 $\mathbf{E}_{O\text{-}PLS}$，即，

$\mathbf{E}_{O\text{-}PLS} = \mathbf{E}_{O\text{-}PLS} + \mathbf{T}_{PLS}\mathbf{P}_{PLS}^T$

使用校正模型所得的 $\mathbf{W}_{ortho}$，$\mathbf{P}_{ortho}$，$\mathbf{W}_{PLS}$ 以及 $\mathbf{P}_{PLS}$ 值，校正新样本，即对每个成分的计算重复如下 25 - 27，以及 28 - 30 步骤。

对 OPLS 成分的计算，

25) $t_{ortho,new} = \mathbf{x}_{new}^T\mathbf{w}_{ortho} / (\mathbf{w}_{ortho}^T\mathbf{w}_{ortho})$

26) $\mathbf{t}_{ortho,new}^T = [\mathbf{t}_{ortho,new}^T\ t_{ortho,new}]$

27) $\mathbf{e}_{O\text{-}PLS,new}^T = \mathbf{x}_{new}^T - t_{ortho,new}\mathbf{P}_{ortho}^T$

设置 $\mathbf{x}_{new}^T = \mathbf{e}_{PLS,new}^T$，返回至第 25 步开始新的计算，否则进入步骤 31。

对 PLS 成分的计算，

28) $t_{new}^T = \mathbf{x}_{new}^T \mathbf{w}_{PLS} / (\mathbf{w}_{PLS}^T \mathbf{w}_{PLS})$

29) $\mathbf{t}_{PLS,new}^T = [\mathbf{t}_{PLS,new}^T \; t_{PLS,new}]$;

30) $\mathbf{e}_{PLS,new}^T = \mathbf{x}_{new}^T - t_{PLS,new}\mathbf{P}_{PLS}^T$

$\mathbf{x}_{new}^T = \mathbf{e}_{PLS,new}^T$，返回至第 25 步开始新的计算，否则进入下一步。

31) $\mathbf{x}_{ortho,new}^T = \mathbf{t}_{ortho,new}^T \mathbf{P}_{ortho}^T$

32) $\mathbf{t}_{PCA\text{-}ortho,new}^T = \mathbf{x}_{ortho,new}^T \mathbf{P}_{PCA\text{-}ortho}$

估计第 24 步中的 PCA 新得分，$\mathbf{x}_{ortho,new}^T = \mathbf{t}_{PCA\text{-}ortho,new}^T \mathbf{P}_{PCA\text{-}ortho}^T + \mathbf{e}_{PCA\text{-}ortho,new}$

若仅移除 PCA 分解后得到的正交成分，则以下式将 $\mathbf{e}_{PCA\text{-}ortho,new}$ 加入到 $\mathbf{e}_{O\text{-}PLS,new}$ 中。

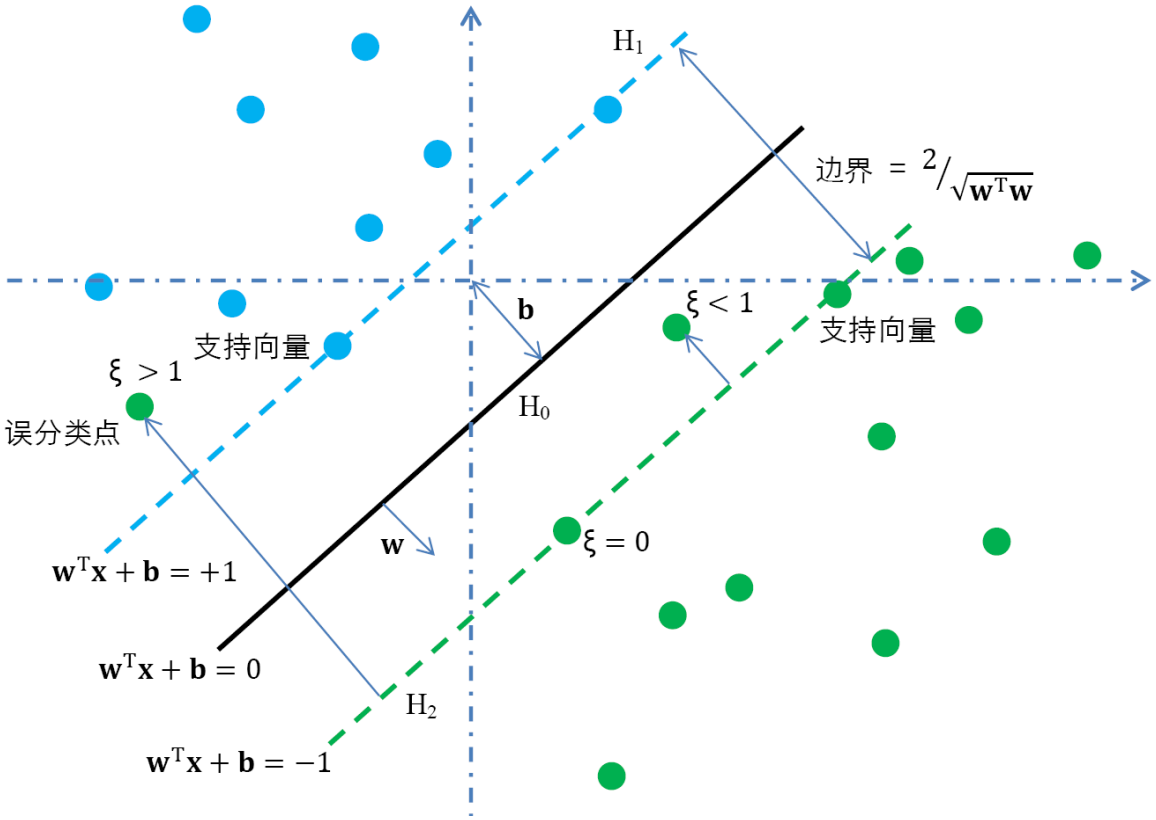
$\mathbf{e}_{O\text{-}PLS,new}^T = \mathbf{e}_{O\text{-}PLS,new}^T + \mathbf{t}_{PCA,new}^T \mathbf{P}_{PCA}^T$

33) 通过下式将 PLS 分析得到的成分加入 $\mathbf{e}_{O\text{-}PLS,new}^T$ 中。

$\mathbf{e}_{O\text{-}PLS,new}^T = \mathbf{e}_{O\text{-}PLS,new}^T + \mathbf{t}_{PLS,new}^T \mathbf{P}_{PLS}^T$

17.8. SVM 法

SVM 法的基本原理可由下图表示。从图中可知，该法的核心是在于确定函数 $g(x) = \mathbf{w}^T\mathbf{x} + \mathbf{b}$ 的各个要素，包括如何获得 **w** 和 **b**，如何求得最优的 g(x)，以及最优的标准是什么等。



其中，最优标准以类间的距离，即分类间隔决定，显然间隔越远则类与类越不易混淆。若设定数据点为 $D_i = <x_i, y_i>$，定义几何间隔$\xi_i = \left(1/||\mathbf{w}||\right) g(x_i)$，则图中虚线间的间隔便是几何间隔，$H_1$ 为<**w**, **x**> + **b** = 1，H2 为<**w**, **x**> + **b** = -1。几何间隔与样本误分次数间的关系为误分次数 <= $(2r/\xi)^2$，其中 ξ 为样本集合到分类面的间隔，r 则等于 max $||x_i||$，i = 1, 2, ..., n。基于此，上述 SVC 分类问题转换为求最大 ξ 值的问题。

实因 **w** 为超平面法向量，因而 **w** 实际上仅由在 H_1 平面上的样本点来决定，这些点即为支持向量，支撑样本类别的分界线。从而亦将问题变为这样的目标函数：$\max\left(\frac{1}{||\mathbf{w}||}\right)$，亦即 $\min\left(\frac{1}{2}||\mathbf{w}||^2\right)$，约束条件是：$y_i(\mathbf{w}^T x_i + b) \geq 1, i = 1, 2, \ldots, n.$，其中 *n* 为样本数。

很显然，当||**w**|| = 0 时便达到目标函数最小值，其几何意义为上图中 H_1 和 H_2 间的距离为无限大。

将上述条件最优化问题实因凸二次规划问题，必定存在全局最优解，并可将其转换为拉格朗日极值问题，引入拉格朗日乘子法，得下式。

$$L(\mathrm{w},\mathrm{b},\alpha)=\frac{1}{2}||\mathbf{w}||^2-\sum_{i=1}^{n}\alpha_i(y_i(\mathbf{w}^{\mathrm{T}}x_i+\mathbf{b})-1)$$

求极值可得如下二式，

$$\frac{\partial L}{\partial \mathrm{w}}=0\ \gg \mathrm{w}=\sum_{i=1}^{n}\alpha_i y_i x_i$$

$$\frac{\partial L}{\partial \mathrm{b}}=0\ \gg \sum_{i=1}^{n}\alpha_i y_i=0$$

将上二式带入$L(\mathrm{w},\mathrm{b},\alpha)$可得：

$$L(\mathrm{w},\mathrm{b},\alpha)=\frac{1}{2}\sum_{i,\mathrm{j}=1}^{n}\alpha_i\alpha_j y_i y_j x_i^{\mathrm{T}}x_j-\sum_{i,\mathrm{j}=1}^{n}\alpha_i\alpha_j y_i y_j x_i^{\mathrm{T}}x_j-b\sum_{i=1}^{n}\alpha_i y_i+\sum_{i=1}^{n}\alpha_i$$

$$=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,\mathrm{j}=1}^{n}\alpha_i\alpha_j y_i y_j x_i^{\mathrm{T}}x_j$$

从而将问题更进一步转换为对偶问题，得到如下形式。

$$\max_{\alpha}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,\mathrm{j}=1}^{n}\alpha_i\alpha_j y_i y_j x_i^{\mathrm{T}}x_j$$

$$\mathrm{s.t.},\alpha_i\ \geq 0, i=1,2,\ldots,n;\ \sum_{i=1}^{n}\alpha_i y_i=0$$

最后基于二次优化算法可得拉格朗日乘子，并可获得如下模型，即：

$$f(\mathrm{x})=sgn((\sum_{i=1}^{n}y_i\alpha_i x_i)^{\mathrm{T}}\mathbf{x}+\mathbf{b})$$

其中，$b = -(1/2)[\max_{yi=-1}(\mathbf{w}^{\mathrm{T}}\mathbf{x}) + \min_{yi=+1}(\mathbf{w}^{\mathrm{T}}\mathbf{x})]$。

对于未知样本的预测，更体现该法的精妙之处，即仅需计算其与训练数据集的内积即可，如下式所示。

$$f(\mathrm{x})=\left(\sum_{i=1}^{n}y_i\alpha_i x_i\right)^{\mathrm{T}}\mathbf{x}+\mathbf{b}=\sum_{i=1}^{n}\alpha_i y_i<\mathrm{x}_i,\mathrm{x}>+\mathrm{b}$$

而对于线性不可分的问题，该法的求解方法是基于核函数将低维数据映射到更高维空间，以便将新数据进行线性划分。

本软件提供如下表所述的 4 种核函数。

序号	核函数名称	公式	参数
1	线性核函数	$K(x_i, x_j) = x_i^{\mathrm{T}} x_j$	无可设参数。
2	多项式核函数	$K(x_i, x_j) = (\gamma x_i^{\mathrm{T}} x_j + r)^d, \gamma > 0$	*d*-多项式最高项次数，γ-通常使用类别数的倒数，r-通常为 0。
3	RBF 核函数	$K(x_i, x_j) = \exp(-\gamma \lvert\lvert x_i - x_j \rvert\rvert^2), \gamma > 0$	γ-通常使用类别数的倒数。
4	Sigmoid 核函数	$K(x_i, x_j) = \tanh(\gamma x_i^{\mathrm{T}} x_j + r)$	γ-通常使用类别数的倒数，r-通常为 0。

本软件提供 *C*-SVC 和 *nu*-SVC 二种方法，它们没有本质区别，参数 *C* 在[0 ∞]之间，基于数据中的噪声信息选择，或采用交互检验或网格搜索的方法优化；而 *nu* 则在[0 1]之间，为分类错误样本所占比例的上界，支持向量所占比列的下界，增大 *nu* 值意味着增大分类边界，允许更大的错误率。

第十八章　安装说明

本章暂略，相关内容将在其他文件中详细介绍。

第十九章　参考文献

本软件的完成涉及大量参考文献，将其中部分文献列举如下。

1. Croux, C.; Filzmoser, P.; Oliveira, M. R.: Algorithms for Projection - Pursuit Robust Principal Component Analysis. Chemometrics and Intelligent Laboratory Systems 2007, 87, 218-225.

2. Nielsen, N.-P. V.; Carstensen, J. M.; Smedsgaarda, J.: Aligning of Single and Multiple Wavelength Chromatographic Profiles for Chemometric Data Analysis using Correlation Optimised Warping. Journal of Chromatography A 1998, 805, 17-35.

3. Abbaspour, A.; Mirzajani, R.: Application of Spectral Beta-Correction Method and Partial Least Squares for Simultaneous Determination of V(IV) and V(V) in Surfactant Media. Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy 2006, 64, 646-652.

4. Osten, D. W.; Kowalski, B. R.: Background Detection and Correction in Multicomponent Analysis. Analytical Chemistry 1985, 57, 908-917.

5. Hu, Y. G.; Jiang, T.; Shen, A. G.; Li, W.; Wang, X. P.; Hu, J. M.: A Background Elimination Method Based On Wavelet Transform for Raman Spectra. Chemometrics and Intelligent Laboratory Systems 2007, 85, 94-101.

6. Zhang, Z. M.; Chen, S.; Liang, Y. Z.: Baseline Correction using Adaptive Iteratively Reweighted Penalized Least Squares. Analyst 2010, 135, 1138-1146.

7. Chen, T.; Martin, E.: Bayesian Linear Regression and Variable Selection for Spectroscopic Calibration. Analytica Chimica Acta 2009, 631, 13-21.

8. Nounou, M. N.; Bakshi, B. R.; Goel, P. K.; Shen, X. T.: Bayesian Principal Component Analysis. Journal of Chemometrics 2002, 16, 576-595.

9. Rajalahti, T.; Arneberg, R.; Berven, F. S.; Myhr, K.-M.; Ulvik, R. J.; Kvalheim, O. M.:

Biomarker Discovery in Mass Spectral Profiles by Means of Selectivity Ratio Plot. Chemometrics and Intelligent Laboratory Systems 2009, 95, 35-48.

10. Laxalde, J.; Ruckebusch, C.; Devos, O.; Caillol, N.; Wahl, F.; Duponchel, L.: Characterisation of Heavy Oils using Near-Infrared Spectroscopy: Optimisation of Pre-Processing Methods and Variable Selection. Analytica Chimica Acta 2011, 705, 227-234.

11. Wold, S.; Sjostrom, M.: Chemometrics, Present and Future Success. Chemometrics and Intelligent Laboratory Systems 1998, 44, 3-14.

12. Karstang, T. V.; Kvalheim, O. M.: Comparison between 3 Techniques for Background Correction in Quantitative-Analysis. Chemometrics and Intelligent Laboratory Systems 1991, 12, 147-154.

13. andersson, M.: A Comparison of Nine Pls1 Algorithms. Journal of Chemometrics 2009, 23, 518-529.

14. Jiang, W.; Zhang, Z.-M.; Yun, Y.; Zhan, D.-J.; Zheng, Y.-B.; Liang, Y.-Z.; Yang, Z. Y.; Yu, L.: Comparisons of Five Algorithms for Chromatogram Alignment. Chromatographia 2013, 76, 1067-1078.

15. Verduandres, J.; Massart, D. L.; Menardo, C.; Sterna, C.: Correction of Non-Linearities in Spectroscopic Multivariate Calibration by using Transformed Original Variables and PLS Regression. Analytica Chimica Acta 1997, 349, 271-282.

16. Tomasi, G.; Van Den Berg, F.; andersson, C.: Correlation Optimized Warping and Dynamic Time Warping As Preprocessing Methods for Chromatographic Data. Journal of Chemometrics 2004, 18, 231-241.

17. Diana, G.; Tommasi, C.: Cross-Validation Methods in Principal Component Analysis: A Comparison. Statistical Methods & Applications 2002, 11, 71-82.

18. Bouveresse, E.; Rutan, S. C.; Vanderheyden, Y.; Penninckx, W.; Massart, D. L.: Detection,

Interpretation and Correction of Changes in the Instrumental Response of Near-Infrared Monochromator Instruments Over Time. Analytica Chimica Acta 1997, 348, 283-301.

19. Luypaert, J.; Heuerding, S.; Massart, D. L.; Vander Heyden, Y.: Direct Orthogonal Signal Correction As Data Pretreatment in the Classification of Clinical Lots of Creams From Near Infrared Spectroscopy Data. Analytica Chimica Acta 2007, 582, 181-189.

20. Kemsley, E. K.: Discriminant Analysis of High-Dimensional Data: A Comparison of Principal Components Analysis and Partial Least Squares Data Reduction Methods. Chemometrics and Intelligent Laboratory Systems 1996, 33, 47-61.

21. Centner, V.; Massart, D.-L.; Noord, O. E. D.; Jong, S. D.; Vandeginste, B. M.; Sterna, C.: Elimination of Uninformative Variables for Multivariate Calibration. Analytical Chemistry 1996, 68, 3851-3858.

22. Goutis, C.: A Fast Method To Compute Orthogonal Loadings Partial Least Squares. Journal of Chemometrics 1997, 11, 33-38.

23. Vogt, F.; Tacke, M.: Fast Principal Component Analysis of Large Data Sets. Chemometrics and Intelligent Laboratory Systems 2001, 59, 1-18.

24. Guo, Q.; Wu, W.; Massart, D. L.; Boucon, C.; De Jong, S.: Feature Selection in Principal Component Analysis of Analytical Data. Chemometrics and Intelligent Laboratory Systems 2002, 61, 123-132.

25. Gourvenec, S.; Capron, X.; Massart, D. L.: Genetic Algorithms (Ga) Applied To the Orthogonal Projection Approach (Opa) for Variable Selection. Analytica Chimica Acta 2004, 519, 11-21.

26. Broadhurst, D.; Goodacre, R.; Jones, A.; Rowland, J. J.; Kell, D. B.: Genetic Algorithms As A Method for Variable Selection in Multiple Linear Regression and Partial Least Squares Regression, with Applications To Pyrolysis Mass Spectrometry. Analytica Chimica Acta 1997, 348, 71-86.

27. Phatak, A.; Dejong, S.: the Geometry of Partial Least Squares. Journal of Chemometrics 1997, 11, 311-338.

28. Debraekeleer, K.; Sanchez, F. C.; Hailey, P. A.; Sharp, D. C. A.; Pettman, A. J.; Massart, D. L.: Influence and Correction of Temperature Perturbations On Nir Spectra During the Monitoring of A Polymorph Conversion Process Prior To Self-Modelling Mixture Analysis. Journal of Pharmaceutical and Biomedical Analysis 1998, 17, 141-152.

29. Serneels, S.; Croux, C.; Van Espen, P. J.: Influence Properties of Partial Least Squares Regression. Chemometrics and Intelligent Laboratory Systems 2004, 71, 13-20.

30. Gugliotta, L. M.; Vega, J. R.; Meira, G. R.: Instrumental Broadening Correction in Size Exclusion Chromatography - Comparison of Several Deconvolution Techniques. Journal of Liquid Chromatography 1990, 13, 1671-1708.

31. Kvalheim, O. M.: Interpretation of Partial Least Squares Regression Models by Means of Target Projection and Selectivity Ratio Plots. Journal of Chemometrics 2010, 24, 496-504.

32. Wentzell, P. D.; andrews, D. T.; Hamilton, D. C.; Faber, K.; Kowalski, B. R.: Maximum Likelihood Principal Component Analysis. Journal of Chemometrics 1997, 11, 339-366.

33. Osborne, S. D.; Jordan, R. B.; Kunnemeyer, R.: Method of Wavelength Selection for Partial Least Squares. Analyst 1997, 122, 1531-1537.

34. Grung, B.; Manne, R.: Missing Values in Principal Component Analysis. Chemometrics and Intelligent Laboratory Systems 1998, 42, 125-139.

35. Li, B. B.; Morris, J.; Martin, E. B.: Model Selection for Partial Least Squares Regression. Chemometrics and Intelligent Laboratory Systems 2002, 64, 79-89.

36. Li, H.-D.; Liang, Y.-Z.; Xu, Q.-S.; Cao, D.-S.: Model-Population Analysis and Its Applications in Chemical and Biological Modeling. Trends in Analytical Chemistry 2012, 38, 154-162.

37. Kasemsumran, S.; Du, Y. P.; Li, B. Y.; Maruo, K.; Ozaki, Y.: Moving Window Cross Validation:

A New Cross Validation Method for the Selection of A Rational Number of Components in A Partial Least Squares Calibration Model. Analyst 2006, 131, 529-537.

38. Gemperline, P. J.; Cho, J. H.; Archer, B.: Multivariate Background Correction for Hyphenated Chromatography Detectors. Journal of Chemometrics 1999, 13, 153-164.

39. Jiang, J. H.; Wang, J. H.; Chu, X.; Yu, R. Q.: Neural Network Learning To Non-Linear Principal Component Analysis. Analytica Chimica Acta 1996, 336, 209-222.

40. Hassel, P. A.; Martin, E. B.; Morris, J.: Non-Linear Partial Least Squares. Estimation of the Weight Vector. Journal of Chemometrics 2002, 16, 419-426.

41. Ter Braak, C. J. F.; De Jong, S.: the Objective Function of Partial Least Squares Regression. Journal of Chemometrics 1998, 12, 41-54.

42. Leng, C. L.; Wang, H. S.: On General Adaptive Sparse Principal Component Analysis. Journal of Computational and Graphical Statistics 2009, 18, 201-215.

43. Gil, J. A.; Romera, R.: On Robust Partial Least Squares (Pls) Methods. Journal of Chemometrics 1998, 12, 365-378.

44. Wold, S.; Anttia, H.; Lindgrenb, F.; Hman, J. O.: Orthogonal Signal Correction of Near-Infrared Spectra. Chemometrics and Intelligent Laboratory Systems 1998, 44, 175-185.

45. Henseler, J.: Partial Least Squares. Wiley Encyclopedia of Management 2015.

46. Krishnan, A.; Williams, L. J.; Mcintosh, A. R.; Abdi, H.: Partial Least Squares (Pls) Methods for Neuroimaging: A Tutorial and Review. Neuroimage 2011, 56, 455-475.

47. Hinkle, J.; Rayens, W.: Partial Least Squares and Compositional Data: Problems and Alternatives. Chemometrics and Intelligent Laboratory Systems 1995, 30, 159-172.

48. Barker, M.; Rayens, W.: Partial Least Squares for Discrimination. Journal of Chemometrics 2003, 17, 166-173.

49. Xu, Q. S.; De Jong, S.; Lewi, P.; Massart, D. L.: Partial Least Squares Regression with Curds and Whey. Chemometrics and Intelligent Laboratory Systems 2004, 71, 21-31.

50. Boulesteix, A. L.; Strimmer, K.: Partial Least Squares: A Versatile Tool for the Analysis of High-Dimensional Genomic Data. Briefings in Bioinformatics 2007, 8, 32-44.

51. Wold, S.; Sjostrom, M.; Eriksson, L.: Pls-Regression: A Basic Tool of Chemometrics. Chemometrics and Intelligent Laboratory Systems 2001, 58, 109-130.

52. Liland, K. H.; Indahl, U. G.: Powered Partial Least Squares Discriminant Analysis. Journal of Chemometrics 2009, 23, 7-18.

53. Wold, S.; Esbensen, K.; Geladi, P.: Principal Component Analysis. Chemometrics and Intelligent Laboratory Systems 1987, 2, 37-52.

54. Guo, J. A.; James, G.; Levina, E.; Michailidis, G.; Zhu, J.: Principal Component Analysis with Sparse Fused Loadings. Journal of Computational and Graphical Statistics 2010, 19, 930-946.

55. Chau, F.-T.; Chan, H.-Y.; Cheung, C.-Y.; Xu, C.-J.; Liang, Y.; Kvalheim, O. M.: Recipe for Uncovering the Bioactive Components in Herbal Medicine. Analytical Chemistry 2009, 81, 7217-7225.

56. Hoy, M.; Steen, K.; Martens, H.: Review of Partial Least Squares Regression Prediction Error in Unscrambler. Chemometrics and Intelligent Laboratory Systems 1998, 44, 123-133.

57. Kruger, U.; Zhou, Y.; Wang, X.; Rooney, D.; Thompson, J.: Robust Partial Least Squares Regression - Part Iii, Outlier Analysis and Application Studies. Journal of Chemometrics 2008, 22, 323-334.

58. Gonzalez, J.; Pena, D.; Romera, R.: A Robust Partial Least Squares Regression Method with Applications. Journal of Chemometrics 2009, 23, 78-90.

59. Kruger, U.; Zhou, Y.; Wang, X.; Rooney, D.; Thompson, J.: Robust Partial Least Squares Regression: Part I, Algorithmic Developments. Journal of Chemometrics 2008, 22, 1-13.

60. Kruger, U.; Zhou, Y.; Wang, X.; Rooney, D.; Thompson, A.: Robust Partial Least Squares Regression: Part Ii, New Algorithm and Benchmark Studies. Journal of Chemometrics 2008, 22, 14-22.

61. Vanden Branden, K.; Hubert, M.: Robustness Properties of A Robust Partial Least Squares Regression Method. Analytica Chimica Acta 2004, 515, 229-241.

62. Barros, A. S.; Pinto, R.; Delgadillo, I.; Rutledge, D. N.: Segmented Principal Component Transform-Partial Least Squares Regression. Chemometrics and Intelligent Laboratory Systems 2007, 89, 59-68.

63. Wold, S.; Trygg, J.; Berglund, A.; Antti, H.: Some Recent Developments in PLS Modeling. Chemometrics and Intelligent Laboratory Systems 2001, 58, 131-50.

64. Helland, I. S.: Some Theoretical Aspects of Partial Least Squares Regression. Chemometrics and Intelligent Laboratory Systems 2001, 58, 97-107.

65. Du, Y. P.; Liang, Y. Z.; Jiang, J. H.; Berry, R. J.; Ozaki, Y.: Spectral Regions Selection To Improve Prediction Ability of PLS Models by Changeable Size Moving Window Partial Least Squares and Searching Combination Moving Window Partial Least Squares. Analytica Chimica Acta 2004, 501, 183-191.

66. Ghasemi, J.; Niazi, A.: Spectrophotometric Simultaneous Determination of Nitroaniline Isomers by Orthogonal Signal Correction-Partial Least Squares. Talanta 2005, 65, 1168-1173.

67. Niazi, A.; Yazdanipour, A.: Spectrophotometric Simultaneous Determination of Nitrophenol Isomers by Orthogonal Signal Correction and Partial Least Squares. Journal of Hazardous Materials 2007, 146, 421-427.

68. Benoudjit, N.; Francois, D.; Meurens, M.; Verleysen, M.: Spectrophotometric Variable Selection by Mutual Information. Chemometrics and Intelligent Laboratory Systems 2004, 74, 243-251.

69. Sun, J.: Statistical Analysis of Nir Data: Data Pretreatment. Journal of Chemometrics 1997, 11, 525-532.

70. Chen, Z. P.; Li, L. M.; Yu, R. Q.; Littlejohn, D.; Nordon, A.; Morris, J.; Dann, A. S.; Jeffkins, P. A.; Richardson, M. D.; Stimpson, S. L.: Systematic Prediction Error Correction: A Novel Strategy for Maintaining the Predictive Abilities of Multivariate Calibration Models. Analyst 2011, 136, 98-106.

71. Pravdova, V.; Boucon, C.; De Jong, S.; Walczak, B.; Massart, D. L.: Three-Way Principal Component Analysis Applied To Food Analysis: An Example. Analytica Chimica Acta 2002, 462, 133-148.

72. Alexandridis, A.; Patrinos, P.; Sarimveis, H.; Tsekouras, G.: A Two-Stage Evolutionary Algorithm for Variable Selection in the Development of Rbf Neural Network Models. Chemometrics and Intelligent Laboratory Systems 2005, 75, 149-162.

73. Afanador, N. L.; Tran, T. N.; Buydens, L. M.: Use of the Bootstrap and Permutation Methods for A More Robust Variable Importance in the Projection Metric for Partial Least Squares Regression. Analytica Chimica Acta 2013, 768, 49-56.

74. Rao, R.; Lakshminarayanan, S.: Variable Interaction Network Based Variable Selection for Multivariate Calibration. Analytica Chimica Acta 2007, 599, 24-35.

75. Gong, F.; Wang, B. T.; Liang, Y. Z.; Chau, F. T.; Fung, Y. S.: Variable Selection for Discriminating Herbal Medicines with Chromatographic Fingerprints. Analytica Chimica Acta 2006, 572, 265-71.

76. Despagne, F.; Massart, D. L.: Variable Selection for Neural Networks in Multivariate Calibration. Chemometrics and Intelligent Laboratory Systems 1998, 40, 145-163.

77. Balabin, R. M.; Smirnov, S. V.: Variable Selection in Near-Infrared Spectroscopy: Benchmarking of Feature Selection Methods On Biodiesel Data. Analytica Chimica Acta 2011, 692, 63-72.

78. Alsberg, B. K.; Woodward, A. M.; Winson, M. K.; Rowland, J. J.; Kell, D. B.: Variable Selection in Wavelet Regression Models. Analytica Chimica Acta 1998, 368, 29-44.

79. Wiklund, S.; Johansson, E.; Sjostrom, L.; Mellerowicz, E. J.; Edlund, U.; Shockcor, J. P.; Gottfries, J.; Moritz, T.; Trygg, J.: Visualization of Gc/Tof-Ms-Based Metabolomics Data for Identification of Biochemically Interesting Compounds using Opls Class Models. Analytical Chemistry 2008, 80, 115-122.

80. Jiang, J.-H.; Berry, R. J.; Siesler, H. W.; Ozaki, Y.: Wavelength Interval Selection in Multicomponent Spectral Analysis by Moving Window Partial Least-Squares Regression with Applications To Mid-Infrared and Near-Infrared Spectroscopic Data. Analytical Chemistry 2002, 74, 3555-3565.

81. Liang, Y. Z.; Kvalheim, O. M.; Rahmani, A.; Brereton, R. G.: A 2-Way Procedure for Background Correction of Chromatographic Spectroscopic Data by Congruence Analysis and Least-Squares Fit of the Zero-Component Regions - Comparison with Double-Centering. Chemometrics and Intelligent Laboratory Systems 1993, 18, 265-279.

82. Vigneau, E.; Bertrand, D.; Qannari, E. M.: Application Of Latent Root Regression For Calibration In Near-Infrared Spectroscopy. Comparison With Principal Component Regression And Partial Least Squares. *Chemometrics And Intelligent Laboratory Systems* **1996**, *35*, 231-238.

83. Belousov, A. I.; Verzakov, S. A.; Von Frese, J.: Applicational Aspects Of Support Vector Machines. *Journal Of Chemometrics* **2002**, *16*, 482-489.

84. Yang, Z. R.; Chou, K. C.: Bio-Support Vector Machines For Computational Proteomics. *Bioinformatics* **2004**, *20*, 735-U549.

85. Wehrens, R.; Vanderlinden, W. E.: Bootstrapping Principal Component Regression Models. *Journal Of Chemometrics* **1997**, *11*, 157-171.

86. Pierna, J. A. F.; Baeten, V.; Renier, A. M.; Cogdill, R. P.; Dardenne, P.: Combination Of Support Vector Machines (Svm) And Near-Infrared (Nir) Imaging Spectroscopy For the Detection Of Meat And Bone Meal (Mbm) In Compound Feeds. *Journal Of Chemometrics* **2004**, *18*, 341-349.

87. Amendolia, S. R.; Cossu, G.; Ganadu, M. L.; Golosio, B.; Masala, G. L.; Mura, G. M.: A Comparative Study Of K-Nearest Neighbour, Support Vector Machine And Multi-Layer Perceptron For Thalassemia Screening. *Chemometrics And Intelligent Laboratory Systems* **2003**, *69*, 13-20.

88. Hsu, C. C.; Lin, J.; Chao, C. K.: Comparison Of Multiple Linear Regression And Artificial Neural Network In Developing the Objective Functions Of the Orthopaedic Screws. *Computer Methods And Programs In Biomedicine* **2011**, *104*, 341-348.

89. Galtier, O.; Abbas, O.; Le Dreau, Y.; Rebufa, C.; Kister, J.; Artaud, J.; Dupuy, N.: Comparison Of Pls1-Da, Pls2-Da And Simca For Classification By Origin Of Crude Petroleum Oils By Mir And Virgin Olive Oils By Nir For Different Spectral Regions. *Vibrational Spectroscopy* **2011**, *55*, 132-140.

90. Li, Y. K.; Shao, X. G.; Cai, W. S.: A Consensus Least Squares Support Vector Regression (Ls-Svr) For Analysis Of Near-Infrared Spectra Of Plant Samples. *Talanta* **2007**, *72*, 217-222.

91. Kim, H. C.; Pang, S.; Je, H. M.; Kim, D.; Bang, S. Y.: Constructing Support Vector Machine Ensemble. *Pattern Recognition* **2003**, *36*, 2757-2767.

92. Wollenweber, M.; Polster, J.: Evaluation Of Data Produced By Optode Arrays Under Flow Injection Analysis (Fia) Conditions Using A Partial Least Squares Method (Pls2). *Journal Of Biochemical And Biophysical Methods* **1999**, *41*, 1-11.

93. Esbensen, K. H.; Waskaas, M.; Matveyev, I. H.; Wolden, K. E.; Lode, J. G.; Lied, T. T.; Halstensen, M.: Feasibility Of Emf-Induced Pipeline Wall Friction Reduction By Pls2 Intercalibration Of Acoustic Chemometrics And Reference Laser Velocimetry. *Journal Of Chemometrics* **2001**, *15*, 241-+.

94. Raudys, S.: How Good Are Support Vector Machines? *Neural Networks* **2000**, *13*, 17-19.

95. Hua, Z. S.; Zhang, B.: A Hybrid Support Vector Machines And Logistic Regression Approach For Forecasting Intermittent Demand Of Spare Parts. *Applied Mathematics And Computation* **2006**, *181*, 1035-1048.

96. Capparuccia, R.; De Leone, R.; Marchitto, E.: Integrating Support Vector Machines And Neural Networks. *Neural Networks* **2007**, *20*, 590-597.

97. Wu, Q.; Law, R.: An Intelligent Forecasting Model Based On Robust Wavelet V-Support Vector Machine. *Expert Systems With Applications* **2011**, *38*, 4851-4859.

98. Ding, C. H. Q.; Dubchak, I.: Multi-Class Protein Fold Recognition Using Support Vector Machines And Neural Networks. *Bioinformatics* **2001**, *17*, 349-358.

99. Komura, D.; Nakamura, H.; Tsutsumi, S.; Aburatani, H.; Ihara, S.: Multidimensional Support Vector Machines For Visualization Of Gene Expression Data. *Bioinformatics* **2005**, *21*, 439-444.

100. Sousa, S. I. V.; Martins, F. G.; Alvim-Ferraz, M. C. M.; Pereira, M. C.: Multiple Linear Regression And Artificial Neural Networks Based On Principal Components To Predict Ozone Concentrations. *Environmental Modelling & Software* **2007**, *22*, 97-103.

101. Jandel, M.: A Neural Support Vector Machine. *Neural Networks* **2010**, *23*, 607-613.

102. Hable, R.; Christmann, A.: On Qualitative Robustness Of Support Vector Machines. *Journal Of Multivariate Analysis* **2011**, *102*, 993-1007.

103. Trygg, J.; Wold, S.: Orthogonal Projections To Latent Structures (O-Pls). *Journal Of*

Chemometrics **2002**, *16*, 119-128.

104. Niu, L. F.: Parallel Algorithm For Training Multiclass Proximal Support Vector Machines. *Applied Mathematics And Computation* **2011**, *217*, 5328-5337.

105. Molfetta, F. A.; Bruni, A. T.; Rosselli, R. P.; Da Silva, A. B. E.: A Partial Least Squares And Principal Component Regression Study Of Quinone Compounds With Trypanocidal Activity. *Structural Chemistry* **2007**, *18*, 49-57.

106. Mage, I.; Menichelli, E.; Naes, T.: Preference Mapping By Po-Pls: Separating Common And Unique Information In Several Data Blocks. *Food Quality And Preference* **2012**, *24*, 8-16.

107. Guillen-Casla, V.; Rosales-Conrado, N.; Leon-Gonzalez, M. E.; Perez-Arribas, L. V.; Polo-Diez, L. M.: Principal Component Analysis (Pca) And Multiple Linear Regression (Mlr) Statistical Tools To Evaluate the Effect Of E-Beam Irradiation On Ready-To-Eat Food. *Journal Of Food Composition And Analysis* **2011**, *24*, 456-464.

108. Pant, S. D.; Schenkel, F. S.; Verschoor, C. P.; You, Q. M.; Kelton, D. F.; Moore, S. S.; Karrow, N. A.: A Principal Component Regression Based Genome Wide Analysis Approach Reveals the Presence Of A Novel Qtl On Bta7 For Map Resistance In Holstein Cattle. *Genomics* **2010**, *95*, 176-182.

109. Vigneau, E.; Devaux, M. F.; Qannari, E. M.; Robert, P.: Principal Component Regression, Ridge Regression And Ridge Principal Component Regression In Spectroscopy Calibration. *Journal Of Chemometrics* **1997**, *11*, 239-249.

110. Giacomino, A.; Abollino, O.; Malandrino, M.; Mentasti, E.: the Role Of Chemometrics In Single And Sequential Extraction Assays: A Review. Part Ii. Cluster Analysis, Multiple Linear Regression, Mixture Resolution, Experimental Design And Other Techniques. *Analytica Chimica Acta* **2011**, *688*, 122-139.

111. Fung, G.; Mangasarian, O. L.: Semi-Supervised Support Vector Machines For Unlabeled Data Classification. *Optimization Methods & Software* **2001**, *15*, 29-44.

112. Pedro, A. M. K.; Ferreira, M. M. C.: Simultaneously Calibrating Solids, Sugars And Acidity Of Tomato Products Using Pls2 And Nir Spectroscopy. *Analytica Chimica Acta* **2007**, *595*, 221-227.

113. Rodriguez, D. M.; Aguilar, F. J. A.; Wrobel, K.: Spectrophotometric Assay For Copper And Iron In Transformer Oil Using Partial Least Squares Regression (Pls2). *Ieee Transactions On Dielectrics And Electrical Insulation* **2006**, *13*, 1272-1277.

114. Hua, S. J.; Sun, Z. R.: Support Vector Machine Approach For Protein Subcellular Localization Prediction. *Bioinformatics* **2001**, *17*, 721-728.

115. Pavlidis, P.; Wapinski, I.; Noble, W. S.: Support Vector Machine Classification On the Web. *Bioinformatics* **2004**, *20*, 586-587.

116. Mangasarian, O. L.: Support Vector Machine Classification Via Parameterless Robust Linear Programming. *Optimization Methods & Software* **2005**, *20*, 115-125.

117. Shamim, M. T. A.; Anwaruddin, M.; Nagarajaram, H. A.: Support Vector Machine-Based Classification Of Protein Folds Using the Structural Properties Of Amino Acid Residues And Amino Acid Residue Pairs. *Bioinformatics* **2007**, *23*, 3320-3327.

118. Devos, O.; Ruckebusch, C.; Durand, A.; Duponchel, L.; Huvenne, J. P.: Support Vector Machines (Svm) In Near Infrared (Nir) Spectroscopy: Focus On Parameters Optimization And Model Interpretation. *Chemometrics And Intelligent Laboratory Systems* **2009**, *96*, 27-33.

119. Li, H. D.; Liang, Y. Z.; Xu, Q. S.: Support Vector Machines And Its Applications In Chemistry. *Chemometrics And Intelligent Laboratory Systems* **2009**, *95*, 188-198.

120. Ben-Hur, A.; Ong, C. S.; Sonnenburg, S.; Schokopf, B.; Ratsch, G.: Support Vector Machines And Kernels For Computational Biology. *Plos Computational Biology* **2008**, *4*.

121. Lin, Y.; Lee, Y.; Wahba, G.: Support Vector Machines For Classification In Nonstandard

Situations. *Machine Learning* **2002**, *46*, 191-202.

122. Bellotti, T.; Crook, J.: Support Vector Machines For Credit Scoring And Discovery Of Significant Features. *Expert Systems With Applications* **2009**, *36*, 3302-3308.

123. Peng, N. B.; Zhang, Y. X.; Zhao, Y. H.: Support Vector Machines For Quasar Selection. In *Software And Cyberinfrastructure For Astronomy*; Radziwill, N. M., Bridger, A., Eds., 2010; Vol. 7740.

124. Singh, K. P.; Basant, N.; Gupta, S.: Support Vector Machines In Water Quality Management. *Analytica Chimica Acta* **2011**, *703*, 152-162.

125. Zavaljevski, N.; Stevens, F. J.; Reifman, J.: Support Vector Machines With Selective Kernel Scaling For Protein Classification And Identification Of Key Amino Acid Positions. *Bioinformatics* **2002**, *18*, 689-696.

126. Hernandez, N.; Talavera, I.; Biscay, R. J.; Porro, D.; Ferreira, M. M. C.: Support Vector Regression For Functional Data In Multivariate Calibration Problems. *Analytica Chimica Acta* **2009**, *642*, 110-116.

127. Decoste, D.; Scholkopf, B.: Training Invariant Support Vector Machines. *Machine Learning* **2002**, *46*, 161-190.

128. Peng, X. J.: Tsvr: An Efficient Twin Support Vector Machine For Regression. *Neural Networks* **2010**, *23*, 365-372.

129. Vaira, S.; Mantovani, V. E.; Robles, J. C.; Sanchis, J. C.; Goicoechea, H. C.: Use Of Chemometrics: Principal Component Analysis (Pca) And Principal Component Regression (Pcr) For the Authentication Of Orange Juice. *Analytical Letters* **1999**, *32*, 3131-3141.

130. Thiessen, U.; Van Brakel, R.; De Weijer, A. P.; Melssen, W. J.; Buydens, L. M. C.: Using Support Vector Machines For Time Series Prediction. *Chemometrics And Intelligent Laboratory Systems* **2003**, *69*, 35-49.

131. Steel, S. J.; Uys, D. W.: A Variable Selection Proposal For Multiple Linear Regression Analysis. *Journal Of Statistical Computation And Simulation* **2011**, *81*, 2095-2105.

132. Usdun, B.; Melssen, W. J.; Buydens, L. M. C.: Visualisation And Interpretation Of Support Vector Regression Models. *Analytica Chimica Acta* **2007**, *595*, 299-309.