附件：化学计量学指导原则草案公示稿（第一次）

化学计量学指导原则

化学计量学是一门化学分支的交叉学科，它应用数学和统计学方法并借助计算机技术，设计和选择最优的测量方法和实验，通过解释分析数据，以最优的方式获取关于物质系统的有关信息。化学计量学方法具有通用性，已广泛应用于药物研发、药品质量控制、药品打假、中药材产地和属性的识别等领域。本指导原则介绍化学计量学方法在数据处理和分析中的应用，简化了化学计量学方法的原理与算法，着重阐述化学计量学方法的选择和使用；同时提供了化学计量学方法模型全生命周期管理的基本原则及关键内容。

化学计量学基于多变量数据对研究对象进行表征，建立数学模型，通过模型参数理解研究对象的特征，实现复杂信号的分辨、类属的判别以及定量信息的校正等，进而完成表征、鉴别或定量等分析任务。

化学计量学方法与传统数据分析方法的显著区别在于“多变量分析”。 当传统数据分析方法不适用时，可尝试采用化学计量学方法获得解决方案。多变量数据 $\mathbf{X}$ 一般是指同时利用多个样品的多个测量属性或变量的数据，用 $m \times n$ 的数据表或矩阵表示（m 是样品数，n 是每个样品测量数据的变量数），如 $\mathbf{X}$ 是 m 个样品在 n 个波长（或波数）的近红外光谱数据。样品的性质参数（组分的含量、类别等）一般用向量表示，把某一性质按照样品的顺序排列，形成一个向量或 $m \times 1$ 的矩阵 $\mathbf{y}$。化学计量学利用数学和统计学方法对 $\mathbf{X}$ 进行分析，或者建立 $\mathbf{y}$ 与 $\mathbf{X}$ 之间的定量关系，实现定性或定量分析。

化学计量学定性分析是通过多个变量的数学变换和统计分析得到样本的类别和特征，或者建立判别模型进行类别的判断和鉴别。化学计量学定性分析方法分为无监督方法和有监督方法。无监督方法在处理数据时无需样本性质参数 $\mathbf{y}$，通过多变量数据 $\mathbf{X}$ 来衡量样本间的相似性从而对各样本进行类别划分，得到样品的类别信息和每类样本的特征信息。常用的无监督方法为聚类分析方法，如系统聚类法、k-均值算法等，能够识别不同样本之间的共性及差异，可应用于药物发现、质量控制等。有监督方法是利用一组已知样品性质信息的样本，

建立多变量数据 **X** 和样本性质参数 **y** 之间的模型，最后将未知样品的多变量数据代入所建模型实现判别。常用的有监督方法为判别分析方法，如线性判别分析（LDA）、偏最小二乘-判别分析（PLS-DA）、支持向量机（SVM）等，可应用于产品的类别判定，如质量筛查、假药识别等。

化学计量学定量分析采用多元校正方法，同时使用多变量数据 **X** 中的多个自变量建立与性质参数 **y** 之间的数学关系，通过不同自变量以线性或非线性形式的组合实现对未知样品性质参数 y 的预测。化学计量学定量分析方法均为有监督方法，常用的多元校正方法包括多元线性回归（MLR）、偏最小二乘回归（PLSR）、支持向量回归（SVR）、人工神经网络（ANN）、深度学习（DL）等，可应用于日常检验或过程分析技术中特定组分含量的快速预测，如药物活性成分、水分含量等。

无论是化学计量学定性分析方法还是定量分析方法，模型的建立与使用都需要遵照一定的流程和规范，即进行模型全生命周期管理，包括数据质量保证、建模方法、模型评估与验证、日常使用中的模型监控等。

一、化学计量学方法

1 数据预处理技术

原始测量数据往往包含噪声、背景等与样品性质无关的信息。采用数据预处理技术可有效地滤除噪声、扣除背景、校正光谱基线等，消除干扰并增强数据与样品性质的相关性。常用的数据处理技术包括尺度调整、平滑滤噪、背景扣除、散射校正、变量选择等。

尺度调整

包括中心化、标准化、归一化等三种基本方法。在建立模型时，通常采用中心化增加样品光谱之间的差异，从而提高模型的预测能力（灵敏度）；采用标准化处理差异较大甚至具有不同量纲的数据，使自变量之间具有相同的权重；采用归一化消除变量之间的相对大小对后续分析带来的影响，如向量归一化、面积归一化、最大归一化、平均归一化等。尺度调整会损失一定程度的某些信息，如中心化会损失信号强度信息，标准化会损失部分差异信息。

平滑滤噪

平滑是指去除信号中无规律的随机干扰信号或周期性的高频干扰信号。滤噪是指去除与分析物无关或者不随分析物浓度改变而变化的信号。平滑和滤噪通常被联合使用，使用时不做区分。常用的平滑滤噪方法有移动窗口平均（MWA）法、Savitzky-Golay（SG）平滑和小波变换（WT）技术等。这些方法都需要对信号两端的数据点做特殊处理（如插值法），以消除边缘效应带来的计算失真。

背景扣除

背景是指与分析物无直接关系的响应信号。背景的扣除方法一般根据分析的原理、检测器的响应性能、样品的性质等确定。除采用空白实验外，还可根据响应曲线的形状估算信号中的背景成分，或对背景信号进行计算扣除，如在色谱分析中采用多项式拟合估计背景成分，在光谱分析中采用导数计算扣除光谱中的背景信息等。常用的导数计算方法包括直接差分法、傅里叶变换方法、SG 导数和 WT 法等。

散射校正

漫反射和透反射光谱中由样品的颗粒度、厚度、装样量等因素导致的光谱背景畸变，即为散射效应，覆盖整个谱区且与波长相关，对化学计量学模型的

影响较大。一般需对整体光谱进行散射校正。它仅对背景校正，不改变光谱形状。

多元散射校正（MSC）利用散射校正系数对光谱进行校正，适用于消除颗粒大小及分布不均匀产生的散射效应，广泛应用于固体的漫反射光谱和半固体、混悬液、乳浊液的透（反）射光谱。

标准正态变换（SNV）对每一条光谱独立地进行校正，适用于消除固体颗粒大小、表面散射以及光程变化对漫反射光谱的影响，不需要使用平均光谱，计算过程更为简单。

变量选择

通过选择有信息的变量、消除无信息的变量，有效去除干扰，增强模型的稳健性和可解释性，达到精简模型和提高模型质量的目的。光谱分析方法的变量选择分为波长选择和波段选择。

波长选择是将光谱中的每一个波长变量作为基本单位，通过波长变量重要性判据选择相对重要的变量。通常首先采用不考虑波长变量之间的相互作用的变量选择方法，如根据所建立模型的质量评价波长变量的重要性，或利用模型参数判断各单波长对模型的贡献大小。常用方法包括相关系数法、模型系数法、迭代预测加权（IPW）法、变量重要性投影（VIP）方法、竞争自适应加权重采样（CARS）方法等。

当这些方法无法得到理想的结果时，应考虑变量之间的相互作用，可采用以下三类方法：①基于统计学参数的变量选择方法，包括有无信息变量删除（UVE）法、蒙特卡洛-无信息变量删除（MC-UVE）法、随机检验（RT）法、C值法等；②基于变量响应值之间关系的变量选择方法包括正交投影（SPA）算法、互信息（MI）算法等；③基于优化算法的变量选择方法，包括遗传算法（GA）、模拟退火算法（SA）、蚁群优化（CO）、粒子群优化（PSO）算法等，适用于变量排列组合的数目过于庞大的情况。

波段选择是将相邻的多个波长变量作为选择的基本单位，选出建模效果较好的波段组合。分为波段划分和波段优选两个步骤，先将整体光谱划分为几个波段，再优化波段的组合，常用方法有区间偏最小二乘（iPLS）法和移动窗口偏最小二乘回归（MWPLSR）法。

虽然好的变量选择可以精简模型，甚至能提升模型的预测能力，但通常由

于波长变量较多且光谱变量之间具有相关性，变量选择具有一定难度，目前尚没有得到广泛共识的标准方法。不同变量选择方法所选出的波长（波段）可以不同，但模型预测结果的差异不应太大，否则应考虑样本量和样本代表性问题。当样本数足够多代表性足够强时，变量选择对模型预测结果的影响会有所降低。强相关变量连续分布（相邻波长相关性较高）时，波段选择效果较好；而强相关变量分布较为分散（相邻波长相关性较低）时，波长选择效果较好。

2 多元统计方法

多元统计分析是研究多个变量（或多个因素）之间相互依赖关系的一种综合分析方法，它能够在多个对象和多个指标互相关联的情况下分析它们的统计规律。多元统计方法是化学计量学实现定性、判别、分类和定量分析的基础，最常用的主要包括相关分析、多元回归分析和主成分分析。

相关分析

用于研究两个或多个变量间相互变化关系，分为直接相关和间接相关。直接相关反映了变量之间真实的因果关系，间接相关则反映因受到其他因素的共同影响而呈现出对应的变化趋势。利用间接相关通过对趋势的研究获得有效信息更为常见。

应根据研究对象的统计学分布特点选择合适的相关分析方法。如果研究对象的分布服从正态分布，可用协方差或相关系数描述两个变量的线性相关关系，用偏相关系数或复相关系数描述多个变量之间的多元相关关系。如果研究对象的分布不服从正态分布，则应采用非参数方法，如适用于两个变量的 Spearman 秩相关系数或 Kendall 秩相关系数，适用于多元非正态分布的多元 Kendall 非参数方法等。除此之外，还可利用典型相关分析（CCA）研究两组变量间的相关关系。

多元回归分析

是将一个或多个因变量表述为多个自变量的函数，通过函数关系由已知自变量估计或预测因变量的方法，其中研究一个或多个因变量和多个自变量之间的线性关系的多元线性回归（MLR）最为常用。MLR 在制药领域主要应用于两个方面：（1）建立模型来解释多种因素对响应变量影响的重要性或大小，用于实验设计（DoE）筛选或响应分析；（2）建立预测模型来进行定量分析，通常用于过程分析。

MLR 的基本模型是 **y=Xb+e**，其中 **y** 是因变量（向量），由观测值构成，**X** 是自变量（矩阵），表示对应于每个观测值的影响因素，b 是回归系数，**e** 是预测值与观测值之间的残差。因此，MLR 模型描述了一个因变量（观测值）与多个自变量（影响因素）之间的相关关系。当有一组已知的 **y** 和 **X** 时，通过最小二乘等方法可以通过使残差 **e** 最小确定模型的回归系数 b；而模型系数 b 则定量地表达了因变量与自变量之间的相关关系。

MLR 可以扩展到多个因变量，即将一个因变量的 **y** 向量扩展到多个因变量的 **Y** 矩阵（每列表示一个因变量 $\mathbf{y}_i$）。当 MLR 应用于具有多个响应的 DoE 数据时，可利用相同的 **X** 矩阵针对每个 $\mathbf{y}_i$ 变量建立独立的 MLR 模型。计算时，可以采用 $\mathbf{y}_i=\mathbf{Xb}_i+\mathbf{e}$ 独立计算每一个模型，也可以采用 **Y=XB+E** 同时计算多个模型，其中 B 为回归系数矩阵，E 为残差矩阵。一般来说，独立计算模型的结果较好。

在光谱等分析测试数据的定量分析中，用 X 表示光谱响应矩阵，Y 表示样品属性矩阵，当分析信号的基本模型是 **X=YB+E**，即光谱响应矩阵是多组分纯光谱按样品属性的加和，利用此基本模型求得各组份的纯光谱对未知混合物的各组份的浓度进行定量分析，称为 K 矩阵法；而利用 **Y=XB+E** 作为基本模型进行计算时，即直接用样品属性矩阵作为预测目标，由多组分混合物光谱响应矩阵求得回归系数矩阵 B 的方法，则称为 P 矩阵法。

由于药品组分较多，药品质量的影响因素也较多。各组分的相互作用或影响因素之间的相互影响对建模的效果会有不同程度的影响，所以选择合适的变量是关键。例如，当 **X** 中各变量之间存在共线性时，这些变量的线性不具有独立性，则会影响 MLR 模型的稳健性，甚至导致无法建立有效的模型。因此，应尽量选择独立性较强的 X 变量进行建模，以避免变量共线性引起的 **b** 系数不可靠或模型不稳定。MLR 建模对样本数量（**y** 和 **X** 的行数）的要求为独立的样本数必须大于或等于变量的个数，否则无法求解模型的 **b** 系数，即无法得到有意义的模型。

主成分分析

主成分分析（PCA）是用于发现变量之间关系的多变量分析方法，通过数学变换，将由线性相关变量表示的多变量数据转化为少数几个线性无关的新变量（主成分），同时保证尽可能多地保留原变量的数据特征而不丢失信息，达到简化数据结构（数据降维）的目的。

在数学上，主成分分析可理解为一种数学变换或投影方法，通过寻找 x 坐标系到 x’坐标系最小信息损失的变换或投影。如图 1，通过寻找解释最大方差的方向（x’坐标轴）将原数据投影到新的空间，更高效地表达数据之间的相互关系。这个过程产生的正交向量（x’坐标轴）称为载荷，原数据在新空间的投影称为得分。因此，主成分分析本质上是将相同的数据在一个新空间中显示，通过它们的投影来揭示样品之间的关系。

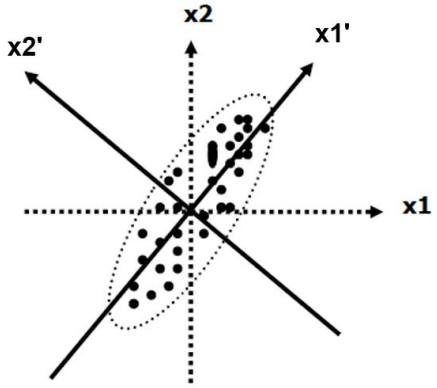

图 1 主成分分析

寻找新空间方向（坐标轴）可通过多种数学方法得以实现，其基本模型是 $\mathbf{X}=\mathbf{T}\mathbf{P}^{\mathrm{T}}+\mathbf{E}$。式中，**X** 是 *m* 行（样品数）*n* 列（变量数）的原始数据矩阵，**T** 和 $\mathbf{P}^{\mathrm{T}}$ 分别是得分和载荷矩阵，当有 *p* 个主成分时，**T** 和 $\mathbf{P}^{\mathrm{T}}$ 分别是 *n* 行 *p* 列和 *p* 行 *m* 列的矩阵，**E** 表示残差，T 表示转置。主成分数（*p*）又称为因子数、潜变量数等。通过主成分分析，原始数据（**X** 矩阵）被转换成一个新的、重新排列的矩阵，新矩阵构成了原始数据的解释部分，残差则是原始数据中无法解释的部分。计算中，新空间的坐标轴由原数据确定，即原数据的线性组合，*p*个方向相互正交且按照所解释方差的大小排列，因此，当原数据的变量数较多时，*p*会远远小于 *n*，达到数据压缩或降维的目的。主成分分析的核心是用一个维数更少的矩阵替代复杂的原始数据矩阵，而信息量仍然与原始数据保持接近，并且从原始数据中提取特征信息，如样本之间的差异。从空间角度来看，一个样本的多个测量值（变量）定义了一个多维空间中的一个点，变量数即空间的维数。因此，相似样品将位于多维空间中的相同区域。通过使用主成分分析，可以在减少维数的同时使相似的样本彼此接近，不同的样本彼此分离。

从统计学的角度，主成分分析的原理是在数据空间中找到描述数据集最大变异的方向，即数据点相距最远的方向，每个方向都是对样本间实际变异贡献

最大的初始变量的线性组合。主成分是相互正交的，通过构建和排序，按包含的信息由多到少对主成分排序。因此应用中会优先对第一主成分进行解释，它包含最大的变异。通常，只有前几个主成分包含有效信息，其他主成分则可能是干扰信息。在实践中，应通过交叉验证或载荷评估等方法建立特定的判断准则来区分噪声和信息，以确定用于分析的主成分数量。残差 E 保存了模型中没有包含的变异，可作为样本或变量与模型拟合程度的度量。主成分分析模型中最终保留的主成分数需综合考虑模型的精简程度、稳健性、拟合度以及性能等。

样本之间的关系可以在一个或几个得分图中得以显示，通常采用前两个主成分的得分，如图 2（A）。载荷是新空间的方向或坐标轴，通过载荷构成分析可以得到载荷与原变量之间的（线性组合）关系，如图 2（B），而原变量之间的关系则可以通过载荷在不同主成分空间的关系得到展现，如图 2（C）。同时，通过得分-载荷双重图，如图 2（D），还可以看出样品与原变量之间的关系，或者描述样品特征的主要变量。

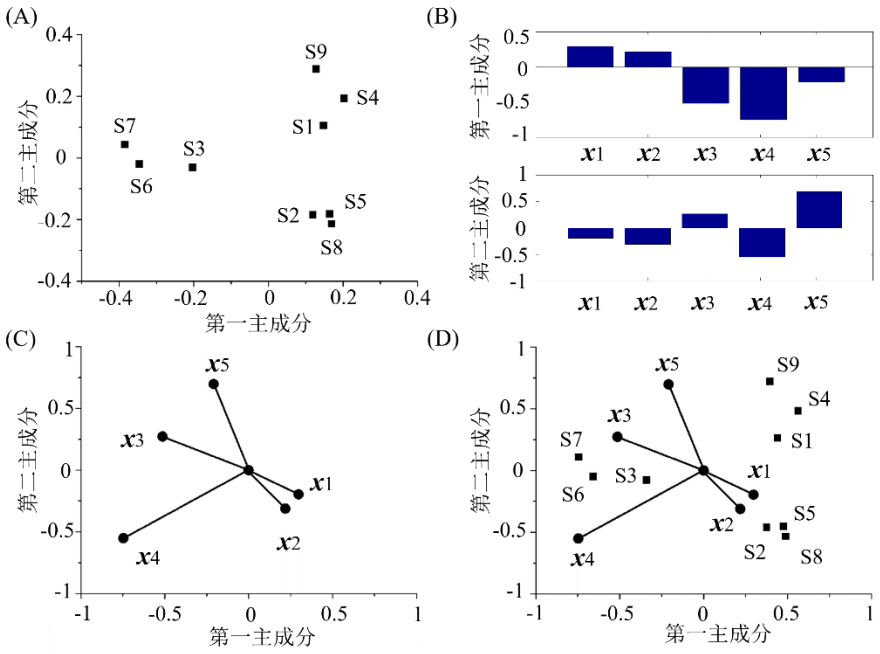

图 2 主成分分析得分图（A）、载荷图构成图（B）、载荷关系图（C）和得分-载荷双重图（第一主成分和第二主成分）（D）

主成分分析是一种无监督的方法，是探索性数据分析的有力工具。通过主成分分析可以显示不同样本的差异，各变量对差异的影响程度，变量之间的相互关系以及样品的特征变量等。值得注意的是，主成分分析捕捉的是数据集中

的主要变化，而相对较小的变化可能无法区分。

3 多元分辨方法

多元分辨（MR）方法是用于处理仪器分析方法（如光谱、色谱、成像技术等）产生的多变量信号的有效工具。多组分体系的测量信号通常可用一个简单的模型来描述，即 $\mathbf{D}=\mathbf{C}\mathbf{S}^{T}$，其中 **C** 为浓度矩阵，$\mathbf{S}^{T}$ 为响应系数矩阵。模型的基本假定是每个组分的测量信号正比于组分的含量，总体测量信号是各组分信号之和。多元分辨技术可从原始测量信号中提取单组分的信息，例如，从 HPLD-DAD 测量数据中提取混合物各组分的光谱和色谱信号，从 GC-MS 测量数据中提取混合物各组分的质谱和色谱信号等。常用的多元分辨方法包括化学因子分析、多元曲线分辨等。

化学因子分析

是通过对数据矩阵进行特征分析、旋转变换等操作，获取混合物体系中各组分的响应信号的 MR 技术。其中，特征分解是所有因子分析法的共通步骤，得到不具有明确的物理或化学意义的抽象解，再根据数据的特点，通过变换得到各组分的浓度、光谱等有实际意义的解。化学因子分析（CFA）在解决多变量问题时具有显著的优点。例如，可处理多因素相互影响的复杂体系，能快速地对大量数据进行处理，可压缩数据，提高数据质量，能研究多种类型的问题。在对原始数据了解甚少甚至对数据的本质一无所知的情况下，仍然可应用化学因子分析方法。更重要的是可获得对测量数据的解释。通过因子分析可对样品或变量进行分类，能够为体系建立完整的有物理意义的模型并以此来预测新的数据点。

化学因子分析已广泛用于色谱、光谱、质谱和化学成像等数据的处理，对待测体系进行定性定量分析。化学因子分析还可用于研究平衡及动力学问题，以及许多其他化学计量学问题，如曲线分辨、数据校正、模式识别等。当分析数据与理论模型（$\mathbf{D}=\mathbf{C}\mathbf{S}^{T}$）有所偏离，如测量数据中存在较严重的基线漂移、较大的噪声干扰、组分信号受实验条件变动较大或者组分之间存在明显相互作用时，CFA 方法的计算结果会存在偏差甚至完全失效。此时，建议对分析测试方案进行调整，或者采用其他的多元分辨方法。

多元曲线分辨

又称为自模型曲线分辨或端元提取，是一种基于测量数据基本模型进行重

叠信号解析的多元分辨技术。多元曲线分辨（MCR）的求解常采用交替最小二乘（ALS）算法，从测量数据 **D** 得到具有化学意义的单个组分的信号 **C** 和 $\mathbf{S}^T$，实现多元分辨，因此也称为 MCR-ALS 方法。与主成分分析寻找最大方差和相互正交的方向相比，多元曲线分辨的目标是发现组分的真实信号 **C** 和 $\mathbf{S}^T$，分别被称为 MCR 得分和 MCR 载荷。

MCR-ALS 适用于具有良好的线性或可转变为线性的测量数据。当被分析组分的测量响应之间具有选择性时，该方法优势在于每个被分析组分只需要 1 个标准样的测量信号作为初始估值。而当测量数据的线性和被分析组分的选择性存在问题时，则每个被分析组分可能需要更多的标准样来校准。

通常，MCR-ALS 的计算结果存在不确定性，且只能得到 MCR 载荷的归一化结果，MCR 得分只是组分之间的相对大小。因此需要使用简单的线性回归方法将 MCR 得分转换为真实的物理量，如药物制剂中有效活性成分和辅料的含量，此时至少一个组分的实际含量应为已知。当两个或两个以上的化学成分变化在某种程度上相互关联时，会出现亏秩现象，例如消耗一种组分而形成另一种组分。在这种情况下，同时分析不同条件下的独立实验数据，或使用两种测量技术的联合测量，通常会获得更好的结果。

4 多元校正方法

多元校正是化学计量学定量分析技术的统称，目的是建立物质浓度（或其他物化性质）与分析数据之间的数学关系，即定量校正模型。对于复杂体系，无法获取与定量目标相关的选择性信号时，必须采用多元校正技术进行定量分析。常见的多元校正方法有多元线性回归、主成分回归、偏最小二乘回归、支持向量机、人工神经网络等，其中主成分回归和偏最小二乘回归是在多元线性回归的基础上发展而来的方法。

主成分回归

实际应用中，通常因变量数较多而达不到 MLR 模型“样品数必须大于变量数”的要求。主成分回归采用 PCA 对自变量矩阵 **X** 进行降维，利用主成分得分建立 MLR 模型，不仅保留了原数据中大部分的信息，也科学地解决了 MLR 模型对样品数的要求。

主成分回归建模的关键是选择合适的主成分数。尽管存在一些方法确定主成分数，主成分数的物理意义也很清楚，但数据中存在的噪声、背景、组分间

的相互作用等干扰因素有时会给主成分数的确定带来困难。因此在实际应用中，一般通过观察残差 **E** 随主成分数的变化，将残差最小（或较小）时对应的主成分数确定为合适的主成分数。主成分数过低或者过高都会影响模型的准确性，主成分数不足时，模型的预测能力不够，主成分数过高会带来模型的过拟合现象。

主成分回归的缺点在于：1）主成分回归只对于自变量矩阵 **X** 进行主成分分析，保留的信息并不一定与因变量 Y 具有较好的相关性；2）主成分回归在主成分数的选择过程中，可能会忽略高阶主成分，导致有用信息的丢失。因此，在光谱分析中，主成分回归通常需比偏最小二乘回归使用更高的主成分数，一般不作为光谱数据定量建模的首选方法。

偏最小二乘回归

是一种潜变量回归分析方法，基于 PCA 从测量数据（自变量 **X**）和预测目标（因变量 **Y**）中分别提取潜变量，并使之尽可能相互正交，从而克服了共线性问题，同时也保留了测量变量中的最大相关信息。偏最小二乘回归（PLSR）不仅对测量数据 **X** 矩阵进行正交分解，而且对因变量 **Y** 矩阵也进行正交分解，并且在分解因变量 **Y** 矩阵的同时也考虑了测量数据 **X** 矩阵的因素，从而加强了 **X** 和 **Y** 矩阵相关性，可以得到最佳的回归效果。通过建立 **X** 矩阵潜变量与 **Y** 矩阵潜变量之间的数学关系构建 PLSR 模型，用来描述 **X** 和 **Y** 矩阵之间的关系。

潜变量也称为因子。使用偏最小二乘回归的一个关键步骤是因子数的确定。因子数选择太小将不能充分解释训练数据集的可变性，而因子数太大将导致过拟合和模型稳健性下降。因此，在模型的校准验证期间应进行因子数的评估。有多种方法可用于模型因子数的考察，最常用的简便方法是观察模型的验证误差随因子数的变化。验证误差是校正集或验证集的预测误差，其中校正集的预测误差又称为校准误差或自验证误差。模型因子数对模型性能的影响如图 3 所示，校准误差随因子数增加呈下降趋势（曲线 A），最佳因子数一般根据预测误差随因子数变化趋势选择：当预测误差随因子数增加呈先下降后上升趋势并出现最小值时（曲线 C），则最小预测误差对应的因子数为最佳因子数；而当预测误差随因子数增加而下降但无最小值时（曲线 B），可选择预测误差不显著降低时对应的因子数作为合适的因子数。

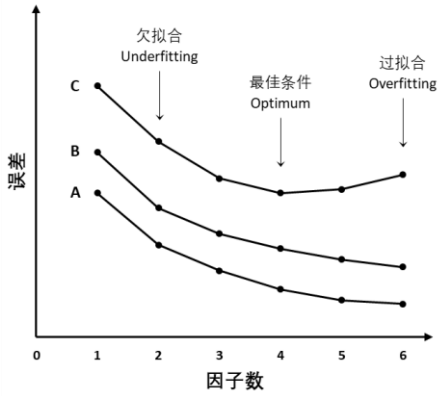

图 3 因子数量对模型性能的影响[A：校准误差随因子数变化曲线；B：预测误差随因子数变化曲线（无最小值）；C：预测误差随因子数变化曲线（有最小值）]

偏最小二乘回归相较于主成分回归，能够更好地描述因变量和测量数据变量的特征。这种方法建立的模型更简单，因子使用更少，还提供了更好的解释可能性和可视化诊断，以优化校准性能。此外，偏最小二乘回归可以消除因变量和测量变量数据中的噪声干扰，是光谱定量分析的主流方法。

5 聚类与判别方法

聚类和判别统称为模式识别，是化学计量学定性分析的常用方法。聚类分析（CA）是将一批样品或变量，按照其性质上亲疏远近的程度进行分类，性质相似的聚成一类，相异的聚为不同的类；判别分析是根据预先设定的分类用校正集数据建立判别函数或模型，待测数据代入判别函数或模型进行类别的判定。

分类的实质是寻找样本之间的差异或相互关系。样本的特征通常用一组能够描述其特征的指标变量表示，按尺度划分为间隔尺度、有序尺度、名义尺度，例如药片的有效成分含量、硬度为间隔尺度，风险的高中低为有序尺度，药物剂型为名义尺度等。分类问题一般通过距离和/或相似性这两个统计量描述指标变量之间的关系，“距离”越小、“相似系数”越大，样本之间越相似。

距离与相似性

用于估算样本之间距离和相似度的方法有很多。不同类型的指标变量在定义距离和相似系数时有很大差异，此处仅涉及间隔尺度的指标变量。常用的描述距离的统计量有欧氏距离、马氏距离，有时还会用到街区距离、明氏距离等。常用的描述相似性的统计量有夹角余弦、相关系数等。由于不同的方法各有侧重，不同方法的计算结果可能存在一定的差异，因此采用合适的方法至关重要。

欧氏距离：是最常用的距离，衡量多维空间中点与点之间的绝对距离。计算公式是：

$$ed_{i,j} = \sqrt{\sum_{k=1}^{m}\left(x_{i,k} - x_{j,k}\right)^2} \quad (1)$$

数据点 i 与数据中心之间的欧氏距离$ed_{i,c}$可用下式计算：

$$ed_{i,c} = \sqrt{(\boldsymbol{x}_i - \mu)^{\mathrm{T}}\left(\boldsymbol{x}_j - \mu\right)} \quad (2)$$

式中 μ 为样本指标参数（变量）的均值，T 表示转置。

马氏距离：也是一种常用的距离公式，可以看作是欧氏距离的一种修正，修正了欧氏距离中各个维度尺度不一致且具有相关性的问题。

单个数据点（$\boldsymbol{x}$）的马氏距离（距中心点的距离）为：

$$DM(\boldsymbol{x}) = \sqrt{(\boldsymbol{x} - \mu)^T \sum^{-1} (\boldsymbol{x} - \mu)} \quad (3)$$

数据点$\boldsymbol{x_i}$和$\boldsymbol{x_j}$之间的马氏距离为：

$$DM(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \sum^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)} \quad (4)$$

式中 Σ 是多维变量的协方差矩阵，μ 为样本均值，T 表示转置。如果协方差矩阵是单位阵，即各个样本间相互独立同分布，马氏距离则变成了欧氏距离。

夹角余弦：是计算数据映射为空间中向量间的余弦值来衡量相似性。此方法在任何维度的向量空间中都适用，计算方法如下：

$$\mathrm{c} = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}} \quad (5)$$

相关系数：常见的相关系数为简单相关系数，反映的是两个变量之间变化趋势的方向及程度。计算公式为：

$$r(x,y) = \frac{cov(x,y)}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - x)(y_i - y)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (6)$$

欧氏距离只适用于表示变量不相关时数据点之间的相似性或差异性。当变量之间存在相关性，则数据空间的实际维数小于变量数，此时可计算马氏距离，但马氏距离要求样本数必须大于变量数。

在主成分空间里计算距离具有更高的效率。由于主成分的正交性，在主成分空间中可以使用少数几个变量表达高维原始数据中的信息，并且可以消除数据中非关键信息的干扰。当采用的主成分累计代表率足够高时，利用主成分得

分计算的距离和用原始变量计算的距离几乎一致。因此，主成分分析并没有改变数据，只是在保持原数据信息的基础上提取了新的潜变量。由于主成分的正交性，马氏距离与采用归一化的得分计算的欧氏距离具有相同的含义，只是在数值上相差一个倍率。

聚类分析

是研究类别关系的一种多变量分析方法，通过样本的分类指标把性质相近或相似的样本归为一类。聚类分析可根据距离或相似性将样本集划分成若干个不同的子集，这些子集称作类（或簇）。这些类（或簇）不是事先给定的，而是根据数据特点进行划分，使得同一簇中的样本彼此相似，不同簇中的样本彼此不同。聚类分析用于解释或验证分析实验数据、优化分析过程。

聚类分析是一种非常实用的探索性数据分析方法，它通过将具有相同特征的样本分组来帮助理解样本的构成，还能从大量原始数据中提取隐藏的信息，以寻找各变量之间的关联、趋势和关系。例如，应用聚类分析能够对中药材进行考察，以区分不同基原与混伪品、不同产地、不同药用部位、不同采收年限与时间等；对药品生产过程进行分析，以评价工艺合理性和产品一致性；对药品与生物体相互作用进行探索，以阐释药品有效性和作用机理。

最常用的聚类分析法是系统聚类分析（HC）法。系统聚类用树形结构来表示样本之间的关系，又称为层次聚类分析。系统聚类算法是一种无监督的学习方法，其算法是将 n 个样本各自看成一类，定义样品之间的距离和类与类之间的距离，将距离最近的两类合并成一个新类，重新计算新类与其他类的距离，再将距离最近的两类合并，如此每次减少一类，直至所有的样本成为一类。其中，类与类之间的聚类的定义通常有最短聚类法、最长距离法、中间距离法、重心法等。

聚类过程标示图称为聚类图，如图 4。系统聚类通过聚类树进行类别的确定，可以由下向上对小的类别进行聚合，即凝聚法，也可以由上向下把大的类别进行分割，即分裂法。

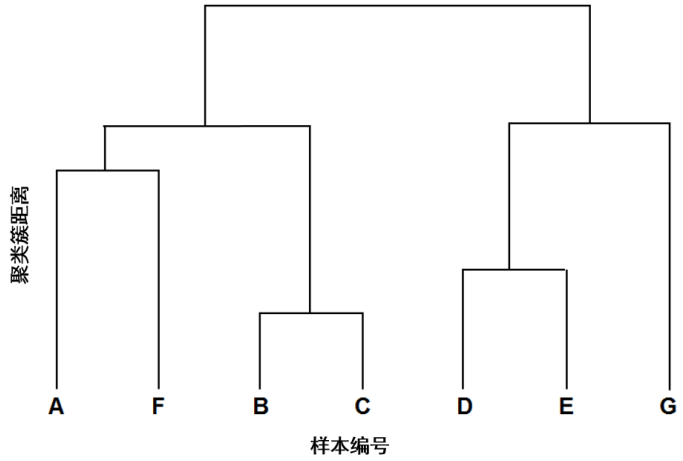

图 4 聚类树示意图

判别分析

是一种有监督的学习方法，是指根据多个因子（观测值或指标参数）对研究对象进行分类的一种多元统计分析方法。与聚类分析不同，判别分析分类的对象要求事先有明确的类别空间。其原理是按照一定的判别准则，使用已知类别研究对象的观测值或指标参数建立一个或多个判别函数或模型，根据未知类别样本的观测值或指标参数对样本的类属进行预测。例如，应用判别分析能够进行药品质量筛查，根据已有真药的特性建立判别准则，鉴别药品的真伪和/或药材的道地性，还可以根据药品与生物体相互作用，进行药品有效性和作用机理研究。

线性判别分析（LDA），又称为 Fisher 判别分析，是一种经典的判别分析算法。LDA 通过寻找降维空间使降维后各样本的“类内差异最小化，类间差异最大化”，与 PCA 具有相似之处，但区别在于寻找的最优投影方向具有更好的样本区分能力而不仅仅是最大化地保留样本信息。应用 LDA 时，可以先将样本点投影到一维空间，若效果不明显，可以考虑增加维度，即投影至二维或更高维空间中。二次判别分析（QDA）法是 LDA 的变体，允许数据的非线性分类。QDA 与 LDA 的区别在于投影面的形状不同。应用 QDA 时需对每个类的协方差矩阵进行估计。LDA 适用于多类问题，但是当类分布不均衡时应谨慎使用，可考虑用 QDA 建立非线性关系。

SIMCA 分类法是一种基于 PCA 的判别方法。其原理为利用先验分类知识对每一种类别建立一个主成分回归模型，用于判断未知样本的类别归属。由于 SIMCA 为基于 PCA 的判别法，其方法验证应遵循 PCA 法的准则。此外，应用

SIMCA 还应考虑不同类的重叠问题。SIMCA 比 PCA 更适用于难区分类别的判别，可用于近红外光谱、质谱等数据的分析，也可以用于色谱和化学成像等数据的分类。

偏最小二乘判别分析（PLS-DA），将多元校正模型中的因变量 **y** 表示类别即可建立预测类别的模型，使用中 **y** 有两种表示方法，一种是采用数字对样本的类别进行编码，另一种方式是用“0”或“1”表示对应样本的类别。预测时，根据模型的预测值进行类别的判定。

TQ 对数法是基于 PCA 和 SIMCA 方法发展起来的一种基于近红外光谱进行药品一致性判别的方法。该方法的原理是：先对某已知类别样品的光谱进行主成分分析，得到各光谱的得分；再由得分计算 Hotelling T^2 值，作为第一统计值 T^2；继而利用主成分分析模型计算每条光谱的残差，作为第二统计值 Q^2；最后对两个统计值进行对数变换后，计算置信椭圆作为模型。判别时，将未知类别样品的光谱投影到模型的主成分空间，通过计算得到该样品的两个对数判别值，再与置信椭圆进行比较，在置信椭圆以内即判别为具有“一致性”，否则，判别为“不一致”。在实际应用中，判别结果可以转化为一个判别参数，即距椭圆边缘的远近，如果某光谱在椭圆以内，判别参数小于 1，在椭圆上等于 1，在椭圆外大于 1，并且越远数值越大。通过赋予判别参数实际意义，如过程控制中的 CQA，则可实现在生产中异常批次判别。

6 模型转移

在模型应用过程中，样品、测试环境或仪器都可能发生改变，如样品温湿度变化、样品形态改变、仪器更换、老化和附件更新等，采集的光谱往往会随之变化，如吸收峰的偏移、展宽、吸收强度的非线性变化和波长的漂移等，导致原有的模型不再适用，预测结果发生偏差。在近红外光谱分析中，为解决此类问题，可通过建立不同条件下样本光谱的函数对光谱进行校正，或通过模型系数或预测结果的校正，使原来的模型具有适用性，此类方法称为模型转移，或模型传递、光谱标准化等，是化学计量学中的一个重要研究领域。

通常，将建立原有模型所使用光谱的测量仪器（或测量条件）称为主机，待转移光谱的测量仪器（或测量条件）称为子机或从机，对应的光谱称为主机光谱和子机光谱（或从机光谱）。子机可以有多台（或多种测量条件）。

基于光谱校正的模型转移是通过主机和子机光谱之间的联系拟合出对应的转移函数，以保证多台仪器预测结果的一致性和准确性。用来建立转移函数的光谱称为标准光谱，采集标准光谱所用的样品为标准样品。标准样品的选择、制备、运输、存储和采集均需按照一定的规范进行。

同一组样品在主机和子机上测量得到的光谱称为标准（样品）光谱。标准光谱是建立模型转移校正函数的重要依据。根据主机和子机标准光谱进行模型转移的方法称为有标（样）模型转移。在实际应用中，如无法获取标准样品或难以在所有仪器上采集标准光谱，则可采用无标（样）模型转移方法。大部分模型转移算法为有标算法。

按照函数校正的对象，模型转移方法可分为三类：（1）光谱校正方法，如 Shenk’s 算法、直接校正（DS）算法、分段直接校正（PDS）算法、光谱空间变换（SST）法、高阶数据分解和多级同时成分分析方法等，基于 WT 和正交信号校正（OSC）的信号分解算法也可用于光谱的校正；（2）模型系数校正方法，如两步偏最小二乘法、线性模型校正法等；（3）预测结果校正方法，如斜率/截距（S/B）算法、线性插值（LI）法、双模型法等。不同模型转移方法在原理上存在较大差异，在应用效果上也存在差异，在实际应用中，需针对具体情况选择合适的方法。

基于模型系数校正的模型转移方法一般是利用少量的子机光谱对原来的模型进行修正，常用于无标样模型转移。无标样模型转移方法的一般思路是建立广义的“通用模型”。利用小波变换、正交信号校正等信号处理方法可将主机和子机光谱的差异进行扣除，消除两台仪器的光谱差异。也可用双模型法实现无标样模型转移，一个模型用于主机光谱模型的建立，另一个模型用于校正光谱差异带来的偏差。

建立稳健的通用模型也是解决模型转移问题的策略之一。此类方法通常利用光谱预处理、变量选择和全局建模等方法建立起对所有仪器和条件都适用的稳健模型。其中，通过合理的光谱预处理结合合适的建模方法、多模型共识别等方法可达到稳健建模、减小仪器间预测误差的目的。全局建模是指将所有包含差异因素的光谱都加入校正集中，建立一个统一的模型。全局模型能够有效改善不同条件下模型预测结果，但是准确性往往不如单独建立的模型。

二、化学计量学模型全生命周期管理

化学计量学模型全生命周期管理是建立模型、验证模型、使用模型和维护模型的全过程。首先要根据分析的目的和需要达到的目标设定预期目标，并按照一定的规范和流程进行。预期目标的设定可参考分析目标概要（ATP），对测量对象、方法设计、结果需达到的性能水平、风险评估、方法的验证与监控、方法的更新等进行描述。化学计量学模型的开发和使用是一个循环迭代过程，如图 5。在模型开发阶段，主要涉及样本和数据集的选择、建模方法及相关参数的优化、根据预期目标进行模型评估和模型验证以及模型更新和维护。而在模型的日常使用阶段，则需要不断监控并反馈模型性能。模型开发期间获得的任何知识及对性能指标要求都可用于模型监控的性能评估，并用于制定模型的更新条件和维护方案。模型维护方案应在模型进入使用阶段之前确定，并预设模型更新的启动条件。一旦发生可能影响模型性能的情况，立即启动模型更新流程。当测量设备测试地点发生变更时，也应启动模型更新流程。模型更新后需进行再验证，验证内容根据模型更新的程度确定。

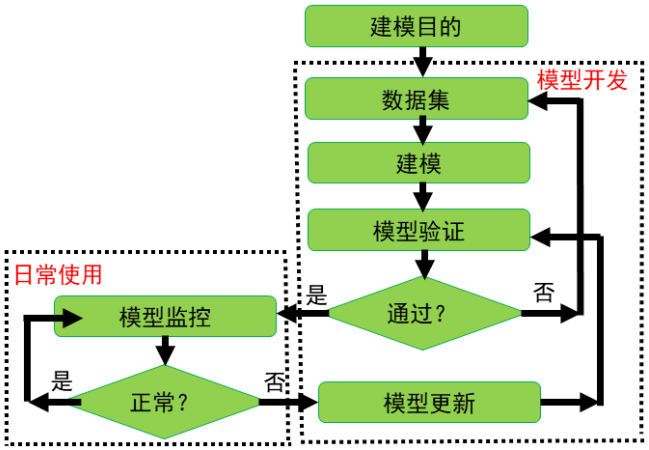

图 5 化学计量学模型的全生命周期管理

1 数据集的建立

在化学计量学模型的建立和分析中涉及多个数据集：校正集、验证集和预测集。校正集是指用于建立模型的数据集，也称为训练集。验证集是指用于模型验证的数据集，主要用于建模过程中的模型评价，目的是得到模型的最优参数。预测集是指模型建立以后用于模型验证的数据集，也称为测试集。校正集和验证集中的目标值（或分类）是已知的，通常由标准方法或参考方法获得，因此也称为参考值。理论上，预测集的目标值（或分类）是未知的，但在实际

应用过程中一般也通过标准方法或者参考方法进行测量，用于模型预测准确性的评价。需要特别说明的是，预测集样本的选择和参考值的测量必须独立于校正集，因此预测集也被称为独立验证集。

数据的评估

在多变量数据分析之前，可使用统计工具评估数据的质量，建议使用统计图表对数据进行观察，评估数据的科学性和合理性。例如用直方图、箱线图评估数据分布，用散点图检测相关性等。描述性统计量（如均值、标准差、方差、中值、最小值、最大值和下/上四分位数）有助于在多变量分析之前对每个变量分别进行快速评估，并检测超出范围的值和异常值、异常分布或不对称性，揭示数据集中的异常情况；相关性检测可以揭示数据中变量之间的相关性。此外，对于大量多变量数据可以采用主成分分析进行降维后再对数据进行考察。

异常样本是指模型无法很好地进行描述的样本。异常样本可能源于原始数据中的意外干扰或测量误差，导致其自变量和/或因变量的参考值失真。模型预测的异常值除由测量仪器、测量条件或者样本之间的相互影响等干扰引起外，还可能由样本超出模型范围而产生。前者应对异常样本进行剔除，后者则为有价值的信息来源，经确认后用于进一步研究以确定现有模型是否需要更新。对于分类模型研究，应分别对每类样本数据进行异常值考察。

定量分析中，化学计量学模型是基于参考方法（如 HPLC 法、UV 法）提供的参考值的二次定量方法，误差可能来源于参考方法，和/或样本本身。为了尽量减小数据的误差，建议严格按照标准操作规程（SOP）进行数据采集。数据的误差决定了模型的预测准确度和精密度。一般地，建模方法带来的误差往往并不显著，但不能据此评估模型的预测误差是否比参考方法的误差更显著。

建模数据集

数据通过评估后，需进一步进行样本的选择、测试数据的预处理，当测试数据变量较多时还需进行变量选择，以获得建模数据集。一定程度上建模数据集的质量直接决定了模型的质量。

校正集样本的选择是模型开发的关键，应科学合理地选择具有代表性的样本，同时需关注样本变量的类型、范围和样本量。选定样本的目标值范围应覆盖模型应用时可能的波动范围。预测值的波动范围可能受到如粒径、批次、人员、日间以及其他变化的影响。大批量生产时，校正集应包含不同规格的多批

次样本，可考虑加入中试批次样品以扩大样本集范围。建议采用基于风险的方法来确定校正集样本中变量的影响因素。

选择校正集的样本量取决于样本的复杂性，即样本的变量类型和强度（如样本的组成、外观等），有时也取决于参考值的分析方法。一般地，样本量越大，在整个校准范围内获得正确结果的可能性越高。当获取新的建模样本困难时，运用 DoE 和/或历史数据库方法可作为构建模型的替代方法。除校正集外，验证集和预测集也需要具有一定的代表性和样本量。

此外，样本选择时还应注意样本分布的均匀性，可以基于目标值的分布情况，选择目标值分布尽量均匀的样本，也可以根据测试数据，选择测试数据分布尽量均匀的样本。Kennard-Stone（KS）方法是最常用的方法之一。当测量数据变量数较多时，可在主成分空间中使用 KS 算法进行选择。

2 模型的建立

是根据数据集的性质选择合适的建模方法并对模型参数进行优化的过程。模型优化是一个繁琐的反复过程，对需要优化的各种因素进行探索，并对模型进行验证，优选出最佳方法和最佳参数。模型优化时需要注意的是避免模型的过拟合。

化学计量学方法的选择

许多化学计量学方法均可用于模型的建立。由于不同化学计量学方法在原理上存在明显的区别，在具体应用中需根据具体的任务和数据集选择相对合适的方法。此外，样本的选择、数据预处理方法、变量的选择以及化学计量学方法的相关参数均会影响模型的性能。因此，方法的选择应综合考虑目标任务、可获得的软件以及对方法的熟悉程度等。但对于特定的数据集，通常很难提前预知哪种算法最为合适，需要进行探索。

采用化学计量学方法建立模型时需要注意的是，许多算法在使用中都具有一定的经验性，特别是在确定模型参数时，仍缺乏科学意义上的标准，需要借助经验进行判断。因此，无论采用什么建模方法，所建立的模型都需要进行严格的评估，以确保模型给出的结果具有正确性并且模型性能符合预设目标。一般需通过模型验证技术对模型进行评估，并最终由独立验证数据集进行验证。

模型的优化

是建立化学计量学模型的必要步骤。建模数据集和方法确定后，仍需要进

行数据预处理、选择合适的变量选择、确定合适的模型参数。尽管通过数据的评估已发现并剔除了数据集中的异常样本，但对建模过程中新发现的异常样本，也应判断是否需剔除。

通常，原始数据可能不是最佳的分析数据，通常需首先进行数据预处理，以突出感兴趣的变异并减少与目标值无关的变异。在实际应用中，数据预处理方法经常与样本选择、变量选择等方法一起交替使用，通过不断地循环得到最优的方法和参数。

当自变量数目较大时，一般采用变量选择技术选择对模型有较大贡献的变量。应谨慎使用预处理技术。任何数据处理技术都会使原始数据发生某方面的改变，甚至丢失部分信息。只有使用得当才可以产生数据增强的效果，反之则可能导致有用信息的丢失。

不同化学计量学方法模型参数的优化的方法各异，对于常用的 PLS 法，仅涉及一个参数（即因子数或潜变量数），而某些算法则涉及多个参数，如 SVM 法中的核函数类型和敏感因子，而卷积神经网络（CNN）等 DL 方法则涉及更多的参数。

模型的整体性能取决于各个环节的匹配，例如进行不同的数据预处理可能会导致模型最优参数和异常样本的差异。变量的选择对模型参数也有影响，选择变量数的差异会导致模型的最优参数的变化。为了得到各环节的最佳方法和参数，模型优化往往需要反复进行。若要得到最佳的方法和参数，则需对所涉及到的方法和参数进行排列组合，逐一尝试。但通常经验也可以帮助获得相对优化的模型。例如，在药品 NIR 光谱的定量模型中，一般采用一阶或二阶导数消除变动的背景，使用 SNV 或 MSC 进行散射校正，采用 PLS 方法建模绝大多数情况下均可得到较满意的结果。

防止过拟合

过拟合是化学计量学建模分析中常见的问题之一，是指由于过度的优化使所建立的模型对校正集数据具有很好的拟合效果，但不能对预测集（独立验证集）样本进行较好的预测。避免过拟合十分重要。

过拟合产生的原因主要来自两个方面：1）校正集样本的代表性，当校正集样本数量不足，尤其代表性无法全面涵盖预测样本时，容易建立过拟合的模型；2）模型的过度优化，由于几乎所有的建模方法都是基于“最小二乘”的原理，过

度地拟合校正集数据形成过度优化的模型。过拟合会使模型失去推广应用价值，因此建模时，一方面要保证校正集样本的数量和代表性，另一方面要避免过度优化。

一般地，建模时，可通过比较模型校正集和验证集的预测误差判断模型是否过拟合：验证集的预测误差与校正集参考值的误差应处于同一水平；当验证集的预测误差显著大于校正集参考值的误差时，则模型过拟合。模型优化后，应进一步采用预测集进行验证，三者的预测误差应处于同一水平，否则该模型过拟合，应重新调整模型参数。

3 模型的验证

由于数据处理的特殊性，除对测量方法进行验证外，还须对化学计量学模型进行验证。但在 3Q 认证的保障下通常只需对化学计量模型进行验证。

模型验证的目的是评估模型并获得优化的模型。模型验证包括建模时的内部验证和模型优化时的外部验证。内部验证包括基于校正集的交叉验证和基于验证集的验证，用于模型参数的调整和优化。外部验证为基于独立验证集（预测集）的验证，用于模型性能的评价，亦称为独立验证。

验证数据集

验证数据集的设计与建模数据集相同，应遵循“代表性”原则，并保证验证数据的可靠性。当从同一数据集中划分验证集和校正集时，验证集数据大小应为校正集数据大小的 20~40%。独立验证集的设计与内部验证集（校正集和验证集）一致。一般地，当具有足够样本代表性时，验证集越大（＞40%），模型验证结果越可靠。

交叉验证

是按固定比例重复地将校正集数据划分为校正集和验证集，用校正集建立模型，用验证集验证获得预测值，并通过预测值和参考值的比较获得预测误差的内部验证方法。

交叉验证的实现形式，通常采用 k 折交叉验证法，即将校正集数据划分为 k 个大小相同（或相近）的子集进行交叉验证，在每一次的迭代计算中，一个子集用于模型验证，其余 k-1 个子集用于建立模型；重复 k 次直到所有子集均作为验证集输出预测值。其中，数据划分可以随机分组，也可按照参考值的排序进行划分，还可以根据先验信息划分数据子集，例如将多批次样品的数据集按批

次划分为不同子集。

最简单的 k 折交叉验证为留一交叉验证（LOOCV），即每次取出一个样本用于验证，其他样本用于建模；LOOCV 一般用于校正集样本数较少的情况，其结果易受异常值影响。交叉验证的常用方法是自举法，即蒙特卡洛交叉验证（MCCV），在每次迭代中随机从校正集中抽样一定比例的数据用于建立模型，而其余的数据用作预测集；经多次重复计算，通常为 100 次或更多，以获得较为稳定的预测误差；每次计算随机取样的比例可以控制在 50~80%。

由于交叉验证中没有使用校正集以外的数据，为内部验证，主要用于模型参数的优化，只能表明模型对校正集数据具有较好的拟合能力，不能保证模型的广泛适用性。特别是当校正集数据的代表性不足时，所建模型有过拟合的风险。

验证集验证

是指利用固定的验证集进行模型验证。验证集一般是从校正集中抽取。样本抽取的原则是尽量使验证集具有代表性，应与剩余的校正集数据具有相同或相似的分布。尽管验证集验证时采用了固定的验证集，有利于模型优化误差大小的比较，其仍属于内部验证。

外部验证

又称为独立验证集（或预测集）验证，是采用独立准备的验证集进行验证的模型验证方法。用于外部验证的验证集应独立于校正集。外部验证的目的是验证模型是否具有实用性，即是否达到了建模的预期目标（ATP 的要求）。因此预测集中的样本应确保包含模型预期目标要求的使用范围内的所有变异。

独立验证集的“独立性”首先取决于建模的预期目标，特别是模型的使用范围，如某品种药品的活性药物成分（API）定量模型、某标示厂家药品各组分含量的定量模型、某品种某厂家某规格药品的质量一致性模型等。对于某品种药品的通用模型，变异因素应考虑到生产厂家、配方、规格、批次等，对于药品的质量一致性模型，只要考虑生产过程的工艺变动即可。其次，样本的来源应具有“独立性”。对于实验样本，要求独立验证集的样本不能来自同一组实验。尽管实验条件完全相同，独立验证集的实验（包括样品的制备和实验数据的测量）需独立于校正集。对于来源于生产批次的样本，独立验证集样本的批次不能与校正集样本相同，最佳的选择是使用模型建立后新生产批次的样本进行外

部验证。

模型性能的评价指标

一般采用一组指标参数或统计图表对模型验证中模型的性能进行评价。指标参数可直接反映模型的预测性能，统计图表则可直观地对验证集样本的预测准确性进行评价。由于一个指标参数只能反映模型性能的某一方面，通常需通过多个指标参数对模型性能进行综合评价。此外，采用不同的作图方式还可以表征出模型对不同样本集预测性能的差异。

对于定量模型，误差是最常用的表征模型性能的指标。方法的预期目标决定了误差的性质，如定性方法的误分类率和定量方法的预测误差。在交叉验证过程中通常有两种指标用于描述验证的误差，即校准均方根误差（RMSEC）和交叉验证均方根误差（RMSECV）。RMSEC 又称为自验证误差，用于直接表征模型对所有校正集样本预测结果的评价；RMSECV 用于表征交叉验证中模型对所有校正集样本预测结果的评价。通过这两个误差值随某个模型参数的变化可以对模型的性能与参数之间的关系进行考察，例如对过拟合的判断。在验证集验证和外部验证中，通常使用预测均方根误差（RMSEP）对模型进行评价。

其中，均方根误差的计算公式为：

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (7)$$

式中 n 为验证集样本数量，$\hat{y}_i$和 y_i 分别为第 i 个验证样本的预测值和参考值。交叉验证中，验证集为校正集。

误差的大小用于描述预测结果的准确性，误差越小说明预测结果与参考值越接近。此外，预测值与参考值之间的相关性也可以用于预测结果的评价，相关性越高表明预测结果越好。相关系数（r）可用于描述相关性。另一个描述相关性的参数是决定系数（R^2），定义为模型已解释变差（回归平方和，SSR）占总变差（总平方和，SST）的比重，公式为：

$$R^2 = \frac{S_{SSR}}{S_{SST}} = \frac{S_{SST} - S_{SSE}}{S_{SST}} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (8)$$

式中 $S_{SSR}=\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$，表示模型已解释离差平方和，$S_{SST}=\sum_{i=1}^{n}(y_i - \bar{y})^2$，表示参考值的总离差平方和，SST 中模型未解释变差为残差平方和（SSE），$S_{SSE}=\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$，表示预测值与参考值差值（或预测残差）的平方和。

在数值上，决定系数（R^2）是相关系数（r）的平方。由 R^2 定义可知，残差

S_{SSE} 越小，R^2 越大，模型预测结果与参考值越接近。S_{SSE} 趋近于 0 时，R^2 趋近于 1，S_{SSE} 与 S_{SST} 相等时，R^2 等于 0。

预测相对标准偏差（RPD）也常用于模型预测性能的评价，定义为参考值的平均标准偏差与预测均方根误差（RMSEP）的比值，公式为：

$$RPD = \frac{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2/n}}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2/n}} \quad (9)$$

式中的 n 有时用（n-1），即用自由度代替样本数。从公式可知，分子部分为 S_{SST} 的均方根，而分母部分为 S_{SSE} 的均方根，因此 RPD 表示总体偏差与预测偏差的比值。该比值越大则预测结果越好。实际应用中，定性模型的 RPD 值应大于 3，定量模型的 RPD 值应大于 5。

利用上述指标参数虽可对验证集的整体预测结果进行评价，但无法观察每个样本的预测结果。在模型评价时，建议采用预测值与参考值的关系图、误差分布图以及误差与参考值的关系图对模型性能进行考察（图 6）。图 6（A）中增加整体预测结果的统计值及趋势线（实线）和对角线（虚线），能更直观地观察预测值与参考值的关系。更重要的是，上述三个图可以有效表征每个样本的预测效果，特别应注意误差较大的样本，可能为异常样本。例如，图 6（B）中，某些样品呈较大的正偏差或负偏差，可根据模型的预期目标增加允许的误差限，将超出允差的样本判别为异常样本；图 6（C）中显示的预测误差与参考值之间的关系，预测误差与参考值无明显的相关性，但在参考值 200 附近的样本误差分布较宽。当预测误差与参考值存在明显的相关性时，说明模型的拟合效果不好，其原因可能来自多个方面，常见原因之一是测量数据与参考值之间存在非线性关系。

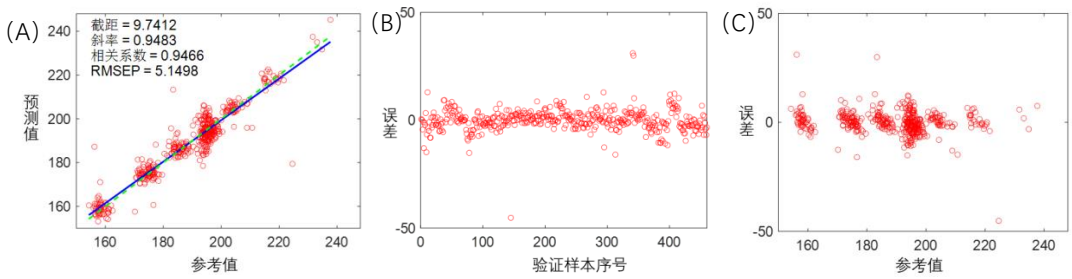

图 6 验证集样本预测值与参考值的关系图（A）、验证集样本预测结果的误差分布图（B）、验证集样本预测误差与参考值的关系图（C）

对于定性模型，一般采用真阳性率（TPR）和假阳性率（FPR）对预测误

差进行评价，也可同时使用真阴性率（TNR）、假阴性率（FNR）和准确性（ACC）三个参数评价。上述参数公式如下：

TPR = TP / P = TP / (TP+FN)，即：真阳性样本数 /阳性样本总数；

FPR = FP / N = FP / (TN + FP)，即：假阳性样本数/ 阴性样本总数；

TNR=TN/ N = TN / (TN + FP)，即：真阴性样本数/ 阴性样本总数；

FNR=FN / P = FN / (TP + FN)，即：假阴性样本数 /阳性样本总数；

ACC = (TP + TN) / (P + N)，即：（真阳性样本数+真阴性样本数）/总样本数。

TPR 有时也称为灵敏度（SEN）、检出限（LOD）等，TNR 有时也称为专属性。

其中，P、N、TP、TN、FP、FN 分别表示阳性样本数、阴性样本数、阳性样本被正确识别的数量、阴性样本被正确识别的数量、误报的阳性样本数量（被模型预测为阳性的阴性样本）和误报的阴性样本数量（被模型预测为阴性的阳性样本）。受试者工作特征曲线（ROC）常用于对定性模型的综合评价。如图 7，ROC 曲线以 FPR 为横坐标，TPR 为纵坐标。横轴越接近零准确率越高，纵轴越大代表准确率越好。曲线越接近左上角（横越小，纵越大），预测准确率越高。曲线把整个图划分成两部分，曲线下方部分的面积被称为 AUC，用来表示预测准确性，AUC 值越高，即曲线下方面积越大，则预测准确率越高。当完全随机分类时，定性模型为无用决策过程，其 ROC 曲线为对角线，AUC 值为 0.5。

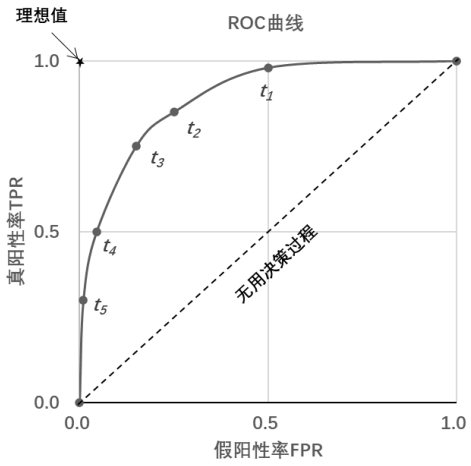

图 7 ROC 曲线

4 模型的评估

一般是指经过优化建立模型以后对模型性能进行的综合评价，即评估是否达到建模的预期目标，应使用独立验证集评估模型的性能。模型评估时建议重新对建模的方法和优化的参数进行科学性和合理性分析，包括建模算法、选择的变量、预处理方法及相关的参数等。除对模型本身的评估外，还应参考分析方法评估的相关内容。

定性模型的评估

对于定性模型，一般对专属性和稳健性两个关键的参数进行评估。

① 专属性

指模型的辨别能力，通过正向验证和反向验证评估。验证集样本必须是代表样本典型差异的一组代表性样本，其中正向验证集样本除相同化学组成外，还应包括其他参数（如样本的多态性、粒度、水分等）的样本；反向验证集样本应包括模型识别对象（阳性样本）以外的各种可能的阴性样本，尤其是具有混淆风险的样本，如具有相似外观的样本、化学组成和结构相似的样本以及具有不同物理特性的样本。定性模型应具有辨别正向和反向验证集样本的能力。若模型的专属性不足，应继续优化模型参数并重新验证方法。

每当引入可能影响模型辨别能力的新影响因素时，应重新验证模型的专属性。重新验证可以仅限于新的影响因素，不一定对模型进行全面的重新验证。

建议使用 ROC 曲线评估专属性。

② 稳健性

又称为粗放性。为了验证模型的稳健性，验证集设计时应综合考虑样本、仪器和测试方法等方面的因素，包括各方面的变动因素，并综合考虑关键的参数（如温度、湿度、分析设备的仪器性能），通过改变这些参数来验证分析方法的适用性，可应用 DoE 进行各参数的设计。

同样建议使用 ROC 曲线评估稳健性。

此外，SEN 或 LOD 也是反映分析方法和定性模型性能的参数，必要时也应进行评价。

定量模型的评估

对于定量模型，一般对专属性、线性、范围、准确度、精密度和稳健性等性能参数进行评估。如另有需要，也可进行其他方面的评估。

① 专属性

用于评估模型的适用范围。验证集样本中应包括不同含量干扰物（如结构相似物质）的样本，考察干扰物的含量对模型预测准确性和精密度的影响。验证集也可以包括掺假的样本、未经过认证的样本等。通常较难直接获得模型预测范围外的认证样本（不合格样品），但可以通过模拟制作的方式，向验证集加入该类样本，验证模型的适用性，以提高专属性评价的置信度。定量模型应包含排除模型不适用样本的判断标准。此外，专属性验证时还需要对异常样本进行判断，可参照“数据的评估”和“模型性能评价指标”中的方法，也可以使用其他识别异常样本的方法。

② 线性

线性是指模型的预测结果与参考值之间的相关性评价，评价指标包括决定系数 R^2、斜率、截距以及对角线的重合程度（图 6）。线性验证的独立验证集应涵盖模型的整个范围。可使用交叉验证进行线性的验证，但不应取代使用独立验证集的评估。

③ 范围

范围是指模型的适用范围，包括测量方法和模型两个层面各自的检测限和定量限。验证集中应该包括低于检测限的样本，以确保此类样本首先被正确识别。在模型定量范围内的样本，预测结果应符合预期目标中确立的准确度和精密度要求。

④ 准确度

模型的准确性可用“模型性能的评价指标”中的误差进行评估。验证集的参考值应尽量覆盖模型整个范围。可使用交叉验证评估模型的准确性，但是不应取代使用独立验证集的评估。采用独立验证集进行评估时，RMSEP 应与 RMSEC 和 RMSECV 在同一水平，且符合预期目标的要求。

⑤ 精密度

模型的精密度可通过计算模型预测结果的标准偏差来评估。精密度验证可以包括预测结果的不确定性分析，并对每个变异因素对不确定度的贡献进行分析。精密度评估应包括测量重复性（同一模型对同一样本重复测量结果的相对标准偏差）、方法精密度（平行样本测量结果的相对标准偏差）以及日内精密度、日间精密度、中间精密度（测试人员间、仪器间和实验室间测量结果的相

对标准偏差）等。

⑥ 稳健性

定量模型稳健性的评估原则与方法同定性模型，但需谨慎对待各参数对准确度和精密度稳健性的影响。同时，建模方法与模型参数，如预处理方法、选择的变量、PLS 模型的因子数等，对预测结果的影响也应该考虑在稳健性的评估范围。

也可使用参考值超出范围的样本或不同类型的挑战样本对模型稳健性进行评估，模型应将这些样本正确地识别为异常样本。

5 模型的监控、更新与再验证

模型的性能可能会因样本因素、仪器因素、工艺变更等发生改变。因此，在模型的日常使用中，应制定持续的模型监控方案，按计划进行模型性能的考察，必要时进行模型更新和再验证。

持续保障模型性能的措施包括：模型持续监控和模型审查、风险评估、评估模型使用中的变异因素和预设的模型维护方案、按需进行模型更新和再验证。

模型监控

应贯穿于模型全生命周期管理的全过程，应制定和记录在模型开发和日常使用中检查模型性能的可控方案——明确模型持续监控及性能评价的必要因素，制定监控、分析和调整模型的计划，并设置执行次数，以识别模型的关键性能和相关的偏差。

模型维护可视为风险评估方案的一部分，并与其关键程度相匹配。如模型适用，且相关测试仪器与测试条件未发生变化，则应遵照持续模型持续验证方案进行模型监测。模型持续验证的触发条件为：1）预设的时间间隔；2）出现可能影响模型性能的事项，包括原材料或制造商的变更和/或上游工艺改变及引起的样本质量变化（如工艺设备或操作设置改变引起模型预测结果的偏离或超标）；3）模型判别和相关操作规则发生改变。

模型诊断可以保障模型用于新样本预测时的有效性，不应出现离群值。注意：异常值不一定是不合格产品，通常为无效结果。多变量模型可能比单变量模型更容易受到异常数据的影响，应：1）从统计上证明多变量模型诊断的合理性；2）核实模型开发和方法验证过程中的数据；3）将多变量模型诊断作为可控方案的一部分加以监控；4）定期开展模型预测结果与参考方法结果的比较，

作为触发事项日常审查的重要环节。

模型审查用于决定是否需要延长模型维护的周期，可根据风险评估的结果决定，也可根据预设的模型维护流程决定。模型维护主要包括重新界定模型的使用范围、调整校正集（添加、替换或移除样本）、甚至重建模型等。模型维护及其理由须科学合理并记录存档。

模型监测通常需要一个或多个质控样品。质控样品应尽可能与校正样品相似，参考值的测定方法、测量条件均应与校正样品一致，测量数据应与模型相匹配，并在模型的范围之内。

模型更新

随着时间的推移，仪器的变化可能会使模型逐渐无法正常使用，而仪器维护或仪器更新往往会导致模型性能减退甚至失效。在这种情况下，需要更新模型或开发新的模型。模型更新的理由主要可以分为两类：一类是校正集需扩展，另一类是测量系统发生变化。前者由于校正空间的扩大，如新成分的添加或其他校正集中未包含的变异，使原始模型不再有效，需用新的样本补充新的变异，扩展校正空间。后者通常是由于测量仪器的老化、维护或更新，如更换光源、光学元件性能衰减、波长校正等。模型更新为模型验证的一部分，应在模型的开发和日常使用中记录其运行状况及更新的内容和理由。

实践中，有多种模型更新技术可用于模型更新。如何选择适当的模型更新技术需基于对潜在原因的充分理解。应优先采用更便捷的模型更新方法。最常用的模型更新技术是模型转移和拓展校正集。当模型转移和拓展校正集无法奏效时，则考虑重新建立模型。

在扩展校正集时，应考虑待添加的新样本数量及其对校正集整体构成的影响。必要时，可移除部分原校正集样本，或利用样本选择方法重新确定新的校正集样本。同时，拓展校正集后还需要对异常样本进行识别和剔除。如果新样本与原有校正集样本在空间形成异常的分布或聚类，则说明新样本不适用于更新原有模型，应使用新样本重新建模。

模型再验证

指模型更新后重新对模型进行验证的过程。模型再验证时对验证集样本的要求与原始模型验证时相同。作为分析方法验证的一部分，模型再验证应基于模型更新的理由和措施，从科学论证和实验方法等方面证明模型的有效性，以

819 保障再验证后模型的性能。

起草单位：中国食品药品检定研究院、南开大学、沈阳药科大学、广东省药品检验所、上海市食品药品检验研究院、海南省药品检验所、河南省药品医疗器械检验院
主要起草人：赵瑜、邵学广、尹利辉、李清、邱蕴绮、杨永健、陈露、王骁、王立萍
复核单位：江苏省食品药品监督检验研究院
联系电话：010-53851546，010-53851547

化学计量学指导原则增订说明

一、背景目的

与传统数据分析方法相比，化学计量学方法利用数学和统计学方法对多变量数据进行计算，通过多个变量的数学变换和统计分析得到样本的类别和特征，进而实现定性和定量分析。美国药典、欧洲药典均收录了化学计量学指导原则，将化学计量学作为过程分析技术的一个重要和关键的组成部分，在药品生产和质量控制中推广和应用。目前《中国药典》尚未收载化学计量学通用技术要求。此次针对我国药品生产、检验和监控等工作的需求，探索建立适合我国制药行业发展现状的化学计量学指导原则，指导分析实践活动中的数据质量控制、分析方法的建立及分析方法的验证，以保障多变量分析方法的科学性和分析结果的可靠性，推动化学计量学在制药领域的应用。

二、起草过程

本指导原则从我国制药行业的发展现状出发，考察化学计量学在我国制药工业水平下的应用需求和方向，在传统数据分析方法的基础上，提出化学计量学方法及其在药品生产、检验和监控等工作中对数据分析的特殊要求。指导原则起草参考了国外药典、国家标准、行业标准等化学计量学相关的标准、规范及指导原则，并适当引导鼓励化学计量学新技术新方法应用于药品生产和质量控制实践的数据分析。

三、主要内容

本指导原则结合国内药品领域在研发、质量控制和生产过程实际需求，介绍了如何使用化学计量学处理和分析典型多元数据，为相关应用的数据分析和解释提供科学合理的、原则性的指导原则和实践指南，包括数据处理、多元统计、多元分辨、多元校正、聚类与判别以及模型转移等，适用于使用不同分析技术（如光谱、色谱等）和用于不同目的（如鉴别、分类、含量预测等）的应用。

化学计量学方法的应用属多变量分析方法，与常规分析方法的开发和验证有所不同，本指导原则就化学计量学实践中如何开发、验证化学计量学模型并

进行适当的生命周期管理，来确保方法的质量和性能进行了详细介绍，包括数据的预处理、模型的建立、模型的验证、模型的评价、模型的监测、更新与再验证等。

四、需要说明的问题

1. 本指导原则介绍的内容不是化学计量学的全部内容，而是可能与药品生产和质量控制实践相关的用于数据分析的化学计量学方法。

2. 本指导原则简化了对化学计量学方法详细原理与算法的描述，更多地强调了指导从业人员在药物分析中如何使用化学计量学方法（如数据的准备、参数的设置、稳健性和一致性的评价等），并对方法的注意事项和局限性提出指导意见。

3. 对于本指导原则未涉及的其他有效的化学计量学方法，不受本指导原则限制，均可采用。