代谢组学数据处理——主成分分析十个要点问题

阿基业*, 何 骏, 孙润彬

(中国药科大学药物代谢动力学重点实验室, 代谢组学研究室, 药物分子设计与成药性优化重点实验室, 天然药物活性组分与药效国家重点实验室, 江苏 南京 210009)

摘要: 代谢组学研究所产生多变量数据常采用主成分分析方法进行处理和评价, 主成分分析涉及抽象的空间模型、复杂的理论计算、精细的数据转换, 需要准确理解和把握主成分分析算法原理和特点。本文从主成分、主成分得分、主成分载荷、缩放与权重、偏最小二乘关联分析与判别分析、隐结构正交投影分析、隐结构双向正交投影分析、S-形图、共享与特有化合物结构分析、模型验证等十个方面, 以简洁、易懂的语言介绍了代谢组学数据处理常用的主成分分析方法中的重点和难点问题, 方便广大代谢组学研究人员更好地熟悉和了解代谢组学数据处理方法, 以合理选择数据处理模式、规范数据处理程序、熟练解析数据处理结果, 并得出可靠结论。

关键词: 代谢组学; 主成分分析; 系统生物学; 多变量数据分析; 主成分

中图分类号: R969 **文献标识码**: A **文章编号**: 0513-4870 (2018) 06-0929-09

Multivariate statistical analysis for metabolomic data: the key points in principal component analysis

A Ji-ye*, HE Jun, SUN Run-bin

(*Jiangsu Province Key Laboratory of Drug Metabolism and Pharmacokinetics, Laboratory of Metabolomics, Jiangsu Key Laboratory of Drug Design and Optimization, State Key Laboratory of Natural Medicines, China Pharmaceutical University, Nanjing 210009, China*)

Abstract: Metabolomics data contains multiple variables usually processed and evaluated by means of principal components analysis. The statistical analysis of the multivariate data is involved in abstract, elusory fitting for the model of hyperspace, complicated theoretical arithmetic and sophisticated transformation of the data matrix. It is crucially important to understand the arithmetic mechanism and the properties of the models fully. In this article, we reviewed the key and puzzling issues in principal components analysis of the metabolomics data, including the principal components, the scores and loadings of a principal components, scaling and weighting, partial least square projection to latent structures, partial least squares discriminant analysis, orthogonal projection to latent structure, orthogonal bidirectional projections to latent structures, S-plot, shared and unique structure plot, and the validation of the model. Hopefully, this article provides a better understanding of data processing mode, model selection, procedure standardization, and data interpretation for a reliable conclusion.

Key words: metabolomics; principal component analysis; system biology; multivariate statistical analysis; principal component

收稿日期: 2017-12-25; 修回日期: 2018-03-25.

基金项目: 中国新西兰政府间科技合作重点项目 (采用中药治疗耳鸣: 以代谢网络为目标, 2017YFE0109600); “十三五”国家重大新药创新专项“基于药动−药效结合的组分中药与创新中药研发关键技术”(2017ZX09301-013).

*通讯作者 Tel: 86-25-83271081, E-mail: jiyea@cpu.edu.cn

DOI: 10.16438/j.0513-4870.2017-1288

代谢组学[1, 2]研究各种生物体受个体自身基因控制、蛋白质作用和系统调控下内源性小分子随生长、发育、变化 (突变)、衰老, 或受到外源性因素, 如病原、环境、药物或毒物刺激而导致的变化。代谢组学所研究的内源性小分子既是执行生化代谢反应的核心物质、反应底物或代谢产物, 也是生物体分子信号传导、信息传递与反馈的关键物质, 反映了基因组、转录组、蛋白组、外环境、物质等多因素作用下综合的终末效应。代谢组学所研究的内源性小分子变化可提示与机制相关的代谢标志物、代谢通路、基因表达、酶活性与功能, 借助代谢组学、基因组学、转录组学、蛋白组学等系统生物学、网络药理学结合手段, 有望为阐释生命科学规律、解析疾病发病机制、发现药物作用靶点和研究药物作用机制提供强有力手段。经过20年的发展, 代谢组学广泛应用于生命科学研究领域并已经取得了显著成就[3−5], 与基因组学、蛋白组学有所不同, 代谢组学所研究的内源性小分子结构与理化性质差异巨大、种类繁多、浓度相差悬殊、检测难度大、化合物鉴定和分析困难。利用现有先进的仪器手段, 如基于质谱 (MS) 检测和核磁共振 (NMR) 技术手段, 已经能检测和鉴定其中数千个小分子[6, 7]。由于代谢组学所研究的内源性小分子涉及到许多代谢通路, 而不同代谢通路中内源性小分子结构、性质差异很大, 在代谢组学研究开始阶段, 多采用非靶向代谢组学方法, 即对多类小分子或多个代谢通路中物质尽可能进行全面检测[8]。与之对应, 在进行非靶向代谢组学研究后, 或者在其他研究工作基础上, 确定目标代谢通路或目标小分子后, 可以采用靶向代谢组学方法, 对一类、一组、一群目标分子进行定量或半定量分析, 有利于更精准、全面地了解该类物质或该通路中有关物质的变化规律[9, 10]。一般来说, 非靶向代谢组学可以采用 NMR、气相色谱质谱、液相色谱质谱、傅立叶变换质谱等方法进行研究, 而靶向代谢组学可以利用毛细管电泳、电化学检测、配备有特殊型号和性能色谱柱的气相色谱质谱、液相色谱质谱进行分析[11−15]。

无论是采用基于 NMR 还是 MS 测定技术的代谢组学平台, 对观测样品/观测变量 (observations) 进行检测后都会获得包括检测变量 (variables) 和信号响应 (responses) 的高通量数据, 其中包涵丰富的小分子信息, 这些小分子化合物少则数十个, 多则几千个。每个样品的数据都包含至少二维信息[16], 对 NMR 测定来说, 包括不同化学位移值 (对应于某个小分子) 及其响应强度 (对应其浓度水平); 对 MS 测定来说, 包括不同保留时间 (对应于某个小分子) 及其响应强度 (对应其浓度水平)。如何对这些含有多变量数据矩阵进行处理是代谢组学所面临的最关键、最核心的问题。虽然理论上可以逐个对每个变量进行分析, 获取有价值的信息, 但这种方法效率低、浪费时间, 而且无法基于整体数据对所测样品的优劣、差异进行综合评价。代谢组学通常借助一些软件 (如 SIMCA p、Matlab 等) 对样品类别、相似性、差异性进行分析, 找出造成样品差异或组间差异的分子, 进一步进行深入的代谢通路分析、生物标志物分析、生物学意义挖掘[17−19]。代谢组学所研究的大量内源性分子及其数据分析对上述研究至关重要, 而多变数数据处理中的主成分分析方法抽象性强[20, 21], 只有准确理解和把握主成分分析原理、方法和特点, 才能正确选择数据处理方式、熟练解析数据处理结果, 得出可靠结论。本文针对代谢组学研究常采用的主成分分析方法中的一些重点、难点问题进行阐述。

1 主成分与主成分分析

采用不同代谢组学工具对生物样品进行分析, 通常可以获得一个数据矩阵, 其中含有 N 个观测样品/观测变量、K 个检测变量/检测化合物及其对应的信号响应强度, 图 1A。主成分分析是多变量数据分析的一个常用的重要方法, 理论上将上述含有 N 个观测变量、K 个检测变量的数据矩阵看成一个 K 维空间的数学模型, N 个观测变量/样品分布在这个空间模型中[22, 23]。从数据分析的本质目的看, 数据分析是为了了解观测变量之间的差异性或者相似性, 为最终的决策提供参考。因此, N 个观测变量分布在 K 维空间中, 总存在某一个维度的方向, 能最好、最大程度地描述观测变量的差异性。基于偏最小二乘法原理, 可以计算得到这个轴线, 使所有观测变量距离该直线的总残差最小, 而投影在此直线方向的方差最大, 即: 这个方向的轴线可以最好、最大程度地描述观测变量的差异性或相似性, 图 1B。并且, 沿着这条轴线方向可计算观测变量总体离散程度参数, 获得描述观测变量总体差异性或相似性的参数, 此为第一个主成分。在此基础上, 可以在与上述轴线垂直的平面上找出第二个最重要的轴线方向, 描述观测变量第二显著的差异性, 获得第二主成分; 以此类推可以获得第三、四、…主成分[16]。

主成分分析最重要的作用是建立低维平面或空间 (通常 2～5 维), 以此分析和概览整个数据集, 并从中揭示出数据集中观测变量的分组、趋势以及离群值。通过分析可以发现观测变量与检测变量, 以及观

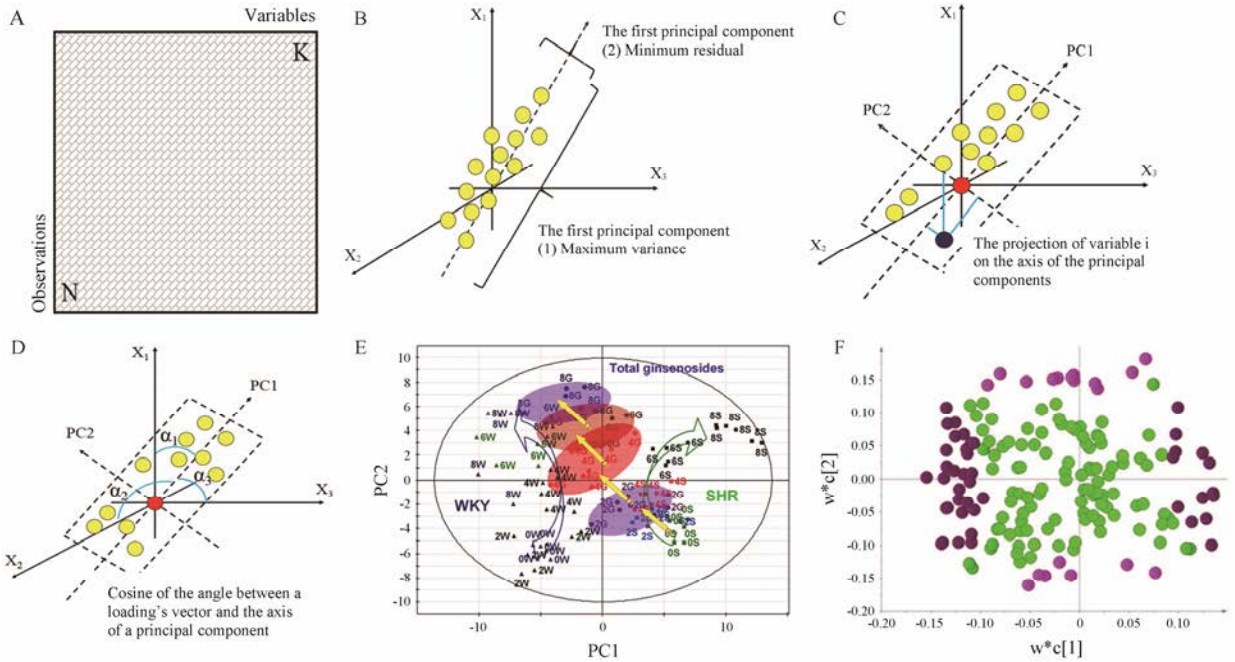

Figure 1 The mathematic model of the principal component analysis. A: Multi-variates data and the fitting of a space model; B: A special axis and the first component based on partial least squares projection to latent structure; C: The projection and scores of each variable in the axis for a specific principal component; D: The loading refers to cosine intersection angle between the two vectors of the variable and a specific principal component; E: A typical scores plots based on a principal component analysis[23]; F: A typical loadings plots based on a principal component

测变量之间、检测变量之间的相似性或差异性关系。由于模型建立之初含有 *K* 维空间，所以该分析主要采用降维手段，重点关注方差大、残差小的前几个主成分所对应的几维空间。由于人类想像很难超过三维的空间，因此，为了便于直观观测，通常取 2 个或者 3 个主成分，对应于二维平面图或者三维立体图进行直观观察和分析。

2 主成分得分与得分图

主成分得分 (principal component score) 是 PCA 分析结果中的一个重要参数，简称得分 (score)，根据主成分得分绘制的图称为主成分得分图。在医药相关的代谢组学研究中，检测变量往往是动物、细胞或者临床来源样品，因此，也称为样品分布散点图。从样品分布散点图中可以直观看出不同观测样品之间距离，据此可判别样品相似性或差异性。这就是多变量数据分析或主成分分析的主要优势与优点。即：将大量的多维度数据转换为可以直观观察的散点图，从散点图可以看出各个观测变量或各组观测变量靠近程度。一般在保证模型的有效性情况下，各个观测变量或各组观测变量之间的距离反映了观测变量相似和差异程度。观测变量之间距离越靠近，其相似性越高，检测变量值相近；反之，如果观测变量之间距离越远，其差异性越大，检测变量值差异也较大。

如图 1C 中，X_1～X_3 为所有 *K* 个变量中第 1、2、3 变量的方向，即 PC1、PC2、PC3 主成分方向。对 *N* 个观测变量中的第 *n* 个，做其垂直于第 *i* 主成分方向的投影，即以该主成分方向为坐标方向，得到观测变量 *n* 在主成分 *i* 上的坐标值，称为 *n* 在第 *i* 主成分上的得分。每个观测变量在每个主成分上都可能有不同的得分。在一项相同来源血浆、血清的代谢组学差异研究中[24]，发现血清与血浆样品总体上分布在平面图的左右两侧，即在第一主成分 PC1 方向血浆和血清样品有近似得分值 (图 2)。但血清和血浆样品随放置/制备时间不同，均出现自上而下移行变化的趋势，说明各组样品在第二主成分 PC2 方向差异明显。即第一主成分方向主要反映了血清与血浆两组样品之间的总体差异，而第二主成分方向反映了放置/制备时间对血清和血浆样品中代谢组所造成的影响。提示在代谢组学研究中血浆和血清样品必须严格区分，而样品放置/制备时间也应相对固定，否则取样时间差异会对研究结果造成不利影响。另外，在 PCA 中一个需要注意的问题是：大部分研究中首先看样品分布散点图 (主成分得分图)，无论是 PCA 还是偏最小二乘判别分析 (partial least squares-discriminant analysis，PLS-DA) 的第一个和第二个主成分对应平面分布图。虽然在大部分情况下这个平面图可以描述

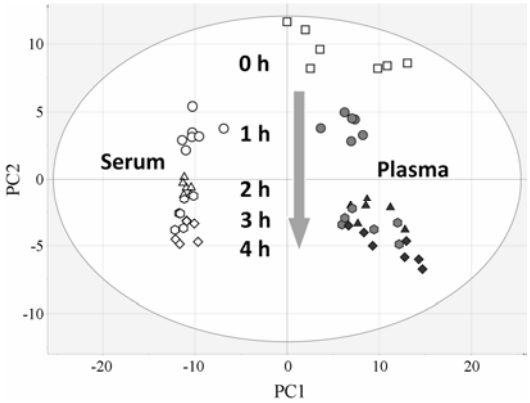

Figure 2　Scores plots of the sera and plasma samples originated identically

模型的主要特点，但如果存在异常样品 (outliers)，或者某个主成分下模型有效性不高 (参考第 10 部分：模型验证)，就可能出现对模型的错误判断和错误解读。

3　主成分载荷与载荷图

样品分布散点图 (主成分得分图) 可以直观反映代谢组学研究中的观测样品/观测变量的相似性或差异性。与之相对应，检测变量/小分子化合物在观测样品中浓度水平的相似性或差异性可以由变量分布散点图 (或主成分荷载图，principle component loading and loadings plot) 展示。几何学上，*K* 个检测变量对应于 *K* 维空间，载荷实质上表征了模型 (线、平面或超平面) 中的每个检测变量在 *K* 维变量空间的矢量方向贡献程度，而主成分载荷表征了主成分模型 (线、平面或超平面) 中每个检测变量的对此贡献。

图 1D 中，X_1、X_2、X_3…为所有 *K* 个变量中第 1、2、3…*k* 变量的矢量方向。对第一个主成分 PC1 来说，其方向与原始 *K* 个变量方向的关系，可以用 PC1 分别与矢变量 X_1、X_2、X_3 等的矢量夹角 α1、α2、α3 等余弦值来表征。这些数值体现了变量 X_1、X_2、X_3 等对 PC1 的贡献，称为载荷。每个变量对每个主成分都有各自唯一的载荷值。因此，每个变量对应于各主成分的载荷决定了其在 *K* 维模型中从原点向外的一条射线方向，再通过计算每个变量对模型的贡献程度大小，就可以确定每个变量载荷在 *K* 维变量空间的确切方位，以各主成分上各变量的载荷作图即为载荷图。

在一项利用自发性高血压大鼠 (SHR) 进行中药整体代谢调节研究中，采用主成分分析方法对 SHR、人参总皂苷给药 2、4、6、8 周后 SHR 及正常对照大鼠 (WKY) 代谢表型进行研究[23]，得到以主成分 1 和 2 (PC1 和 PC2) 构成的二维平面图，即观测变量得分图/样品分布散点图 (图 1E) 与对应的检测变量载荷图/变量散点分布图 (图 1F)。可以分析观测变量/样品分布与检测变量的关系、各个检测变量对观测变量的影响大小和程度以及检测变量之间的相关性。首先，无论观测变量得分图/样品分布散点图 1E，还是检测变量载荷图//变量散点分布图 1F，其中变量越靠近说明相似性越高，矢量方向一致的变化趋势一致，矢量方向相反的变化趋势相反[22]。同样可以对应分析 1E、1F 图中观测变量与检测变量之间的关系。位于 1F 图 0 度轴右侧方向的检测变量在 1E 图对应位置观测变量/样品中浓度水平高，相反，在矢量反方向，即 180 度左侧轴线方向观测变量/样品中浓度水平最低。其高低差异程度取决于各变量位置与原点绝对距离。距离越远，差异程度或倍数越大。在得分图和载荷图上，总是存在这样的规律，距离原点近的变量对模型贡献较小、距离远的贡献大。

4　缩放与权重

在含有多个变量的研究和实际数据分析中，变量数值常常具有不同的数量级。相对而言，数量级大的变量方差数量级也较大，数量级较小的变量方差数量级也较小。主成分分析主要基于方差数据进行投影和计算模型，因此，对于有较大方差的变量，其在主成分分析模型中占有的权重比方差较小的变量更加显著。但是，多变量数据分析中各个变量的实际重要性并不依赖于上述数量级大小和方差大小，即有可能数量级较小的变量更加重要。因此，按照原始观测数据，数量级小的变量方差权重较低，其重要性就不能在模型里充分表达。解决上述问题的方法，是对各变量进行标准化，使各变量的数值范围按一定标准进行调整。这种对变量数据进行规整化的处理称为缩放或者权重 (scaling or weighting)。

常用的缩放标准是令各变量等方差，这样的缩放方式称为等方差缩放 (unit variance scaling, UV-scaling)[22]。等方差缩放的代数方法是计算每一个检测变量的标准偏差 (standard deviation, Sk)，然后对此变量乘以 1/Sk 进行缩放。经过如此缩放后，每个变量均具有相同的方差，图 3A 中方差不同的变量集，经过 UV 缩放后，变为 C 中具有相同方差的变量集。当然，进行缩放/权重处理的方式并不只有等方差缩放，可以根据不同数据处理要求，采取其他的数据缩放或权重方式，如 Pareto 方式等。

另外，在数据处理过程中，除了需要对数据进行缩放/权重处理，还默认对数据进行平均值中心化处理[16]。即计算每个变量的平均值，用变量数据减去该

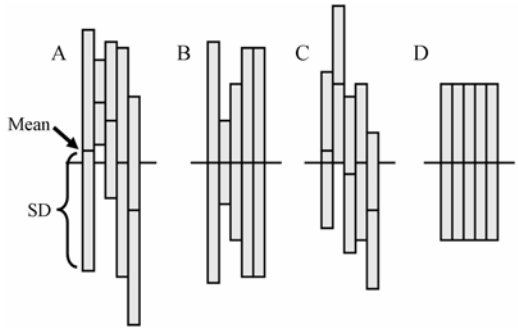

Figure 3 Scaling or weighing, mean centering of the multivariates data

平均值, 使得数据均匀分布在0点附近, 如图3中由A到B这个过程, 称作平均值中心化 (mean-centering)。通过这个过程可提高主成分分析模型的解释能力。经由缩放与平均值中心化的处理, 可由平均值、方差大小十分离散的变量集, 变为平均值和变量大小整齐的变量集 (图 3D), 以提高数据处理质量。在选择一些特殊模式进行数据分析时, 可能需要对应的数据权重手段, 如进行 S-plot 分析, 一般可以选择对数据进行 Pareto 方式处理。

5 偏最小二乘关联分析与偏最小二乘判别分析

偏最小二乘关联分析, 全称为基于偏最小二乘投影的潜在结构关联分析 (partial least squares projection to latent structure, PLS) 是代谢组学数据处理过程中常用的方法之一。PLS 作为一种有监督的学习方法, 是建立在 X (自变量) 与 Y (因变量) 矩阵基础上的双线性模型, 可以看作是由外部关系 (即独立的 X 矩阵和 Y 矩阵) 和内部关系 (即两矩阵间的联系) 构成。与PCA分析的原理相同, PLS利用偏最小二乘法对数据结构进行投影分析[16]。

PLS-DA 是常用的有监督模式多变量统计分析方法。判别分析是通过代谢组学检测到的若干变量值, 判断研究对象如何分类的统计分析方法。在代谢组学数据处理过程中只需要一个数据集 X, 但在分析时必须对样品进行指定并分组, 这样分组后模型将自动加上另外一个隐含的数据集 Y[22], 该数据集变量数等于组别数, 其他计算方法与 PLS 相同。这种模型计算的方法强行把各组分门别类, 有利于发现不同组间的异同点。对于组间差异不够明显的样品, 采用 PCA 方法常常无法区分样品的组间差异, 这种情况下采用 PLS-DA 模型可能更加有效。PLS-DA 模型在需要同时观察多组别样品相似性和差异性时体现更大价值。需要指出的是 PLS-DA 是有监督的学习方法, 在代谢组学数据处理时往往由于主成分过多、分组过于复杂而出现过拟合现象 (over-fitting), 造成模型失真, 在实际数据分析时应注意验证模型有效性和可靠性。

6 隐结构正交投影

2002 年, Trygg 和 Wold[25]在 PLS 算法基础上, 建立了一种新的多变量分析方法, 隐结构正交投影 (orthogonal projection to latent structure, OPLS), 全称为基于偏最小二乘的隐结构正交投影 (partial least squares orthogonal projection to latent structure)。与 PLS 相比, OPLS 的观测变量矩阵 X 中与预测变量矩阵 Y 中无关联的“噪音”变量会被滤除/忽略, 从数学角度理解, 就是除去 X 数据变量中与 Y 变量无关或正交的变异因素。

PLS 的主要目的是最大程度地解释观测值并实现对响应变量的预测, 其目标函数是最大化 X 与 Y 的协方差。通俗的说, PLS 算法将观测值中观测变量 X 的变异以方差的平方和表示, 并将此分成可解释的系统变化 R_2X 和残差 E。对 X/Y 关联进行解释并实现可视化是 PLS 算法的一大优点。但当模型复杂性上升的时候, 通过模型参数对 X/Y 关联进行解释将会变得非常困难。OPLS 算法相比 PLS 算法, 除残差部分 E 之外, 进一步地将观测值中的观测变量 X 的系统变异分解为两个部分, 即与 Y 变异相关的, 可预测 Y 变异的部分 R_2X_{pred}; 与 Y 变异不相关, 即与 Y 变异正交的部分 R_2X_{orth}。在与 PLS 有相同的模型拟合和预测能力的基础上, OPLS 将预测成分 (predictive OPLS component) 和正交成分 (orthogonal OPLS component) 进行区分, 使模型有了更好的可解释性 (图 4)[26]。

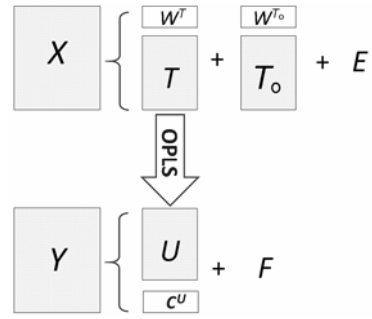

Figure 4 The arithmetic structure of orthogonal projection to latent structure (OPLS)

OPLS 的简化数学模型如下:

$$X = TW^T + T_oW^{T_o} + E$$

$$Y = UC^U + F$$

其中 X 是 $N\times K$ 的观测因子矩阵, Y 是 $N\times M$ 的预

测因子矩阵; T 和 T_o 分别为 X 的预测成分得分矩阵和正交成分得分矩阵, W^T 和 W^{T_o} 分别是 X 的预测成分载荷矩阵和正交成分载荷矩阵。U 为 Y 数据得分矩阵, C^U 是 Y 的预测成分权重矩阵; E 和 F 分别是 X 和 Y 数据残差 (噪音), 上述算法均假设检测变量呈独立的随机正态分布。由于增设了 X 正交方向、与 Y 不相关部分函数, 相当于消除了与 Y 变量变化趋势无关的数据, 所以模型预测能力和有效性可能得到较大提高。

7　隐结构双向正交投影

PLS 与 OPLS 都是用于解决两个独立数据集 (X−Y) 潜在关联分析方法, 即基于 X 变量数据信息, 建立 Y 变量预测模型 (X→Y)。在 OPLS 算法/方法提出不久, Trygg 等[27]又提出了隐结构双向正交投影 (orthogonal bidirectional projections to latent structures, O2PLS)。这一算法的目的是建立两个独立数据集 (X−Y) 的整合模型, 明确 X 与 Y 中共同对应变化, 以及 X 与 Y 中正交于彼此 (即互不相关) 的变化, 以实现两个变量的双向预测, 即 X↔Y。目前, O2PLS 逐渐受到重视, 并在代谢组学、流程监控、定量构效关系研究中应用越来越广泛。

O2PLS 模型的核心矩阵结构如图 5 所示。X 与 Y 是所研究的两个数据集, 两者共同的变化以 X 得分 T, Y 得分 U, X 载荷 W^T 和 Y 载荷 C^U 表示。X 特有的变化, 即正交成分以得分 T_o 和载荷 W^{T_o} 表示, 以此类推 Y 的正交成分得分 U_o 和载荷 C^{U_o}。正交成分 (即矩阵 T 与 U) 的分析除 OPLS 算法, 也可同时用主成分分析 (PCA) 等算法建立相关。非系统性变化部分则以残差 E 和 F 表示。

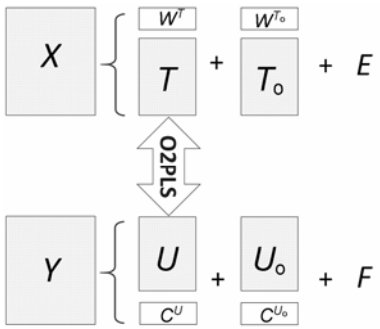

Figure 5　The arithmetic structure of orthogonal bidirectional projections to latent structures (O2PLS)

O2PLS 的简化数学模型表示如下:

$$X = TW^T + T_oW^{T_o} + E$$

$$Y = UC^U + U_oC^{U_o} + F$$

与 OPLS 模型相同, X、Y 分别为观测因子矩阵和预测因子矩阵; 其中, TW^T 和 UC^U 为 X、Y 数据矩阵主成分模型信息。T 与 U 为 X、Y 数据集的得分向量矩阵, W^T 和 C^T 为 X、Y 数据集载荷向量矩阵。T_o 与 U_o 为 X、Y 数据集正交方向得分向量矩阵, W^{T_o} 和 C^{U_o} 为 X、Y 数据集正交方向载荷向量矩阵。E、F 涵义与 OPLS 模型一致。由于同时增设了 X 和 Y 正交方向的函数, 可同时消除 X 与 Y 变量变化趋势无关的数据, 所以模型预测能力和有效性可以得到进一步提高。

8　S 形图

OPLS 可对两组/类样品的代谢组学数据进行判别分析, 并清晰展现组间样品差别。S 形图 (S-plot) 可将观测变量清晰展示在二维平面图上, 当数据经过 Pareto 或 Ctr 方式校正后, 其形状非常像英文字母 S。特别是位于“S”两翼的观测变量是两组间差异最为显著的化合物, 是具有高研究价值的潜在标志物。

S 形图横坐标和纵坐标分别为多变数数据模型所预测主成分对应的 p1 和 p(corr)1, 分别代表了测定变量在区分模型中差异程度 (协方差大小, modeled covariation) 和可靠程度 (相关性, modeled correlation)[28]。一般来说, 测定变量丰度/数值越大, 其分析结果的可靠性越强, 因为当测定变量丰度/数值较小, 接近噪音水平时, 那么对相关性/可靠程度判断出现错误的风险也越大。所以, 代谢组学研究中, 仅基于相关性选择测定变量作为重点研究的差异变量, 会导致一部分丰度/数值小的变量被选中, 出现假阳性 (I 类错误) 结果, 对应于代谢组学研究中, 就是测定化合物仪器响应较弱的化合物。为减少误判, S 形图对潜在标志物的筛选结合了差异程度和可靠程度两个方面的信息。以图 6a 为例, 在某项疾病研究中, 采用 S-plot 作图方法筛选出差异程度和可靠程度较高的蓝色亮点标示化合物, 提示为潜在生物标志物。其中, 横轴方向上协方差 (p) 接近于 0 的变量 (绿色亮点), 代表置信区间分析不足以支持其作为潜在标志物; 而纵轴方向上可靠程度或相关性 p(corr) 接近 0 的变量 (绿色亮点), 代表在判别分析中以此为据的可靠性较低。因此, 一般选取 S-plot 中位于 S 形两端的变量作进一步分析。

9　共享与特有化合物结构分析

代谢组学研究中, 通常需要研究或比较 2 个 OPLS 模型之间所共同的或独特的差异内源性分子/化合物, 以了解药物干预、基因调控等外因对模型中测定变量/内源性分子的调控作用, 进而研究其作用

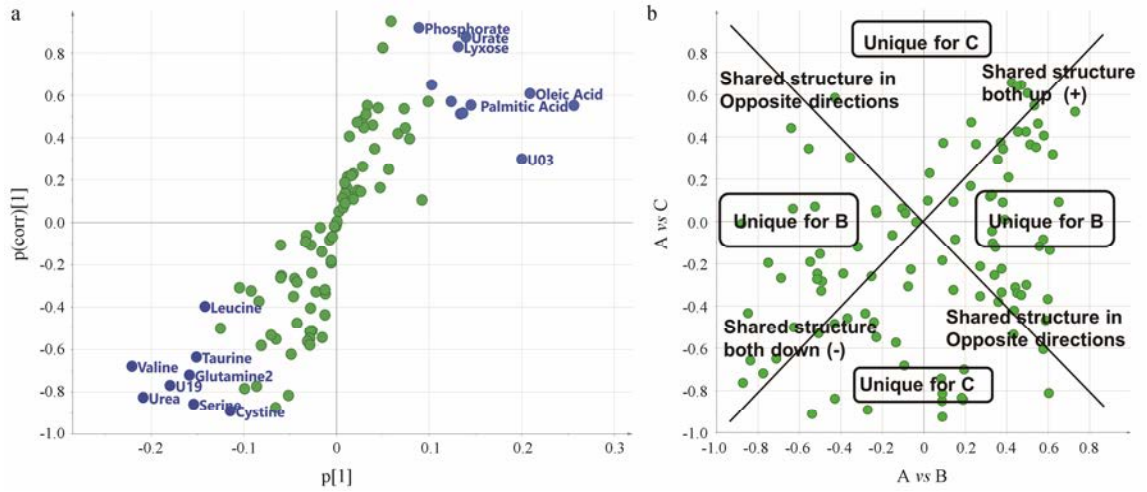

Figure 6 S-plot (a) and shared and unique structure-plot (SUS-plot) (b) visualize and typical molecules in OPLS models

的代谢通路、潜在标志物乃至作用机制。共享与特有化合物结构分析图 (shared and unique structure-plot, SUS-plot) 是实现上述目标、发现相关分子/化合物的很好工具，可以将 2 个 OPLS 模型进行对比分析，研究其中化合物在 2 个模型中的变化趋势的异同性[28]。在两个模型中变化趋势一致或者相反的化合物、2 个模型所特有的化合物分别出现在平面坐标轴不同位置/象限的特定区域，便于发现、归类和分析。

图 6b 显示了 2 个 OPLS 模型中检测变量分布散点图，并包涵了独特信息。横坐标和纵坐标分别为 2 个模型对应的 p(corr)[1]矢量。如果 2 个 OPLS 模型 ***X*** 观测变量与 ***Y*** 分组变量的相似性较高，那么检测变量在图中就趋向于分布在从左下角到右上角 45° 线附近。相反，如果检测变量未分布于此条直线附近，偏离较远，说明这些检测变量在 2 个模型中的差异程度大。极端情况下，检测变量分布在从右下角到左上角 135°直线上，说明在 2 个模型中变化趋势相反。

在 2 个模型的 SUS-plot 图中，检测变量出现在图中的不同区域，代表着变量在 2 个模型中的特异性。以组 A 为参比，对模型 1 (A *vs* B) 和模型 2 (A *vs* C) 这两个 OPLS 模型进行 SUS-plot 对比分析，标注“unique for B”和“unique for C”的区域，分别为对 A 与 B 的判别分析及 A 与 C 的判别分析中具有唯一性、独特性的变量；而 45°斜线附近标注“shared structure both up”和“shared structure both down”位置的变量在两个模型中的变化趋势相同，只是分别为上调和下调趋势。这些变量对两个判别分析模型中都有贡献。对应地，135°斜线附近，标注“shared structure in opposite directions”标记位置的变量，同样对 2 个判别模型中有贡献，只是在 2 个模型中的变化趋势正好相反。

10 模型验证

基于代谢组学矩阵数据所建立数学模型进行评价，通常需要对模型的有效性和可靠性进行验证 (validation of the model)。常规模型验证可以包括如下几个方面。

异常样本剔除是经过多变数数据分析 (通常是主成分分析或 DModx 分析) 所发现的明显不同于大部分样品的离群样品 (通常是 PCA 分析位于椭圆形范围之外)，这些样品在进行主成分分析时一般需要首先加以甄别或剔除[22]。从严格意义上讲，异常样本分析不属于模型验证范畴，但考虑到如果不剔除异常样本会严重影响模型的有效性和可靠性，模型验证也失去其重要意义，因此，阐述异常样本分析在模型验证中的重要性并不为过。当然，并不是每个异常样本都毫无意义，在临床上，分析特殊的异常样本，可能在研究课题和内容之外发现一些特别个体因素，对临床和研究具有重要意义。发现异常样本的方法一般有主成分分析和 DModx 分析 2 种方式[16]，有时利用 *T* 检验 (如 Hotel T’s test) 也可以起到与 DModx 相似的效果。

交叉验证 (cross-validation) 也叫内部验证 (internal validation)，根据分析数据需要，在选择对应模型后，首先选择一个主成分，再逐个增加到 2 个及多个主成分进行模型拟合，观察模型对变量解释度 R^2 以及预测度 Q^2 值变化。根据 SIMCA p 软件对参数的常规设计，一般可以将数据集分为 2 块，设定数据矩阵/观察变量中 6/7 数据用于建立模型，对剩余 1/7 数据进行预测，并且这个过程进行多次重复，最终取得 R^2、Q^2 值。R^2、Q^2 值越接近 1，说明模型可靠性越强，

反之，越接近 0，模型越不可靠。R^2、Q^2 值需要综合考察。对于组别差异明显、有效性和可靠性高的模型，R^2、Q^2 数值会比较接近 1。但一般的模型，随着主成分数目增加，通常情况下 R^2 值不断提升并接近 1，但 Q^2 值只能到达一定程度，然后数值下降。在确定有效主成分数量时，需注意增加主成分数量 Q^2 值是否有较明显增加，如果 Q^2 值停止增加或增加微弱，那么主成分数目就不能再增加。一般认为 Q^2 大于 0.5 说明模型具有较好的可靠性和预测度，大于 0.9 非常优秀，且 R^2、Q^2 值差距不大于 0.2～0.3 为好。因为随着主成分数目增加，模型引入的偏差也增大，如果引入的偏差超过增加主成分后新模型可靠性 (预测度) 增加幅度，那么增加主成分就没有实质性意义，一般来说超过 4 或 5 个主成分的模型需要慎重。

置换检验 (permutation test)：交叉验证可以评价模型的预测能力，但并没有对上述预测能力进行统计检验，验证其合理性和可靠性。在采用 PLS、PLS-DA、OPLS、OPLS-DA 等需要进行数据关联或者判别类的模型进行分析时，可以采用置换检验模式，依据已知测定数据变量 X 对预测变量 Y 进行多次迭代分析，并得出一个对这些变量的统计结果。通过考察所有样品对应 R^2、Q^2 计算值所组成的拟合直线在 Y 坐标轴的截距，表示模型的可靠性和过拟合程度。通常 R^2 截距值应明显小于模型变量解释度，并小于 0.3 (最好小于 0.2，越接近 0 越好)，Q^2 截距值应明显小于模型变量预测度，并小于 0.05。如果 R^2、Q^2 (特别是 R^2) 拟合直线在 Y 轴上的截距接近对应主成分下 R^2、Q^2 值 (直线右侧高点)，那么提示存在过拟合可能。除了模型自身可靠性不佳外，在过多增加无效主成分数目情况下，过度拟合常常会出现。

外部验证 (external validation)：交叉验证采用一组已知的、现有数据评价该模型的预测能力，属于内部验证。但实际应用中，常常需要利用一组已知数据建立的模型 (训练数据，training set)，对另外一组不相关的测试数据 (test set) 进行预测，即“外部验证”。为了获得更高的验证结果，通常情况下，需要训练数据量足够大，具有足够的代表性和可靠性，这样对测试数据才能具有良好的预测性。需要注意的是，外部验证中测试数据需要有明确的已知信息，如属于何种类别，这样进行验证时才能有正确与否的标准。如果没有已知信息，那么就变成了纯粹的预测，对临床疾病分析来说，就成了预测性诊断。

11 总结

主成分分析方法是代谢组学数据分析的常用方法，由于其中的理论计算复杂、构建模型抽象、数据转换多变，专业书籍中相关内容较难理解和直接融入实际数据分析，因此，本文从主成分分析最常用的十个方面入手，以简洁、易懂的语言介绍了代谢组学数据处理常用方法中重点和难点问题，方便研究人员合理选择数据处理模式、规范数据处理程序、熟练解析数据处理结果，并得出可靠结论。

References

[1] Nicholson JK, Lindon JC, Holmes E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli *via* multivariate statistical analysis of biological NMR spectroscopic data [J]. Xenobiotica, 1999, 29: 1181−1189.

[2] Tang HR, Wang YL. Metabonomics: a revolution in progress [J]. Prog Biochem Biophys (生物化学与生物物理进展), 2006, 33: 401−417.

[3] Jia W. Medical Metabonomics (医学代谢组学) [M]. Shanghai: Shanghai Scientific & Technical Publisher, 2011.

[4] Li B, He X, Jia W, et al. Novel applications of metabolomics in personalized medicine: a mini-review [J]. Molecules, 2017, 22: 1173.

[5] Shi J, Cao B, Wang XW, et al. Metabolomics and its application to the evaluation of the efficacy and toxicity of traditional Chinese herb medicines [J]. J Chromatogr B, 2016, 1026: 204−216.

[6] Alden N, Krishnan S, Porokhin V, et al. Biologically consistent annotation of metabolomics data [J]. Anal Chem, 2017, 89: 13097−13104.

[7] Xu GW. Metabolomics: Methods and Applications (代谢组学——方法与应用) [M]. Beijing: Science Press, 2008.

[8] Zhang AH, Sun H, Wang P, et al. Modern analytical techniques in metabolomics analysis [J]. Analyst, 2012, 137: 293−300.

[9] Johnson CH, Ivanisevic J, Siuzdak G. Metabolomics: beyond biomarkers and towards mechanisms [J]. Nat Rev Mol Cell Biol, 2016, 17: 451−459.

[10] Sevin DC, Kuehne A, Zamboni N, et al. Biological insights through nontargeted metabolomics [J]. Curr Opin Biotechnol, 2015, 34: 1−8.

[11] De Raad M, Fischer CR, Northen TR. High-throughput platforms for metabolomics [J]. Curr Opin Chem Biol, 2016, 30: 7−13.

[12] Yin PY, Xu GW. Current state-of-the-art of nontargeted metabolomics based on liquid chromatography-mass spectrometry with special emphasis in clinical applications [J]. J

Chromatogr A, 2014, 1374: 1−13.

[13] Wang Y, Gu H, Lu X, et al. Development of hydrophilic interaction chromatographic hyphenated techniques and their applications [J]. Chin J Chromatogr (色谱), 2008, 26: 649−657.

[14] Tang HR, Xiao CN, Wang YL. Important roles of the hyphenated HPLC-DAD-MS-SPE-NMR technique in metabonomics [J]. Magn Reson Chem, 2009, 47: S157−S162.

[15] Shi XZ, Qiao LZ, Xu GW. Recent development of ionic liquid stationary phases for liquid chromatography [J]. J Chromatogr A, 2015, 1420: 1−15.

[16] Aa JY. Analysis of metabolomic data:principal component analysis [J]. Chin J Clin Pharmacol Ther (中国临床药理学与治疗学), 2010, 15: 481−489.

[17] Cambiaghi A, Ferrario M, Masseroli M. Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration [J]. Brief Bioinform, 2016, 18: 498−510.

[18] Tautenhahn R, Patti GJ, Rinehart D, et al. XCMS online: a web-based platform to process untargeted metabolomic data [J]. Anal Chem, 2012, 84: 5035−5039.

[19] Axelson DE. Data Preprocessing for Chemometric and Metabonomic Analysis [M]. CreateSpace Independent Publishing, 2012.

[20] Fiehn O, Kind T, Barupal DK. Data processing, metabolomic databases and pathway analysis [M] // Annual Plant Reviews: Biology of Plant Metabolomics. Vol 43. West Sussex, UK: Wiley-Blackwell, 2011.

[21] Smilde AK, Westerhuis JA, Hoefsloot HCJ, et al. Dynamic metabolomic data analysis: a tutorial review [J]. Metabolomics, 2010, 6: 3.

[22] Eriksson L, Johansson E, Kettaneh-Wold N, et al. Multi- and Megavariate Data Analysis: Principles and Applications [M]. Umeå: Umetrics Academy, 2001.

[23] Aa JY, Wang GJ, Hao HP, et al. Differential regulations of blood pressure and perturbed metabolism by total ginsenosides and conventional antihypertensive agents in spontaneously hypertensive rats [J]. Acta Pharmacol Sin, 2010, 31: 930−937.

[24] Liu LS, Aa JY, Wang GJ, et al. Differences in metabolite profile between blood plasma and serum [J]. Anal Biochem, 2010, 406: 105−112.

[25] Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS) [J]. J Chemom, 2002, 16: 119−128.

[26] Westerhuis JA, Van VEJJ, Hoefsloot HCJ, et al. Multivariate paired data analysis: multilevel PLSDA *versus* OPLSDA [J]. Metabolomics, 2010, 6: 119−128.

[27] Trygg J, Wold S. O2-PLS, a two-block (X−Y) latent variable regression (LVR) method with an integral OSC filter [J]. J Chemom, 2003, 17: 53−64.

[28] Wiklund S, Johansson E, Sjostrom L, et al. Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models [J]. Anal Chem, 2008, 80: 115−122.