化学计量学中的主成分分析*

刘广军①, 高洪涛②

(第一作者:男,43 岁,副教授;①济宁师范专科学校人事处;②济宁师范专科学校化学系,272025,山东省济宁市)

摘要:主成分分析是对多变量数据进行降维处理的一种线性投影方法,它在尽可能保留原有信息的基础上将高维空间中的样本映射到较低维的主成分空间中,使数据矩阵简化,降低维数,寻找少数几个由原始变量线性组合的主成分(也称潜变量),以揭示数据结构特征,提取化学信息. 主成分分析是化学计量学中的基础方法,广泛用于化学实验数据的统计分析,进行数据降维、变量提取与压缩、确定化学组分数、分类和聚类,以及与其他方法联用进行数据处理.

关键词:主成分分析;化学计量学;投影;数据处理

中图分类号:O6 - 04 **文献标识码**:A **文章编号**:1001-5337(2004)03-0075-04

分析化学正经历着巨大的变革,人们正逐渐认识到分析化学是通过化学量测获得化学信息的科学. 分析化学的发展正使其成为一门创造新概念、新原理的学科,随着数学、物理、计算机和新仪器走进分析化学和学科间的不断交叉融合,产生了新的化学分支学科——化学计量学. 化学计量学是一门交叉学科,它应用数学、统计学和计算机科学的工具和手段及其最新成果来设计和选择最优化化学量测方法,并通过解析化学量测数据以最大限度地获取化学及其相关信息,这使得化学计量学方法越来越得到广大化学工作者的重视.

化学计量学研究有关化学量测的基础理论和方法,分析信号的多元分辨和校正在化学计量学中是非常活跃的一个领域. 化学计量学提供了很多方法来进行多元分辨和校正,常用的有主成分分析(PCA)、偏最小二乘法(PLS)、迭代目标转换因子分析(ITTFA)、渐进因子分析(EFA)、窗口因子分析(WFA)、秩消失因子分析(RAFA)、广义秩消失因子分子(GRAFA)、投影旋转因子分析(PRFA)、直观推导式演进特征投影法(HELP)以及广泛使用的正交投影分辨(OPR)和正交信号校正(OSC)以及残差双线性分解(RBL)等等,在这些方法中,主成分分析(PCA)是多元信号分辨与校正中常用方法,是其他化学计量学方法的基础[1,2].

自 PCA 被引入分析化学计量学以来,已有很多关于 PCA 的文献发表,仅 1996 年至今以来,在网络期刊库 Elsevier Science (http://elsevier.lib.tsinghua.edu.cn) 化学类刊物上有关 PCA 的文章共有 477 篇,其中《Chemometrics and Intelligent Laboratory Systems》上 92 篇,《Analytica Chimica Acta》(ACA) 上有 81 篇,在网络期刊 John Wiley (http://www3.interscience.wiley.com) 上也有 261 篇文章发表. 可以说,PCA 是一种经典的又不断萌生新意的化学计量学方法,被认为是化学计量学方法的基础,只要能够深刻理解 PCA 的实质,能够正确运用 PCA,对于我们理解和运用化学计量学方法解决实际问题以及发展化学计量学新方法都有很大帮助.

主成分分析是对多变量数据进行统计处理的一种数据线性投影方法,它在尽可能保留原有信息的基础上将高维空间中的样本映射到较低维的主成分空间中. 其基本思路是以一种最优化方法浓缩量测数据(用 Y 表示)信息,使数据矩阵简化,降低维数,寻找少数几个由原始变量线性组合的主成分,以揭示数据 Y 结构特征,提取基本信息. 主成分分析主要用于(1)降维(或称数据压缩),寻找几个主成分(也称潜变量)在低维空间表示高维数据;(2)数据的可视化和分类聚类,主成分的投影显示法即可用于分类判别又可用于聚类,可以从投影图中看出样本与样本之间的关系,变量和变量之间的关系;(3)降低随机误差,主成分分析的过程是寻找少数几个相互正交,方差最大的新变量,来重新构造数据,能够有效去除抽出误差;(4)确定化学组分数,从数学意

* 收稿日期:2003 - 09 - 11

义上主成分分析的实质是特征值问题,主成分分析所得到的非零特征值的个数就是矩阵的秩,从化学意义上就是构成数据的化学组分数,确定了矩阵的秩就可以确定体系的组分数;

1 主成分分析原理

主成分计算方法有非线性偏最小二乘(NIPALS)、乘幂法(POWER)、奇异值分解(SVD)和特征值分解(EVD)等等[3]. 若量测数据矩阵记为 X,NIPALS和 SVD 的处理对象是未知量测数据矩阵 X,POWER 和 EVD 的对象是其协方差阵 X * X′(或 X′ * X),他们的原理基本上是基于特征值问题,计算结果也基本相同. 在化学计量学中一般采用的方法是 NIPALS 和 SVD,这两种算法从本质上讲并无不同之处,NIPALS 算法可帮助我们理解主成分分析计算过程[3],由于目前在化学计量学计算广泛采用 Matlab 语言进行科学计算,在 Matlab 中 SVD 方法非常简单,用一句话"[U,S,V] = SVD(X)",就可得到所需结果,故 SVD 方法在化学计量学中得到广泛的应用.

1.1 奇异值分解(SVD)[1,3,4]

SVD 的详细算法并不重要,重要的是 SVD 在数据矩阵表示中的基本原理,它对于理解主成分、因子以及特征值分析之间的关系有所帮助.

SVD 可将可将任意阶实数矩阵分解为 3 个矩阵的乘积,即 X = U * S * V′,其中 S 为对角矩阵,各元素为奇异值或特征值的平方根,对于对称矩阵,奇异值与特征值的平方根是相同的,它收集了 X 矩阵的特征值,U 和 V 分别为标准列正交矩阵和标准行矩阵,收集了这些特征值所对应的列特征矢量和行特征矢量.

设 T = U * S, P = V,则有

$$X = T * P' \tag{1}$$

式(1)是 PCA 的最一般公式,其中 T 称为得分矩阵,P 称为载荷矩阵,原始数据阵可以看成得分矩阵和载荷矩阵两个矩阵的乘积. 这在化学量测上有明确的意义和解释,对于联用仪器(如 HPLC - DAD, GC - MS, LC - MS 等等)能提供两维数据,对于 HPLC - DAD 来说,一维为色谱,另一维为光谱,对于 GC - MS 来说,一维为色谱,另一维为质谱. 对于 OV 型数据,一维为样本,另一维为变量,得分矩阵反映了样本与样本之间的关系,而载荷矩阵反映了变量(如波长、质荷比)之间的关系.

1.2 主成分分析数学和几何意义(投影和方差)

由公式(1)可得 T = X * P,可见得分矩阵 T 的每一个元素实际是每一个样本向量 y_i (i = 1,2, …, n)对荷载矩阵 P 中的每一相互正交的荷载矢量上的投影坐标(内积本质上就是投影),它反映了样本与样本之间的相互关系; 同理可得,荷载矩阵的每一个元素实际是每一个变量向量 y_j (j = 1,2, …, d)对得分矩阵中的每一相互正交的得分矢量上的投影坐标,它反映了变量与变量之间的相互关系. 主成分分析的实质是投影,是高维数据在低维空间的表示. PCA 是最常采用的线性投影方法,通过 PCA 高维数据投影在几个正交的主成分空间上,主成分是原始变量的线性组合,主成分的选择遵循方差最大的原则[1,3~7].

从几何意义上来讲,主成分分析通过对数据重构,通过线性转化(投影),得到互不相关的相互正交的新坐标轴来表示变量,得到新坐标轴的依据是各点到坐标轴的距离最小. 例如测量 2 个变量,20 个样品得到的数据如图 1 所示.

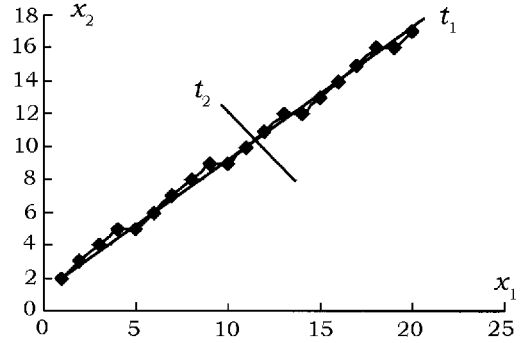

图 1 主成分分析的数学几何表示

主成分的目的就是重新构建坐标轴 t_1 和 t_2, t_1 和 t_2 相互正交互不相关,是原始变量 x_1 和 x_2 的线性组合,用 t_1 和 t_2 来表示数据,使得数据点到 t_1 的距离最小,到 t_2 的距离最大,可以用下式来表示

$$t_1 = c_{11}x_1 + c_{12}x_2,$$

$$t_2 = c_{21}x_1 + c_{22}x_2,$$

其中,第一个主成分轴 t_1 为方差最大的方向,第二主成分轴 t_2 为方差次大的方向.

主成分分析得到互不相关的新变量 t_1 和 t_2(也称潜变量),主成分的确定基于最大方差判据,每一主成分相继描述前一个主成分没能建模的最大方差. 因此,数据的方差大部分包含于第一个主成分中,第二个主成分比第三个主成分所包含的信息多,等等. 最后,根据所有主成分所包含的方差百分数,

可计算所有主成分所包含的方差百分数,可计算出所有需要的主成分.

2　应　用

2.1　数据降维(压缩和变量选择)

主成分分析作为一种投影方法,可以在互不相关相互正交的新坐标轴构成的低维空间可以表示高维数据,进行数据压缩;通过寻找新的,相对于原变量来说数目少得多的潜变量来表示原数据,大大降低了变量的维数,由于主成分分析的方差最大原则,主成分能基本代表数据的结构,换句话说,可以通过少数的新变量重构数据,而并不损失原来的基本数据信息. 在数据压缩、特征提取和变量选择方面得到了广泛的应用[5,8,9].

2.2　分类和聚类

主成分的投影显示法即可用于分类判别又可用于聚类,可以从投影图中看出样本与样本之间的关系,变量和变量之间的关系,在化学模式识别中用于分类、聚类[3,5,8,9]. 在药物生产和质量控制中也有用武之地[10].

2.3　数据重构,去除抽出误差

通过主成分分解,采用主成分重构数据,能够有效地去除化学量测中的随机误差,我们模拟了随机误差的情况,经主成分分解,进行数据重构后,随机误差的影响得到有效去除. 值得注意的是,与量测数据正交无关的误差仅采用主成分分析是无法去除的,梁逸曾将量测误差分为置入误差和抽出误差,将主成分分析能够去除的误差称为抽出误差[1]. 陈宗海等提出了小波平滑主成分分析的方法,能有效滤除量测数据中的随机噪声,提高了信噪比[11].

2.4　确定化学组分数

由于矩阵的秩代表数据矩阵中线性无关的向量的个数,即等于非零的特征值的个数. 如果不存在量测误差,则量测数据阵的秩就等于体系中存在的独立组分数. 在实际过程中,往往存在量测误差,因而,非零特征值的个数比较多,即不能通过其判断化学组分数. 根据主成分的方差最大原则,第一、第二主成分的方差依次大于第二、第三主成分的方差,特征值越大,方差也就越大,前几个特征值的方差达到总方差的一定的比率,就可以用特征值的数目来确定组分数,此时组分数能代表体系的实际组成. 应该注意的是,主成分分析判断组分数的依据是基于特征值和特征向量的,一般情况下,问题不大;但在有些情况下,在实际化学量测中某种组分的特征值并不大,即该组分的方差对总方差的贡献并不大,这时,特征值判据就存在一定的问题,要采取其他的方法进行确定[12,13].

3　主成分分析应注意的几个问题

3.1　数据预处理[1,3]

实际化学量测给出的数据不仅量纲不同,值的大小也有很大的差异,若将原始数据直接用来做主成分分析会有很大的问题,必须对数据进行预处理. 数据预处理的目的是将原始数据转换为更便于处理的数值,通过对原始数据乘、除、加或减一个常数,得到代码数据.

3.1.1　均值中心化　消除数据的常数偏移量,对坐标原点做变换,利用公式(2)从每个变量中减去该列的平均值.

$$x_{ik}^{*} = x_{ik} - \frac{1}{N}\sum_{i=1}^{n} x_{ik}. \tag{2}$$

3.1.2　数据标准化　由于变量(或特征)表示了样品的不同性质,在多数情况下,这些性质差别较大,即列于列之间的数值可能存在很大差别,这就意味着不同变量具有不同的绝对值和变量范围(方差). 这些差异可以通过比例调整进行消除,将他们按比例调整到相似的范围和方差. 两种比例调整方法是:按数值范围进行调整的值域调整法和按标准偏差进行调整的自动调整法.

(1)　值域调整法(range scaling):

$$x_{ik}^{*} = \frac{x_{ik} - x_k(\min)}{x_k(\max) - x_k(\min)},\ 0 \le x_{ik}^{*} \le 1. \tag{3}$$

(2)　自动调整法(autoscaling)

$$x_{ik}^{*} = \frac{x_{ik} - \bar{x}_k}{s_k}. \tag{4}$$

其中 s 为标准偏差.

$$s_k = \sqrt{\frac{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}{n-1}}. \tag{5}$$

(3)　归一化:根据公式(6)将一个数据向量的模长规范化到 1.

$$x_{ik}^{*} = \frac{x_{ik}}{\mathrm{norm}(x_k)}. \tag{6}$$

有时也用各元素除以极大值来进行归一化.

3.2　组分数的确定

主成分分析是基于特征值问题的解,主成分的

判定是基于方差最大原则,如 RE 法,IND 法和 Fisher 方差比较法,可参看文献[1]. 在一般情况下,根据误差理论判断主成分数,结果是正确的. 但在实际量测中,某些变量的特征值的方差比较小,但也是重要的组分,这时人们通常采用交叉验证法(Cross-Validation)来进行组分数的确定,结果是比较可靠的[1,3,11~13]. 所以,组分数的确定并不能完全依靠主成分来判断.

主成分分析在化学计量学中得到了非常广泛的应用,在渐进因子分析(EFA)中,利用了整个矩阵的特征值的全局因子分析(Global PCA)来判断色谱的流入流出信息,这种方法通常适用于一些组分的谱线重叠不严重而且其相对浓度相差不大的混合物体系,计算比较慢,对于重叠严重并且组分相对浓度相差很大的混合物体系一般采用局部主成分分析方法,而且计算速度比全局主成分分析计算要快很多. 如窗口因子分析(WFA)、固定尺寸移动窗口渐进因子分析(FSMWFA)和直观推导式演进特征投影法(HELP)主要采用局部主成分分析,利用实验数据的局部秩变化进行体系的分辨分析.

PCA 算法也有广泛的发展,如快速 PCA,稳健 PCA、柔性 PCA、非线性 PCA 以及应用于高维空间的 NPCA,读者可参看文献[15~19],本文不再赘述.

主成分分析是在分析化学计量学中应用广泛的线性投影方法,在不损失样品基本信息的前提下,它通过寻找相互正交互不相关的主成分来表示数据矩阵,揭示数据特征,提取化学信息,可用于化学实验数据的降维、压缩、变量的选择以及去相关,与其他方法联用进行数据处理. PCA 是化学计量学的基础方法,既经典又不断萌发新意,对 PCA 的正确运用和深入理解对于解决实际问题和发展分析化学计量学新理论和新方法都大有裨益.

参考文献:

[1] 梁逸曾. 白灰黑复杂多组分分析体系及其化学计量学算法[M]. 长沙:湖南科学技术出版社,1996. 12.

[2] Barry K L. Chemometrics[J]. Anal Chem, 2000, 72(91R~97R).

[3] Wu W, Massart D L, Sde J. The kernel PCA algorithms for wide data. Part 1: theory and algorithms[J]. Chemom Intel Lab Syst,1997, 36:165~172.

[4] 邵学广,蔡文生. 化学计量学:统计学与计算机在分析化学中的应用[M]. 徐筱杰译. 北京:科学出版社,2003.

[5] Guo Q, Wub W, Massart D L, et al. Feature selection in principal component analysis of analytical data[J]. Chemom. Intel Lab Syst,2002, 61:123~132.

[6] Steven Z, Fairchild, John H K. PCR eigenvector selection based on correlation relative standard deviations [J]. J Chemometrics, 2001, 15: 615~625.

[7] Daszykowski M, Walczak B, Massart D L. Projection methods in chemistry [J]. Chemom Intel Lab Syst, 2003, 65:97~112.

[8] Helena I B, Edlund P O, Olav M, et al. Screening of Biomarkers in Rat Urine Using LC/ Electrospray Ionization-MS and Two-Way Data Analysis[J]. Anal Chem, 2003, 75: 4784~4792.

[9] Barros A S, Rutledge D N. Genetic algorithm applied to the selection of principal components [J]. Chemom Intel Lab Syst, 1998, 40:65~81.

[10] Max Andre. Multivariate Analysis and Classification of the Chemical Quality of 7-Aminocephalosporanic Acid Using Near-Infrared Reflectance Spectroscopy [J]. Anal Chem, 2003, 75: 3460~3467.

[11] 陈宗海,林祥钦,邵学广. 基于小波变换平滑主成分分析[J]. 分析化学, 2000, 28(8):925~929.

[12] 沈海林,李晓宁,梁逸曾. 化学体系组分数确定的新方法:子空间比较法[J]. 科学通报,2000, 45(6):587~592.

[13] 苏越,郭寅龙. 偏最小二乘法中主成分数确定的新方法[J]. 计算机与应用化学,2001, 18(3):237~240.

[14] Xu Q S, Liang Y Z. Monte Carlo cross validation[J]. Chemom. Intel Lab Syst, 2001, 56:1~11.

[15] Statheropoulos M, Pappa A, Karamertzanis P, et al. Noise reduction of fast, repetitive GC/MS measurements using principal component analysis (PCA) [J]. Anal Chim Acta, 1999, 401: 35~43.

[16] Hubert M, Rousseeuw P J, Verboven S. A fast method for robust principal components with applications to chemometrics [J]. Chemom Intel Lab Syst, 2002, 60: 101~111.

[17] Pravdova V, Boucon C, Walczak B, et al. Three-way principal component analysis applied to food analysis: an example[J]. Anal Chim Acta, 2002, 462: 133~148.

[18] Tortajada-Genaro L A, Camp ńs-Falc ó P, Verd ú Andr é s J, et al. Multivariate versus univariate calibration for nonlinear chemiluminescence data: Application to chromium determination by luminol-hydrogen peroxide reaction [J]. Anal Chim Acta, 2001, 450:155~173.

[19] Cserh ái T, Forg ács E. Use of canonical correlation analysis for the evaluation of chromatographic retention data[J]. Chemom Intel Lab Syst, 1995, 28: 305~313.

(下转第 81 页)

CONTENT DETERMINATION OF GASTRDIN IN DIFFERENTLY PROCESSED TIANMA BY RP-HPLC

YUAN Xiao, *YUAN Ping*, *YOU Min*, *SI Ying*, *WANG You-wei*

(Wuhan Institute of Botany, CAS, 430074, Wuhan, Hubei, PRC)

Abstract: Three processing methods of fresh Tian Ma (Rhizoma Gastrodiae) from traditional Chinese medicinal plant Gastrodia elata Blume were investigated and evaluated. Methods: To Mensurate and analyze the contents of gastrodin in Tian Ma samples processed by three methods of steaming on boiling water, roasting and heating at 125 ℃ and 60 ℃, RP-HPLC was used. Conclusion: The contents of gastrodin in processed Tian Ma were 0.1316 %, 0.1313 % and 0.0270 %, determined by RP-HPLC, respectively. Experimental results: The processing method of steaming on boiling water was mostly suitable for merchandised Tian Ma, which owned a good appearance, roasting at 125 ℃ specially for a large scale of pharmaceutical usage of Tian Ma powder, the processing method of heating at 60 ℃ was not a good option for processing fresh Tian Ma. Additionally, instantly processing in any processing method was an important requirement of keeping the content of gastrodin in Tian Ma stabilizing.

Key words: Gastrodia elata blume; Gastrodin; processing method; RP-HPLC; content

(上接第 78 页)

APPLICATION OF PRINCIPAL COMPONENT ANALYSIS IN CHEMOMETRICS

LIU Guang-jun[①], *GAO Hong-tao*[②]

(①Personnel Bureau Office; ②Department of Chemistry, Jining Teachers' College, 272025, Jining, Shandong, PRC)

Abstract: Principal component analysis (PCA) is the most popular linear projection method in chemometrics. It allows projection of multidimensional data onto few orthogonal features, called principal components (PCs), constructed as linear combination of original variables to maximize description of the data variance. PCA has been applied in data decomposing, feature selecting, classing and clusting. And it is a basic and developing method for chemical experimental data processing.

Key words: PCA; chemometrics; projection; data processing