第44卷第11期 中国科学技术大学学报 Vol. 44, No. 11
2014年11月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Nov. 2014

文章编号:0253-2778(2014)11-0881-06

一种基于残余矩阵方差的主因子数估计方法

陆 玮,邵利民

(中国科学技术大学化学与材料科学学院,安徽合肥 230026)

摘要:提出了一种基于残余矩阵方差的主因子数估计方法(RVR方法),判断标准是扣除不同因子后,残余矩阵方差显现出差异性.列出了RVR对模拟数据和实验数据的处理结果,并且与其他方法进行了比较,结果表明RVR方法对主因子数的估计效果较好.

关键词:主因子数;主成分分析;特征值;特征向量;残余矩阵

中图分类号:O651 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2014.11.001

引用格式: Lu Wei, Shao Limin. Estimation of the number of primary factors based on residual matrix variance ratio[J]. Journal of University of Science and Technology of China, 2014,44(11):881-886.
陆玮,邵利民. 一种基于残余矩阵方差的主因子数估计方法[J]. 中国科学技术大学学报,2014,44(11):881-886.

Estimation of the number of primary factors based on residual matrix variance ratio

LU Wei, SHAO Limin

(*School of Chemistry and Materials Science, University of Science and Technology of China, Hefei* 230026, *China*)

Abstract: A method of estimating the number of primary factors based on residual matrix variance (RVR) was presented, whose criterion is the difference between reduced matrices after deducting different numbers of factors. Meanwhile, the results of the simulated and experimental data obtained by using RVR were presented. A comparison between RVR and several typical methods indicates that RVR has good performance in estimating primary factors.

Key words: number of principal factors; principal component analysis; eigenvalue; eigenvector; reduced matrix

0 引言

仪器联用技术的发展使矩阵形式的二维测量数据在分析化学中渐成主流.这类数据包含了丰富而且复杂的信息,其中信息主要贡献者的数目称为主因子数.根据矩阵数据的主因子数,可以了解复杂样品中的化学组分数.此外,非化学成分也可能向数据矩阵提供信息,例如仪器噪声、仪器脉动以及数据预处理.这些因素被称为"鬼"因子(ghost factors).因此,通过对主因子数的估计,还可以分析仪器的背景信息,评价仪器的性能.

关于主因子数估计方法的研究历史可以追溯到20世纪60年代.经过半个多世纪的发展,各种方法相继出现.这些方法各具特点,但是在处理实际数据时

收稿日期:2014-09-16;**修回日期**:2014-10-21
基金项目:国家自然科学基金(21175123),教育部新世纪优秀人才支持计划(NCET-11-0878)资助.
作者简介:陆玮,女,1989年生,硕士. 研究方向:化学计量学. E-mail: luwei0298@163.com
通讯作者:邵利民,博士/副教授. E-mail: lshao@ustc.edu.cn

均有所欠缺.特别地,如果体系中存在微量组分,噪声水平较高,或者信号重叠严重时,不同估计方法所得到的主因子数差异较大.这种情况下,仅凭一种方法难以对主因子数进行精确地判断.本文提出了一种基于比较残余矩阵方差的方法,以 RVR 方法(residual matrix variance ratio)表示.基本思路是,按照因子的贡献率由大到小的顺序,从原始矩阵中依次扣除因子,得到各残余矩阵的方差,根据残余矩阵方差所表现出的特征差异,来辨别主要因子和次要因子.

1 理论基础

估计主因子数的首要步骤,是对矩阵数据进行主成分分析(principal component analysis, PCA).PCA 有多种数值计算方法,本文采用奇异值分解方法(singular value decomposition, SVD)[1,2].

$$\boldsymbol{D}=\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}} \tag{1}$$

式中,$\boldsymbol{D}$ 表示原始矩阵;矩阵 $\mathbf{V}$ 的列向量是 $\boldsymbol{D}$ 的协方差矩阵的特征向量,按照特征值由大到小的顺序排列;矩阵 $\mathbf{U}$ 的列向量是 $\boldsymbol{D}^{\mathrm{T}}$ 的协方差矩阵的特征向量;$\mathbf{S}$ 为对角矩阵,其主对角线元素是相应特征值的平方根,特征值在数值上等于相应主成分的方差.

如果数据矩阵 $\boldsymbol{D}$ 的秩为 n,s 表示 $\boldsymbol{D}$ 矩阵行数和列数中的最小值,那么数值较大的特征值 $\lambda_1\cdots\lambda_n$ 属于主要因子,来自化学组分以及较大的仪器因子的贡献;数值较小的特征值 $\lambda_{n+1}\cdots\lambda_s$ 属于次要因子,主要来自噪声的贡献[3].

属于主要因子集的特征值既包含化学因子,也包含仪器误差,如基线漂移、非线性、信号展宽等.在特征值的主要因子集和次要因子集之间找到分界线对分析工作者是一种挑战.经验估计方法是根据对大量数据分析实例的总结,归纳出经验判据,以此区别主要因子和次要因子,进而估计主因子数.尽管其数学推证并不严密,但在实际应用中是有效的.其有效性的理论基础是:特征值在数值上等于主成分的方差,而主要因子和次要因子在方差上存在差异,如果特征值的某种组合能够体现这种差异,该组合即为一种有效判据.本文提出的 RVR 方法即属于经验估计方法.

RVR 方法的基本原理如下:分别以 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 表示矩阵 $\mathbf{U}$ 和 $\mathbf{V}$ 的列向量,即 $\mathbf{U}=[\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_s]$,$\mathbf{V}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_s]$.

Step 1 对原始矩阵 $\boldsymbol{D}$ 进行 SVD,如式(1)所示,得到特征向量 $\boldsymbol{v}_i$,然后通过下式获得主成分,

$$\boldsymbol{PC}_i=\boldsymbol{D}\times\boldsymbol{v}_i \tag{2}$$

式中 $\boldsymbol{PC}_i$ 表示第 i 个主成分.

Step 2 按照各因子对应特征值由大到小的顺序,从原始矩阵中依次扣除因子 1、因子 2、…,得到一系列残余矩阵 $\boldsymbol{R}_i$(i 为扣除的因子数),

$$\boldsymbol{R}_1=\boldsymbol{D}-\boldsymbol{PC}_1\times\boldsymbol{v}_1^{\mathrm{T}} \tag{3}$$

$$\boldsymbol{R}_i=\boldsymbol{R}_{i-1}-\boldsymbol{PC}_i\times\boldsymbol{v}_i^{\mathrm{T}}\ (i\geqslant 2) \tag{4}$$

Step 3 对各残余矩阵 $\boldsymbol{R}_i$ 进行 SVD,如下式,

$$\boldsymbol{R}_i=\boldsymbol{U}_i\boldsymbol{S}_i\boldsymbol{V}_i^{\mathrm{T}} \tag{5}$$

提取 $\boldsymbol{S}_i$ 矩阵的对角线元素,将各元素进行平方得到各残余矩阵的特征值向量,以 $\boldsymbol{t}_i$ 表示.

Step 4 将各残余矩阵的特征值向量作为列向量,构成形式如下的矩阵 $\boldsymbol{E}$,

$$\boldsymbol{E}=[\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_i,\cdots] \tag{6}$$

Step 5 计算 $\boldsymbol{E}$ 矩阵中各行的方差,再计算相邻两行的方差比值,以一系列 F 比值表示.根据 F 比值判断主因子数.

假设原始矩阵的主因子数为 n,当 $i\leqslant n$ 时,矩阵 $\boldsymbol{E}$ 前 i 行中,每行元素既包含主要因子的特征值,又包含次要因子的特征值,特征值间相差较大,则该行元素的方差将较大;当 $i>n$ 时,该行元素中仅包含次要因子,特征值之间相差不大,故方差较小.从主因子的角度,主因子数的判断标准应是 F 比值处于局部极大值时对应的因子数;从噪声因子的角度,主因子数应比 F 比值相互接近时对应的最小因子数小 1.由于存在仪器因子等不可避免的因素,根据上述两个标准判断的主因子数有时会出现不一致的情况,这时需要结合实验条件和判断标准对体系的主因子数进行综合判断.与一般方法比,该方法因具有两个判断标准,能起到相互验证的效果,因此提高了结果的可靠性.

根据该算法的计算步骤可以看出,使用该算法时需要对 n 个残余矩阵进行奇异值分解,耗时较长,后期对该算法进行改进,发现只对原始矩阵 $\boldsymbol{D}$ 进行奇异值分解,根据 $\boldsymbol{D}$ 的特征值可以构建出残余特征值矩阵 $\boldsymbol{E}$.下面给出 $\boldsymbol{E}$ 矩阵的推导过程.

根据 SVD 可以得到

$$\boldsymbol{D}=\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}=\sum_{k=1}^{s}\sqrt{\lambda_k}u_k v_k^{\mathrm{T}} \tag{7}$$

$$\boldsymbol{PC}_i=\boldsymbol{D}\times\boldsymbol{v}_i=\Big(\sum_{k=1}^{s}\sqrt{\lambda_k}u_k v_k^{\mathrm{T}}\Big)\cdot v_i=\sqrt{\lambda_i}u_i \tag{8}$$

将以上两式代入残余矩阵 $\boldsymbol{R}_i$ 的表达式,得到

$$\boldsymbol{R}_i=\boldsymbol{D}-\sum_{k=1}^{i}\mathrm{PC}_k\mathrm{V}_k^{\mathrm{T}}=$$

$$\boldsymbol{D}-\sum_{k=1}^{i}\sqrt{\lambda_k}u_k v_k^{\mathrm{T}}=\sum_{k=1}^{s}\sqrt{\lambda_k}u_k v_k^{\mathrm{T}}-\sum_{k=1}^{i}\sqrt{\lambda_k}u_k v_k^{\mathrm{T}} \tag{9}$$

整理式(9),得到

$$\boldsymbol{R}_i=\sum_{i+1}^{s}\sqrt{\lambda_k}u_k v_k^{\mathrm{T}} \tag{10}$$

由式(10)可以看出,矩阵 $\boldsymbol{R}_i$ 的特征值集合为 $\{\lambda_{i+1},\lambda_{i+2},\cdots,\lambda_s\}$,即

$$\boldsymbol{E}=\begin{bmatrix}\lambda_2 & \lambda_3 & \lambda_4 & \cdots\\ \lambda_3 & \lambda_4 & \lambda_5 & \cdots\\ \lambda_4 & \lambda_5 & \lambda_6 & \cdots\\ \lambda_5 & \lambda_6 & \cdots & \cdots\\ \lambda_6 & \cdots & \cdots & \cdots\\ \cdots & \cdots & \cdots & \cdots\end{bmatrix} \tag{11}$$

经计算,改进算法和原始算法计算的主因子数是相等的,且耗时较短.

下面介绍一些普适性较强的方法,如 RESO 方法、NPFPCA 方法等,将它们应用于对模拟数据和实验数据主因子数的估计中,并将 RVR 方法与它们的估计结果作比较,以显现 RVR 方法估计结果的优劣.

RESO 方法[4]是对原始数据分别进行主成分分析(PCA)和平滑主成分分析(SPCA)[5-6],得到两组特征值矩阵. RESO 值为矩阵中对应的特征值比值. 平滑主成分分析在一定程度上起到了滤噪的作用,因此主因子不受 SPCA 的影响,其 RESO 值接近于 1,而次要因子经过 SPCA 后发生较大的变化,RESO 值远小于 1.

NPFPCA[7]的基本原理与 RESO 相似,同样是对矩阵进行主成分分析和平滑主成分分析,但所比较的是特征向量间的差异,以一致性系数表示. 其中主因子的一致性系数接近于 1,其标准偏差接近于 0. 次要因子的一致性系数远小于 1,标准偏差大于 0.

DRMAD 方法[8]是计算一系列因子水平的残余标准偏差(RSD)[9],判断这一系列残余标准偏差中的异常值,根据异常值的数量来估计主因子数.

DRAUG 方法[10]是向原始矩阵中加入秩为 1、无误差的增广矩阵,比较两个矩阵在相邻因子水平的方差比值,以 F 比例表示,来判断主因子数.

VMmad 方法[11]是对数据的每一列向量进行目标测试,即修正的 Vogt-Mizaikoff F 检验[12],经过估计判断得到一系列的主因子数值,通过绝对偏差中位数检测这列数值中的离群值,其中离群值的数目即为主因子数.

2　数据处理结果

2.1　模拟数据

2.1.1　模拟数据计算结果

光谱数据为二乙醚、氨和 β-丙内酯在波数为 750～1 250 cm^{-1} 的光谱,浓度数据模拟了时间为 1～10 min 的 50 个等时间间隔点、以高斯函数形式表示的浓度分布. 将二者结合,再加入噪声水平为 0.01 的正态分布噪声,组成模拟气相色谱-红外联用数据. 理论上该数据估计出的主因子数应为 3. 根据 RVR 方法的判断标准和图 1,可知模拟数据的主因子数为 3,与理论值吻合.

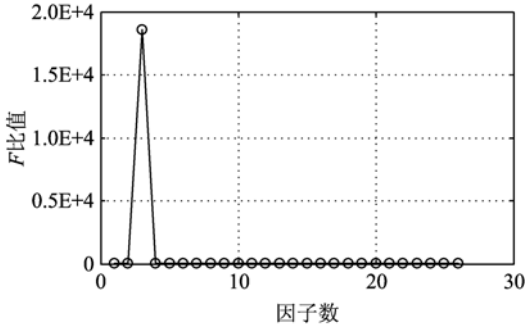


图 1　RVR 方法对模拟数据的处理结果

Fig. 1　The processing result of applying RVR to simulated data

2.1.2　模拟数据的极限情况

对模拟数据的分析主要是处理以下 3 种情况所能容忍的最大程度,即修改氨模拟色谱的峰位置、峰高度以及噪声水平,探究各参数值为多大时,仍能保证模拟数据的主因子数为 3:

(Ⅰ) 保持乙醚(图 2 中曲线 1)和丙内酯模拟色谱(图 2 中曲线 3)的形状和位置不变,仅改变氨模拟色谱(图 2 中曲线 2)的位置. 比较各方法对色谱重叠所能忍受的极限程度,进而比较各方法的处理效果.

(Ⅱ) 保持乙醚(图 3 中曲线 1)和丙内酯模拟色谱曲线(图 3 中曲线 3)的形状和位置不变,仅改变氨模拟色谱(图 3 中曲线 2)的色谱强度,比较各方法的微量程度,进而比较各方法的处理结果.

(Ⅲ) 保证各物质色谱曲线的形状和位置不变,仅改变加入的噪声水平,比较各方法所能承受的最大噪声水平.

由表 1 中第 1 列数据可知,RVR 方法、DRAUG 方法和 RESO 方法所能承受的乙醚和氨的色谱重叠程度较大. 第 2 列中,使用 RVR 和 RESO 方法时,当

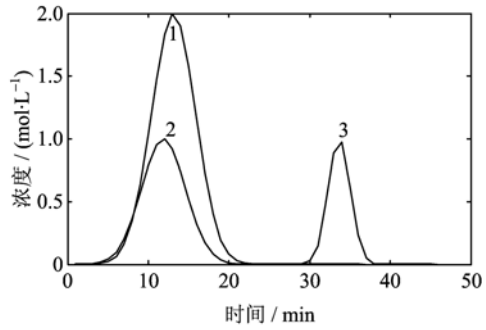


图 2 色谱重叠程度示意图

Fig. 2 The schematic diagram of chromatographic overlapping degree

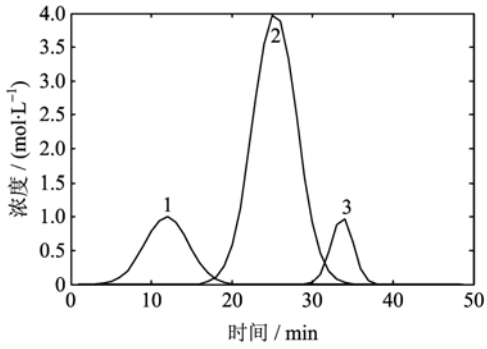


图 3 微量组分示意图

Fig. 3 The schematic diagram of minor component

色谱曲线的微量组分占据最大含量组分的 1/10 时，估计的主因子数仍为 3. 由此可推测出 RVR 方法处理微量组分的能力是极强的. 第 3 列中，RVR 方法所能承受的噪声水平仅次于 DRAUG，承受能力较强. 总体来说，RVR 方法处理模拟数据极限情况的能力较强.

表 1 使用 RESO，NPFPCA，DRMAD，DRAUG 和 RVR 方法分析色谱重叠、微量组分、噪声水平情况的极限

Tab. 1 Limitations of chromatographic overlapping, trace element and noise level in the data when methods RESO, NPFPCA, DRMAD, DRAUG and RVR are valid

方法	Ⅰ position	Ⅱ height	Ⅲ noise level
RESO	3.03	10	0.06
NPFPCA	3.06	5	0.05
DRMAD	3.14	2	0.01
DRAUG	3.04	8	0.08
RVR	3.04	10	0.07

【注】 Ⅰ列数据表示色谱图中峰的位置，数值越接近乙醚的色谱峰位置(3)，色谱重叠程度越大. Ⅱ列数据表示色谱峰的高度，数值与乙醚色谱峰的高度(1)相差越大，所能承受的组分微量程度越大. Ⅲ列数据表示所能承受的噪声水平，数值越大，所能承受的噪声水平越大.

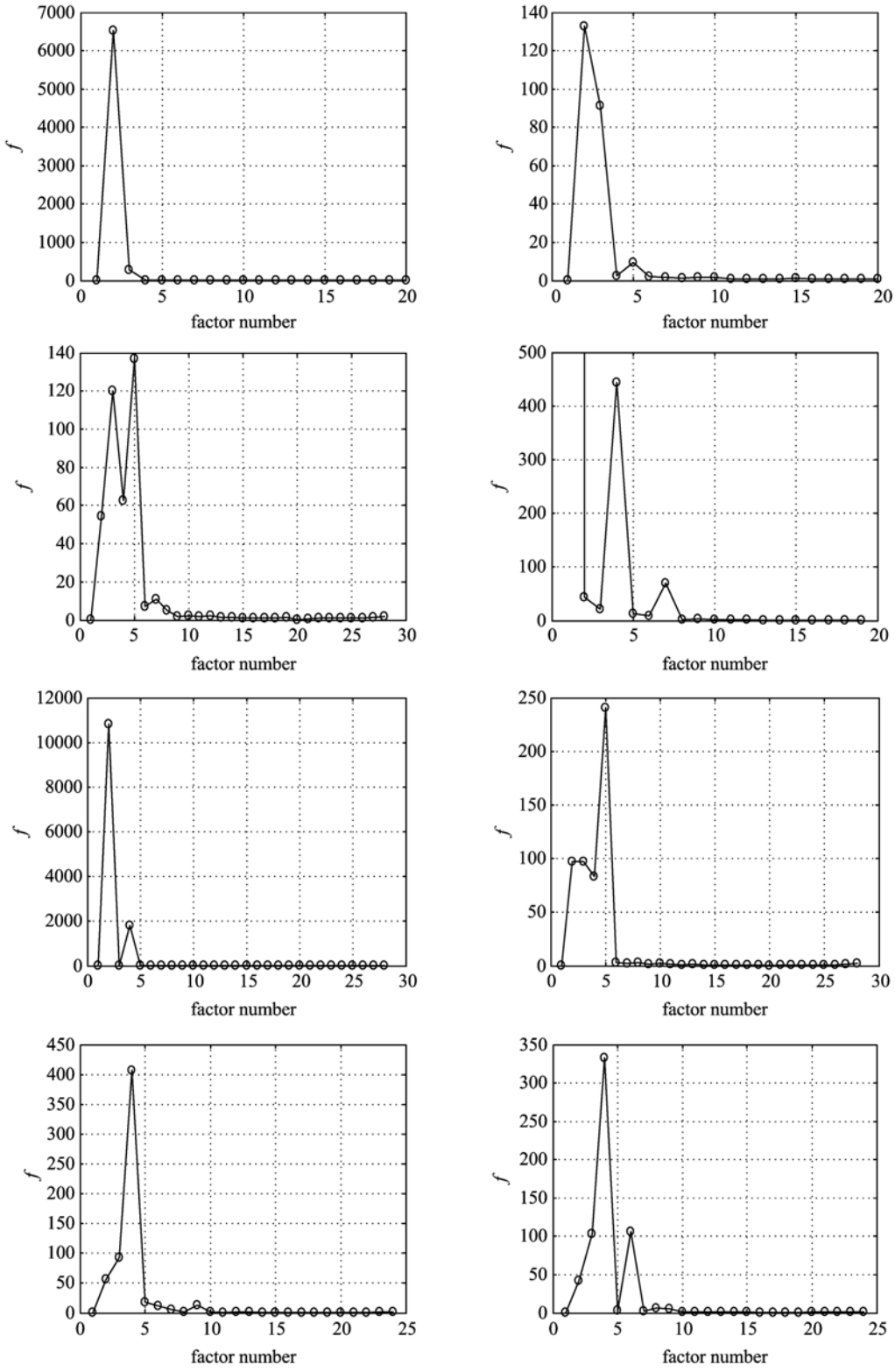
2.2 实验数据

红外光谱实验数据属于大气光谱数据，以“光谱数据 1”和“光谱数据 2”表示. 测量中采用 MDA (Atlanta，GA)型 FTIR 光谱仪，Bomem Michelson 100 型干涉仪，MCT(HgCdTe)型检测器；反射器为立方角阵列反射器；所使用的望远镜的直径为 31.5 cm，可用于产生准直红外光束.

采用稀土元素的 HPLC-DAD 数据作为实验数据，分别制备了二组分(Yb 和 Tm)，三组分(Yb，Tm 和 Er)以及六组分(Lu，Yb，Tm，Er，Ho 和 Tb)的样品. 采用 FL2000 HPLC 产品(Spectra-Physics，USA)型工作站，多波长 UV-Vis 检测器(Spectra-Physics，USA)，其中波长检测范围为 580～720 nm，采样间隔 5 nm，采样时间为 15 min，采样时间间隔约为 0.344 s. “色谱数据 1”和“色谱数据 2”分别选取二组分样品在 4.5～8.6 min 之间的色谱数据，“色谱数据 3”和“色谱数据 4”分别选取三组分样品在 4.5～9.9 min 之间的色谱数据，“色谱数据 5”和“色谱数据 6”分别选取六组分样品在 3.9～12 min 之间的色谱数据.

图 4 是使用 RVR 方法对实验数据的处理结果图，横坐标是因子数，纵坐标为 F 比值. 估计的结果是根据 F 比值差异和实验条件的综合，选择性地使用判断标准得到的. 从主因子的角度，主因子数的判断标准应是 F 比值处于局部极大值时对应的因子数；从噪声因子的角度，主因子数应比 F 比值相互接近时对应的最小因子数小 1. 可能是由于存在部分主因子对体系的贡献较小，它们常被误认为是噪声因子，所以会出现上述两个标准判断不一致的情况. 而噪声因子间的差异是极小的，所以从某种程度上说，使用第二个判断标准估计得到的主因子数更准确些，即主因子数应比 F 比值相互接近时对应的最小因子数小 1. 由图可判断出主因子数为图中呈现出直线时的最小因子数减 1，故从上到下、从左到右，可判断出其主因子数依次为 4，5，8，7，4，5，9，6.

表 2 是 RVR 方法与其他判断方法对实验数据的处理结果比较. 第 1 列是图 4 中的相应数据名称，第 2 列矩阵大小以行数×列数形式表示，第 3～8 列数据表示对应方法估计得到的主因子数. 使用 VMmad 方法估计光谱数据 1 和光谱数据 2 时，显示估计结果为 n/a，表示未估计出结果. 期望值有一个最小值，是指各数据体系中包含的最少物质种类数，大部分方法估计的主因子数比期望最小值大，是



从上到下，从左到右，对应的数据依次为光谱数据 1、光谱数据 2、色谱数据 1、色谱数据 2、色谱数据 3、色谱数据 4、色谱数据 5、色谱数据 6

图 4　RVR 方法对实验数据的处理示意图

Fig. 4　The processing diagrams of experimental data by using RVR

表 2 RESO,NPFPCA,DRMAD,DRAUG,VMmad,RVR 方法对实验数据的处理结果

Tab. 2 The processing results of experimental data by using RESO, NPFPCA, DRMAD, DRAUG, VMmad, RVR

数据	矩阵大小	RESO	NPFPCA	DRMAD	DRAUG	VMmad[a]	RVR	期望值[b]
光谱数据 1	600×1 502	2	2	5	7	n/a	4	≧ 3
光谱数据 2	332×1 502	2	4	4	5	n/a	5	≧ 3
色谱数据 1	717×29	3	3	8	7	5	8	≧ 2
色谱数据 2	932×29	2	3	9	8	6	7	≧ 3
色谱数据 3	717×29	4	3	6	7	3	4	≧ 2
色谱数据 4	932×29	3	3	7	8	4	5	≧ 3
色谱数据 5	1 600×25	3	3	8	7	4	9	≧ 6
色谱数据 6	1 600×25	5	4	8	7	6	6	≧ 6

【注】 a 对于光谱数据 1 和光谱数据 2,VMmad 方法没有给出结果,因为矩阵的行数比列数小,不满足该方法适用条件.
b 期望范围是根据实验情况推测出的样品的主要成分数量.

因为非化学成分也可能向数据矩阵提供信息,例如仪器噪声、仪器脉动以及数据预处理,它们会增加主因子数.各方法估计的主因子数在期望范围内,表示该方法估计结果相对准确.

3 结论

RVR 方法是一种经验判据.通过该方法对模拟数据的处理情况可看出,RVR 方法对模拟数据以及模拟极限的处理情况较好,它处理重叠色谱峰、微量组分的能力以及所能忍受的噪声水平较强.根据对实验数据的处理情况,RVR 方法对实验数据的估计值均在期望范围内.虽然 RESO 方法所模拟的极限能力较强,但是它对实验数据的估计值偏低.NPFPCA,DRMAD,DRAUG 方法所模拟的极限不如 RVR 方法所能忍受的极限强.修正的 Vogt-Mizaikoff F 检验方法的部分估计值偏低,它适用于行数比列数大的矩阵,而且 F 检验需要实验误差满足正态分布,而 RVR 方法对实验误差无要求.总体上,RVR 方法对数据主因子数的估计效果较好,但需要注意的是,RVR 方法具有两个判断标准,选用哪种标准来确定最终的主因子数需要结合实验情况来选择.

参考文献(References)

[1] De Lathauwer L, De Moor B, Vandewalle J. A multilinear singular value decomposition[J]. SIAM Journal on Matrix Analysis and Applications, 2000, 21(4): 1 253-1 278.

[2] Golub G H, Reinsch C. Singular value decomposition and least squares solutions [J]. Numerische Mathematik, 1970,14(5): 403-420.

[3] Abdi H, Williams L J. Principal component analysis [J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2010,2(4): 433-459.

[4] Chen Z P, Liang Y Z, Jiang J H, et al. Determination of the number of components in mixtures using a new approach incorporating chemical information [J]. Journal of Chemometrics, 1999,13(1): 15-30.

[5] Rice J A, Silverman B W. Estimating the mean and covariance structure nonparametrically when the data are curves[J]. Journal of the Royal Statistical Society, 1991,53: 233-243.

[6] Silverman B W. Smoothed functional principal components analysis by choice of norm[J]. The Annals of Statistics, 1996,24(1): 1-24.

[7] Xu C J, Liang Y Z, Li Y, et al. Chemical rank estimation by noise perturbation in functional principal component analysis [J]. Analyst, 2003, 128(1): 75-81.

[8] Malinowski E R. Determination of rank by median absolute deviation (DRMAD): A simple method for determining the number of principal factors responsible for a data matrix[J]. Journal of Chemometrics, 2009, 23(1): 1-6.

[9] Meloun M, Capek J, Miksik P, et al. Critical comparison of methods predicting the number of components in spectroscopic data [J]. Analytical Chimica Acta, 2000, 423(1): 51-68.

[10] Malinowski E R. Determination of rank by augmentation (DRAUG)[J]. Journal of Chemometrics, 2011,25(6):323-358.

[11] Malinowski E R. Adaptation of the Vogt-Mizaikoff F-test to determine the number of principal factors responsible for a data matrix and comparison with other popular methods[J]. Journal of Chemometrics, 2004, 18(9): 387-392.

[12] Vogt F, Mizaikoff B. Dynamic determination of the dimension of PCA calibration models using F-statistics [J]. Journal of Chemometrics, 2003,17(6): 346-357.