基于PCA的代谢组学数据滤噪方法

丁　俊,邓伶莉,许晶晶,董继扬

(厦门大学电子科学系,福建省等离子体与磁共振研究重点实验室,福建　厦门　361005)

摘　要:代谢组学数据不可避免地受到各种刺激因素的作用,如何降低干扰因素的影响是代谢组学数据预处理的一个重要任务。详细分析了代谢组学数据方差的构成及其在特征空间中的分布特点,并在此基础上提出一种滤除未知干扰因素的新方法,提高感兴趣因素的显著性。文中采用真实的代谢组学数据验证新滤波算法的有效性,并与正交信号校正(orthogonal signal correction,OSC)方法进行比较。实验结果表明,新滤波方法可以在抑制未知干扰因素影响的同时,较好地保留感兴趣因素信息以及生物体内在的个体差异信息,降低模型发生过拟合的危险,使后续的统计分析结果更可靠。

关键词:代谢组学;干扰因素;主成分分析(PCA)

中图分类号:O 482.53^{+}2　　**文献标识码**:A　　**文章编号**:1673-1255(2012)01-0060-06

Metabolomics Data Filtering Method Based on PCA

DING Jun, DENG Ling-li, XU Jing-jing, DONG Ji-yang

(Department of Electronic Science, Fujian Provincial Key Laboratory of Plasma and Magnetic Resonance, Xiamen University, Xiamen 361005*, China)*

Abstract: The metabolomics dataset is disturbed by various stimuli inevitably. The main task for metabolomics data preprocessing is to reduce the impacts of the disturbing factors. In present work, the formation of data variance and their distribution in feature space are analyzed. Furthermore, a new method to filtrate unknown disturbing factors is proposed and the significance of interesting factors is improved. The efficiency of the new filtering algorithm is estimated by real metabolomics dataset. Comparing with orthogonal signal correction (OSC) method, the experiment shows that the new method is superior in reducing unknown disturbing factors and retaining useful information and intrinsic individual differences in organisms. In addition, it can also prevent the overfitting of model and make the subsequent statistical analysis more reliable.

Key words: metabolomics; disturbing factors; principal component analysis (PCA)

代谢组学是20世纪90年代末发展起来的一个新兴研究领域[1],它借助核磁共振(NMR)和色谱质/谱联用等现代高通量分析技术,研究生物体内源性代谢物质的整体及其对内因和外因变化应答规律[2]。已广泛应用于药物毒性及安全性评价、疑难疾病诊断、新药研发及药物作用机制研究等生命科学的多个领域[3,4]。

处于复杂环境中的生物体不可避免地受到各种内外界刺激因素的作用,当试图探究某种刺激因素(如疾病、饮食、药物干预等)对生物体代谢过程的作用时,其他不感兴趣的刺激因素便成为了干扰因素。若这些干扰因素对生物体的作用过大,则会影响后续的分析结果的准确性,造成异常代谢通路和相关生物标志物的辨识错误。因此,减少干扰因素的影响是代谢组学数据预处理的一个关键步骤。

收稿日期:2012-02-13

基金项目:国家自然科学基金(81171331;11175149);中央高校基本科研业务费专项资金(2011121046)

作者简介:丁俊(1986-),男,江西上饶人,硕士研究生,研究方向为核磁共振代谢组学;董继扬(1974-),男,福建安溪人,教授,研究方向为核磁共振代谢组学和数字图像处理.

正交信号校正(orthogonal signal correction, OSC)是代谢组学中常用的滤波方法,它通过滤除自变量数据矩阵$\boldsymbol{X}$中与应变量矩阵$\boldsymbol{Y}$正交的部分,以此滤除与应变量$\boldsymbol{Y}$不相关的信息。但由于代谢组学数据中,样本的个体差异和波谱数据噪声是不可避免的,而这些原本与$\boldsymbol{Y}$矩阵相互独立的噪声和个体差异信息经OSC方法校正后,便会与$\boldsymbol{Y}$矩阵具有很强的相关性。这些虚假的相关性将严重影响后续统计分析结果,导致后续的系统建模产生过拟合等现象[5]。

文中提出一种基于主成分分析(principal component analysis, PCA)的滤波方法(记为PCF),抑制未知刺激因素的影响,增强感兴趣因素的显著性,从而简化后续的多变量统计分析。采用真实的^{1}H NMR代谢组学数据验证新滤波方法的有效性,并与OSC方法做比较。结果表明,新方法对未知干扰因素的滤除效果优于OSC,且滤波后数据的PLS-DA模型的解释能力(R^2)和预测能力(Q^2)都得到了提高。

1　理论与方法

1.1　数据的方差分析(ANalysis Of VAriance, ANOVA)

设观测数据矩阵为$\boldsymbol{X}=(x_{lnm})$。其中$l=1,2,\cdots,L$,表示系统受到具有L个因素水平的刺激因素的作用;$n=1,2,\cdots,N$;$m=1,2,\cdots,M$,表示数据集包含N个样本,每个样本有M个变量。则$\boldsymbol{X}$的总离差平方和(sum of squares)可表示为[6]

$$S(\boldsymbol{X})=\sum_{l=1}^{L}\sum_{n=1}^{N_l}\sum_{m=1}^{M}(x_{lnm}-\bar{x}_{\cdot\cdot m})^2,$$

$$\bar{x}_{\cdot\cdot m}=\frac{1}{N}\sum_{l=1}^{L}\sum_{n=1}^{N_l}x_{lnm} \tag{1}$$

其中,N_l表示第l个因素水平下样本数,且$N=\sum_{l=1}^{L}N_l$;$\bar{x}_{\cdot\cdot m}$表示第m个变量在数据集中的均值。记$\bar{x}_{l\cdot m}=\frac{1}{N_l}\sum_{n=1}^{N_l}x_{lnm}$,则$S(\boldsymbol{X})$可表示为

$$\begin{aligned}S(\boldsymbol{X})=&\sum_{l=1}^{L}\sum_{n=1}^{N_l}\sum_{m=1}^{M}(x_{lnm}-\bar{x}_{l\cdot m})^2+\\&\sum_{l=1}^{L}\sum_{n=1}^{N_l}\sum_{m=1}^{M}(\bar{x}_{l\cdot m}+\bar{x}_{\cdot\cdot m})^2+2\times\\&\sum_{l=1}^{L}\sum_{n=1}^{N_l}\sum_{m=1}^{M}(x_{lnm}-\bar{x}_{l\cdot m})(\bar{x}_{l\cdot m}+\bar{x}_{\cdot\cdot m})\end{aligned} \tag{2}$$

式(2)中第三项(即交叉项)值为0。因此$S(\boldsymbol{X})$可进一步表示为

$$S(\boldsymbol{X})=S_E+S_F \tag{3}$$

其中,$S_F=\sum_{l=1}^{L}\sum_{n=1}^{N_l}\sum_{m=1}^{M}(\bar{x}_{l\cdot m}+\bar{x}_{\cdot\cdot m})^2$,

$S_E=\sum_{l=1}^{L}\sum_{n=1}^{N_l}\sum_{m=1}^{M}(x_{lnm}-\bar{x}_{l\cdot m})^2$。$S_F$是由于因素水平不同而引起的方差(组间方差),$S_E$是同一因素水平的样本子集的方差(组内方差),来源于个体差异和噪声方差等。

当数据集不受其他干扰因素影响,且样本量足够大时,通常可假设组内方差S_E正态分布于特征空间中。由于刺激因素的作用通常是局部的,即刺激因素对不同代谢物的影响不同,或只影响一部分的代谢物浓度,因此S_F在特征空间中的分布不均匀。若用PCA进行分析,则S_E将均匀地分布在各主成分(PC)上[7],而S_F将主要集中在前几个PCs上。换句话说,在特征空间中,S_F是各向异性的,而S_E是各向同性的。

当数据集受到未知干扰因素(F′)作用时,其所产生的方差$S_{F'}$将混合在S_E中。设噪声和个体差异的信息用为$S_{E'}$,则组内方差为

$$S_E=S_F+S_{E'} \tag{4}$$

S_F在特征空间的各向异性使得S_E也是各向异性的。利用这一性质可以判断数据集是否受到未知干扰因素的作用,甚至可以对未知干扰因素的作用进行粗略估计。

1.2　基于PCA的滤波方法

当刺激因素(F')未知,即各样本的F'因素水平未知时,常规ANOVA方法很难将式(4)右边的两项$S_{F'}$和$S_{E'}$分离开。理论上,$S_{E'}$在特征空间中正态分布,信息的相关性很弱,S_F的分布为非高斯且具有较强的相关性。若对数据矩阵进行PCA建模,$S_{F'}$将集中在前面几个PCs所张成的子空间中,而$S_{E'}$将集中在最后几个PCs所张成的子空间中。根据这一特性,文中提出一种基于PCA的滤波方法如下:

(1) 按已知因素对数据$\boldsymbol{X}$进行划分,使每一组(X_l, l=1,2,...,L)内的样本具有相同的因素水平。

(2) 对每一组样本$\boldsymbol{X}_l$进行如下PCA分析:$\boldsymbol{X}_l=\boldsymbol{TP}'+\boldsymbol{E}$。从$i$=1开始,依次提取$\boldsymbol{X}_l$的第$i$个主成分,并计算第$i$个次成分所解释的方差比例$\alpha_i$,直到$\alpha_i\leqslant\alpha_0$

为止。

(3) 删除前 i 个主成分，重构数据 $\boldsymbol{X}_l^* = \boldsymbol{X}_l - \boldsymbol{TP}'$。其中，$\boldsymbol{T}=(T_1,T_2,...,T_i)$，$\boldsymbol{P}=(P_1, P_2,..., P_i)$。

(4) 用 $\boldsymbol{X}_l^*$ (l=1,2,…, L)重组新的数据矩阵 $\boldsymbol{X}^*$，抑制未知干扰因素的影响。

上述算法称为基于PCA的滤波方法(简称PCF)。可以通过选取适当的参数值 α_0 ($0 \leqslant \alpha_0 \leqslant 1$)来控制滤波效果。$\alpha_0$ 取值越大，个体差异信息保留越完整，但对未知因素的抑制作用越小。但如果 α_0 取值太小，个体差异信息丢失太多，可能使分析模型发生过拟合现象。α_0 的选择应满足在滤除干扰因素的同时，尽可能保留个体差异信息的原则。

2 实验数据集及其方差构成

2.1 实验数据集

文中数据集来自一个关于素食人群代谢差异的研究项目[8]，由41个普通饮食男性志愿者(OM)、40个普通饮食女性志愿者(OF)、42个奶素食男性志愿者(VM)以及38个女性奶素食志愿者(VF)4个组的尿液样品的^{1}H-NMR谱组成。先利用MestRe-C V2.3软件(http://qobrue.usc.es/jsgroup/MestRe-C)及自编软件对谱数据进行手动调相、基线校正、谱峰对齐、以及去除残余水峰、尿素峰和定标物DSS的谱峰；然后采用等间隔积分方法将0.5 ~ 8.8 ppm区间的谱数据积分为178个点。由于ANOVA分析要求每个实验组应具有相同的样本数，先对OM,OF,VM和VF 4个组的样本分别进行PCA分析，删除落入95% Hotelling T^2置信区间之外的样本，使每一组中均包含35个样本，构成一个140×178的数据矩阵。最后，对该数据矩阵进行GAN[7]行归一化和Pareto列标准化[9]，得到结果矩阵 $\boldsymbol{X}$ 供后续的实验分析。

2.2 方差构成分析

对数据矩阵 $\boldsymbol{X}$ 进行PCA分析，然后对前3个PCs进行ANOVA分解，计算各因素矩阵的方差，其结果见表1。

表1 PCA模型前三个主成分所解释数据 $\boldsymbol{X}$ 方差比例情况

成分	饮食	性别	交叉因素	残值	总计
$\boldsymbol{X}$	19.26	27.93	7.05	45.76	100.00
PC1	16.96	27.42	6.12	18.92	69.42
PC2	0.94	0.02	0.07	8.62	9.65
PC3	1.08	0.11	0.57	3.00	4.76

从表1的的结果可见，$\boldsymbol{X}$ 中饮食因素和性别因素所引起的方差分别只占总方差的19.26%和27.93%，两种因素的交叉作用引起的方差占7.05%，而残余方差占到了总方差的45.76%。2种因素的交叉作用不为零，说明这两种因素不独立，即不同性别的人对饮食的反应存在一定的统计差异。对 $\boldsymbol{X}$ 进行PCA分析，则每一主成分上均混有各种因素的信息。例如：PC1上包含了绝大部分的性别因素(16.96/19.26约88.1%)、饮食因素(27.42/27.93约98.2%)、和分交叉作用因素(6.12/7.05约86.8%)的方差，还包含残余矩阵的方差(18.92/45.76约41.3%)。此外，模型前三个PCs解释大部分的残余矩阵方差((18.92+8.62+3.00)/45.76=66.74%)，且 $p(1):p(2):p(3)$=6:3:1，即残余方差在特征空间上也表现为各向异性，由此可以推测，数据 $\boldsymbol{X}$ 可能还受到其他未知因素的影响。

3 实验及结果分析

3.1 性别因素分析

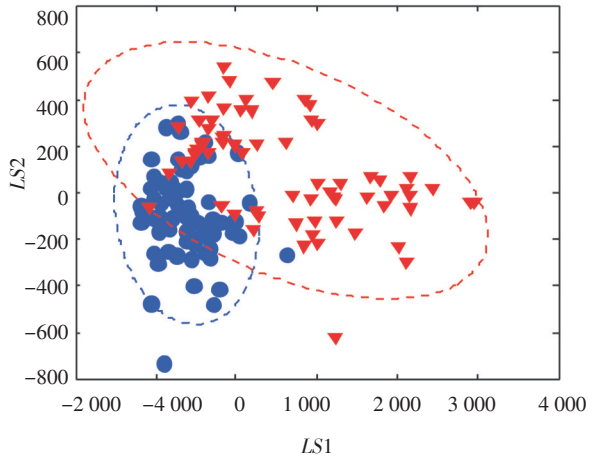
为了验证PCF算法的有效性，实验将性别因素作为感兴趣因素，因此系统的干扰因素包括：饮食因素、性别与饮食的交叉因素以及未知的干扰因素等。从表1可知，在 $\boldsymbol{X}$ 的第一主成分上只有39.50%(27.42/69.42)是性别因素的作用，其他信息包括：24.43%的饮食信息、8.8%的交叉因素信息、以及27.25%残余信息(如噪声、个体差异和其他未知干扰因素)。

取 $\alpha_0=0.5$ 分别对两组样本进行PCF滤波,再对滤波后的数据 $\boldsymbol{X}^*$ 进行PCA分析,观察各种干扰因素的滤除情况,并与OSC(删除一个成分)方法作比较,结果如表2所示。

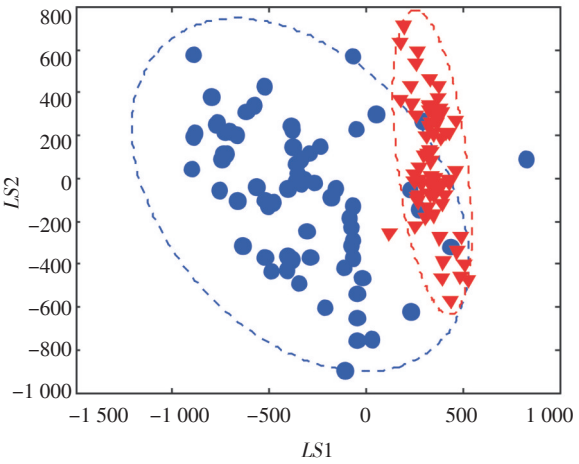
表2　PCA模型前三个主成分所解释滤除后数据方差比例情况

方法	成分	饮食	性别	交叉因素	残值	总计
OSC	$\boldsymbol{X}^*$	19.06	27.94	6.95	40.68	94.63
	PC1	17.06	27.43	6.08	18.83	69.40
	PC2	1.45	<0.01	<0.01	3.81	5.26
	PC3	0.24	<0.01	0.46	3.58	4.28
PCF	$\boldsymbol{X}^*$	2.14	27.54	1.53	30.70	61.91
	PC1	0.40	26.79	0.75	3.76	31.70
	PC2	0.65	0.13	0.26	3.46	4.50
	PC3	0.83	<0.01	0.23	3.41	4.47

从表2中可以看出,经PCF滤波后,饮食因素和交叉因素的信息已经大部分被滤除,残差信息也有部分被滤除,其所占比例分别从45.76%下降到25.97%和30.7%。滤波前残余方差 S_E 在前三个中PCs分别为18.92、8.62和3.00,而PCF滤波后 S_E 的分布变为3.76、3.46和3.41。可见,PCF滤波后残差方差在特征空间中的分布变均匀了,说明残差矩阵中的未知因素得到了很好的抑制。而OSC滤波后,各种干扰因素的方差变化不大,残余方差 S_E 在前三个PCs上的比例也变化不大,仍然具有较强的各向异性,这说明OSC方法(删除一个成分)不能滤除 $\boldsymbol{X}$ 中的干扰因素。对滤波后的数据进行PLS-DA模型分析,计算模型的的得分图,结果如图1所示。



(a) OSC滤波



(b) PCF滤波. (▼: VF+OF, and ●: VM+OM)

图1　性别因素的PLS-DA得分图

由图1可见,在原始数据中,由于饮食等干扰因素的影响,男性、女性两类样本在PLS-DA得分图中有部分重叠。OSC校正后,得分图的重叠现象没有得到改善;CPF滤波后,重叠现象在第一个隐变量(*LS*1)方向上有一定改善,女性样本点更集中;PCF滤波后,得分图中两类样本的重叠现象有明显改善,不仅女性样本点明显集中,男性样本点的集中程度也得到改善,两类样本在(*LS*1/*LS*2)得分图上可分性最大。文中采用Fisher准则[10]对两类样本的线性可分性($\boldsymbol{J}$)进行定量比较

$$J=\frac{|\tilde{m}_1-\tilde{m}_2|^2}{\tilde{S}_1^2+\tilde{S}_2^2} \tag{5}$$

其中,$\tilde{m}_1$,$\tilde{m}_2$ 表示两类样本得分的均值;$\tilde{S}_1$,$\tilde{S}_2$ 表

示两类样本得分的标准差。

表3给出了滤波后PLS-DA模型的解释能力(R^2)、预测能力(Q^2)以及两类样本的线性可分性(J)的比较结果。可见,各种滤波方法滤波后,第一主成分*LS*[1]的解释能力、预测能力以及线性可分性均有不同程度的提高。PCF滤波方法对模型预测能力和样本可分性的改善最为明显,OSC方法的改善不明显。这些结果均与图1和表2的结果相吻合。

表3 PLS-DA模型的解释能力 R^2 和预测能力 Q^2

	LS[1]			*LS*[2]		
	R^2/(%)	Q^2/(%)	J	R^2/(%)	Q^2/(%)	J
$\boldsymbol{X}_0$	40.90	40.10	0.88	17.20	19.50	0.85
OSC	40.90	40.20	0.89	18.00	22.50	1.17
PCF	46.00	45.30	7.99	17.30	23.80	0.15

经OSC或PCF滤波后,PLS-DA模型的解释能力和预测能力均得到提高,且PCF滤波后更为明显。这些结果说明:PCF滤波方法能有效地抑制X_0中的干扰因素,而OSC方法不适合于抑制干扰因素的影响。也就是说,基于PCF的滤波是一种有效抑制干扰因素影响的方法。

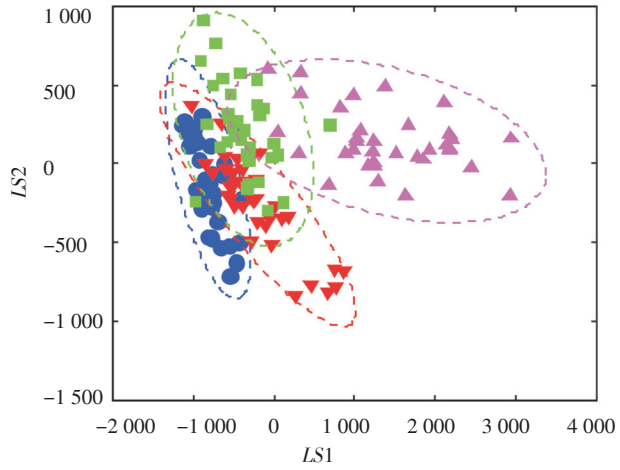
3.2 双因素分析

考虑饮食和性别2个已知因素,残余矩阵在$\boldsymbol{X}$前三个PCs上的方差分布不均匀,说明$\boldsymbol{X}$可能还受到其他未知因素的干扰。将$\boldsymbol{X}$中的样本按因素水平分为素食女性(VF)、普食女性(OF)、素食男性(VM)和普食男性(OM)4组,分别用MCF方法进行滤波($\alpha_0=0.5$),得到新的数据矩阵$\boldsymbol{X}$*。再用ANOVA方法分析数据方差结构的变化,并与OSC(去除一个隐变量)方法的结果进行比较,如表4所示。

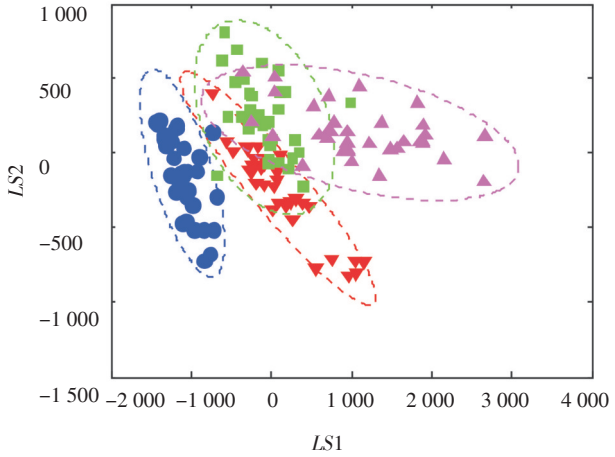
表4可见,OSC滤波后,虽然有一部分(约11.12%)残余方差被滤除,但S_E在前三个PCs上仍然分布不均匀,数据方差仍然大部分集中的PC1上,说明OSC的滤波不是针对未知干扰因素的,也不适合于未知干扰因素的滤除。PCF方法滤除了S_E中的41.26%的信息,而且滤波后S_E的方差均匀地分布在各PCs上。此外,经PCF滤波后,饮食因素和性别因素的显著性得到了很大的提高,例如:在$\boldsymbol{X}$的PC1上,饮食因素和残余方差的比例为16.96:18.92,而PCF滤波后($\boldsymbol{X}$*)的PC1上,饮食因素和残余方差的比例大幅度提高到了17.79:4.54,即残余方差对饮食和性别因素的干扰就很小了,这一点可以从PCF滤波前后的PLS-DA得分图得到直观的验证,如图2所示。

表4 PCA模型前三个主成分所解释滤除后数据方差比例情况

方法	成分	饮食	性别	交叉因素	残差	总计
OSC	$\boldsymbol{X}$*	19.26	27.90	7.05	40.67	94.88
	PC1	17.15	27.40	6.05	16.82	67.42
	PC2	1.57	<0.01	<0.01	3.79	5.36
	PC3	0.24	<0.01	0.46	3.57	4.27
MCF	$\boldsymbol{X}$*	19.26	27.93	7.05	26.88	81.12
	PC1	17.79	26.54	4.96	4.54	53.83
	PC2	0.52	0.69	0.15	3.06	4.42
	PC3	0.29	0.59	0.32	3.01	4.21



(a) 滤波前的得分图



(b) 滤波后的得分图(▼: OF, ●: OM, ■:VM, ▲:VF)

图2　PCF滤波前后的PLS-DA分析结果

从图2中可以看出,素食女性(VF)、普食女性(OF)、素食男性(VM)和普食男性(OM)4组样本在滤波前的PLS-DA得分图上重叠较为严重,而滤波后的4组样本在得分图上重叠现象得到了改善。实验还分析了OSC滤波后数据的PLS-DA模型,结果表明,OSC滤波前后的PLS-DA模型没有明显变化(这里没有提供OSC滤波后的得分图)。

4　总结与讨论

生物体作为一个复杂的系统,不可避免地受到各种来自周围环境和内在基因等刺激因素的作用,当试图分析定量分析生物体对某一个刺激因素的代谢响应时,那些无关因素将对感兴趣因素的识别与分析产生干扰作用。文中通过对代谢组学数据矩阵方差构成进行分析,根据他们在特征空间分布的特点,提出基于PCA的代谢组学数据滤波方法,将数据矩阵按感兴趣刺激因素水平分组,分别对每一组进行次成分分析,并删除主要由干扰因素引起的主成分,得到滤波后的数据用于进一步的数据分析。对素食、普食男女人群尿液一维^{1}H NMR数据分析,分别考察了把其中一种因素作为感兴趣因素,另一种因素作为未知因素进行滤波的效果,及已知2种刺激因素时的滤除效果。结果表明,文中方法对未知刺激因素引起的干扰有较好的滤除效果,同时保留了极大部分感兴趣因素信息和生物样本个体差异信息,有利于下一步对特征代谢物的检测。

参考文献

[1] J K Nicholson, J C Lindon, E Holmes. Metabonomics′: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data, Xenobiotica, 1999, 29 : 1181–1189.

[2] H Tang,Y Wang. Metabonomics: a revolution in progress[J]. Progress in biochemistry and biophysics.2006,33: 401–417.

[3] M Coen, E Holmes, J C Lindon, et al. NMR-based metabolic profiling and metabonomic approaches to problems in molecular toxicology[J]. Chemical research in toxicology, 2008,21: 9–27.

[4] E Y Xu, W H Schaefer, Q Xu. Metabolomics in pharmaceutical research and development: metabolites, mechanisms and pathways, Current opinion in drug discovery & development, 2009,12: 40.

[5] H C Goicoechea, A C Olivieri. A comparison of orthogonal signal correction and net analyte preprocessing methods[J]. Theoretical and experimental study, Chemometrics and Intelligent Laboratory Systems, 2001,56: 73–81.

[6] Z Sheng, S Xie, C Pan. Probability theory and mathematical statistics[M]. Beijing:China Higher Education Press, 2001.

[7] J Dong, K Chen, J Xu, et al. Group aggregating normalization method for the preprocessing of NMR-based metabolomic data[J]. Chemometrics and Intelligent Laboratory Systems ,2011, 32:262–268.

[8] J Xu, S Yang, S Cai, et al. Identification of biochemical changes in lactovegetarian urine using 1 H NMR spectroscopy and pattern recognition[J]. Analytical and Bioanalytical Chemistry, 2010,396: 1451–1463.

[9] H Kubinyi. 3D QSAR in drug design: theory, methods and applications[M]. Kluwer Academic Publishers, 1993.

[10] R O Duda, P E Hart. Pattern classification and scene analysis[M]. A Wiley-Interscience Publication, New York: Wiley, 1973.