Vol. 37 高等学校化学学报 No. 4
2016 年 4 月 CHEMICAL JOURNAL OF CHINESE UNIVERSITIES 626 ~ 632

doi: 10.7503/cjcu20150826

多批次肝衰竭患者呼出气体的电喷雾萃取电离质谱检测及代谢组学数据分析

李鹏辉[1]，邓伶莉[1,2]，罗 娇[3]，李 巍[3]，宁 晶[1]，丁健桦[1]，邬小萍[3]

(1. 东华理工大学江西省质谱科学与仪器重点实验室，南昌 330013;
2. 东华理工大学信息工程学院，南昌 330013; 3. 南昌大学第一附属医院，南昌 330123)

摘要 采用高分辨电喷雾萃取电离质谱(EESI-MS)技术对肝衰竭患者和健康志愿者呼出气体样本进行快速检测，结合多块偏最小二乘分析(MB-PLS)方法，对多批次获取的呼出气体代谢数据进行统计建模分析，并与传统的 PLS 方法进行比较. 结果表明，MB-PLS 方法能有效消除批次差异对统计建模的影响. 此外，利用 MB-PLS 模型变量 VIP 值对变量进行筛选，可降低数据的冗余，消除无关变量对模型的影响，从而有效提高了模型的性能.

关键词 呼出气体; 代谢组学; 电喷雾萃取电离质谱; 多块偏最小二乘分析

中图分类号 O657.6 **文献标志码** A

呼吸是人体基本的生命体征之一，人体呼出气体作为生物媒介携带了大量生理/病理信息，有研究报道在健康人体呼出气体中检测出 3000 多种化合物[1]. 传统的代谢组学方法主要分析生物体液[2~4](血清、尿液、唾液、乳液和组织液等)和生物组织中代谢物水平[5]的变化规律，实际上生物呼出气体也可作为代谢组学的研究对象，用于探索机体生理/病理状态[6,7]. 近年来，在基于代谢组学的疾病研究领域[8~10]，尤其是肝病相关领域，呼出气体因其安全、采样方便且非侵入性、不涉及个人隐私问题以及携带大量代谢信息等特点而备受关注[11,12].

人体呼出气体中代谢物含量极低[13]，因此对检测仪器的灵敏度有很高的要求，这一直是限制呼出气体代谢组学发展的重要原因之一. 随着现代分析技术的快速发展，呼出气体检测技术也逐渐丰富，如气相色谱-质谱联用(GC-MS)[1,12,14]、离子分子反应质谱(IMR-MS)[15]、电子鼻传感技术(EN)[16,17]、激光光谱(LS)[18]、选择离子流动管质谱(SIFT-MS)[19]和质子转移反应质谱(PTR-MS)[20]等技术. 电喷雾萃取电离质谱(EESI-MS)技术是一种新型直接质谱分析方法[21,22]，可在无需样品预处理条件下对复杂机体进行直接快速检测，其检测灵敏度高、响应速度快，能够实现呼出气体中痕量挥发性有机化合物的原位、实时、在线分析[23~25]. 然而，由于呼出气体的有效存储时间短，难以在短时间内完成大量样本的收集与检测. 在数据统计分析方面，为保证结果的可靠性，通常对样本量有一定要求. 多批次样本虽然可以获取足够的数据量，但由于人体呼出气体受环境空气影响较大，不同批次(不同时间或地点)收集的样本其检测结果存在较大的差异. 若直接将不同批次样本数据合并成一个大数据矩阵，采用代谢组学中常用的主成分分析(PCA)[26]方法或偏最小二乘分析(PLS)[27]方法对其进行统计分析，由于批次间差异信息的干扰，通常很难准确提取出有用的特征信息.

代谢组学数据往往非常复杂，因此数据处理已经成为代谢组学研究中的关键技术和瓶颈之一. 不同批次获取的数据存在批次间的变异，致使不同批次的数据难以集成. 虽然有一些样本归一化方法已经被开发来解决批次间差异的问题，例如常数和归一化法[28]、内标法[29]、质量控制法[30]和基于方差的归一化法[31]等，但是每种方法都有其优点和缺点. 多块偏最小二乘分析(Multi-block PLS，MB-PLS)

收稿日期: 2015-10-27; 网络出版日期: 2016-03-18.

基金项目: 江西省重大科技创新研究项目(批准号: 20124ACB00700)、长江学者和创新团队发展计划项目(批准号: IRT13054)和国家自然科学基金(批准号: 21265002)资助.

联系人简介: 邬小萍，女，教授，主要从事传染病临床研究. E-mail: wuxiaoping2823@aliyun.com

是近年来广泛应用的一种基于监督的多块数据分析方法[32]，该方法能利用数据块之间的关联性将数据块进行有效整合，并对数据中相关特征信息进行提取. 因其结果是由多个数据块综合分析得到，故相比于单个数据块的分析结果为更为全面、准确. 本文利用 EESI-MS 技术获取了 4 批次肝衰竭患者和健康志愿者呼出气体的代谢组学数据，根据各批次数据间“变量空间”一致的特点，采用相应的 MB-PLS 方法对其进行整合建模，并与传统的 PLS 方法进行比较.

1　多批次数据的多元统计分析

代谢组学数据分析中的多块数据问题通常包括 2 类（“样本空间”相同但“变量空间”不同，“变量空间”相同但“样本空间”不同）（图 1）. 对于采用 EESI-MS 技术获取的各批次呼出气体代谢组学数据，虽然不同批次的样本不同（即“样本空间”不同），但是所检测的代谢物变量是一致的（即“变量空间”相同），与图 1（B）描述问题相等.

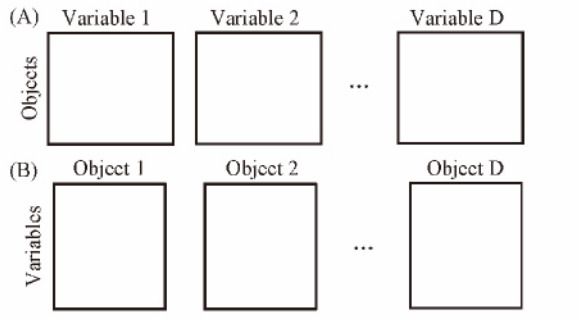

Fig. 1　Two typical multi-block problems

(A) The objects are in common, but the variables measured on these objects are different; (B) the variables are in common, but the objects are different.

若矩阵 $X^b(b=1,2,\cdots,B)$ 表示第 b 批次获取的数据，矩阵每一行对应 1 个样本，所有 B 个数据矩阵均具有相同的变量空间；$y^b(b=1,2,\cdots,B)$ 用来表示第 b 批次样本的类别信息. 采用 MB-PLS 将 B 个批次数据进行分析，为方便起见，将 X^b 和 y^b 进行中心化处理，分别记作 X_0^b 和 y_0^b，MB-PLS 方法的目标函数如下：

$$\max \sum_{b=1}^{B} [\mathrm{Cov}(t_k^b, y_{k-1}^b)]^2 \tag{1}$$

在 PLS 模型中，变量投影重要性指标 VIP（Variable importance in the projection）用于评估各变量在模型中的重要性. 对于 K 个成分的 MB-PLS 模型，变量 i 在该模型中的投影重要性指标 VIP 定义如下：

$$\mathrm{VIP}_i = \sqrt{d\sum_{h=1}^{k}\left[\sum_{b=1}^{B} R^2(y_0^b, t_h^b)\right]p_{ih}^2 \Big/ \sum_{h=1}^{k}\sum_{b=1}^{B} R^2(y_0^b, t_h^b)} \tag{2}$$

式中：d 为数据矩阵的变量维数；$R^2(y_0^b, t_h^b)$ 为第 b 个数据块的类别矢量 y^b 与其对应第 k 个成分的得分矢量 t_h^b 的相关系数的平方；P_{ih} 为第 h 个负载对应的第 i 个变量的权重.

2　实验部分

2.1　仪器与试剂

EESI 离子源（东华理工大学研制）[20,22]；LTQ-Orbitrap-XL 高分辨质谱仪（美国 Finnigan 公司），配有 Xcalibur 数据处理系统；T2PV/L 型 5L-Tedlar® 采样袋（大连德霖气体包装有限公司）；甲醇（色谱纯，SK Chemicals 公司）.

2.2　研究对象与分组

在遵守医学道德准则的相关规定下，分 4 个批次收集就诊于南昌大学第一附属医院感染科的共 35 例肝衰竭患者和 35 例健康志愿者的呼出气体. 肝衰竭患者年龄均在 38 ~ 65 岁之间，排除同时患有糖尿病、脂肪肝、酒精肝、自身免疫性肝病、肾病、呼吸系统疾病、未控制的精神病及活动性感染等疾病的患者；健康志愿者均来自患者家属和医院医务人员，既往无肝病病史，无烟酒嗜好，年龄在 28 ~ 55 岁之间. 各批次样本收集的具体信息如表 1 所示.

Table 1　Four batches of exhaled breath sample

Batch	Date of collection	Liver failure patient	Healthy volunteer	Batch	Date of collection	Liver failure patient	Healthy volunteer
1	2014-07-08	9	4	3	2014-08-25	8	11
2	2014-08-14	9	4	4	2014-09-11	9	16

2.3 质谱条件及呼出气体的收集与检测

高分辨质谱仪在正离子检测模式下工作，扫描范围 m/z 50～700；离子传输管温度 100 ℃；喷雾电压 3.5 kV；雾化气(N_2)压力 1.2 MPa；萃取剂为纯甲醇，流速 5 μL/min；气体样品通过转子流量计控制流速为 800 mL/min；其它条件由 LTQ-Orbitrap-XL 系统自动优化得到. EESI 离子源 2 个毛细管喷雾口之间距离 1～2 mm，夹角为 60°，到质谱进样口的距离为 5 mm，详细参见文献[20,22]. 在高分辨质谱扫描模式下，一级质谱质量分辨率 R = 60000.

用 5L-Tedlar® 采样袋收集呼出气体样本，采样袋在使用前以纯净氮气冲洗 3 次. 所有受试者在采样前 10 h 内禁食、禁烟、禁酒，采样前 24 h 内禁止食用辛辣物，晨起后仅以清水漱口. 受试者在通风条件良好的环境下呼吸 30 min 以上，静息状态下向采样袋内深呼气，直至采样袋充满为止. 收集后在 3 h 内完成 EESI-MS 检测，并获得相应的代谢指纹图谱. 各个批次疾病组(肝衰竭患者)和对照组(健康志愿者)的代谢指纹图谱如图 2 所示.

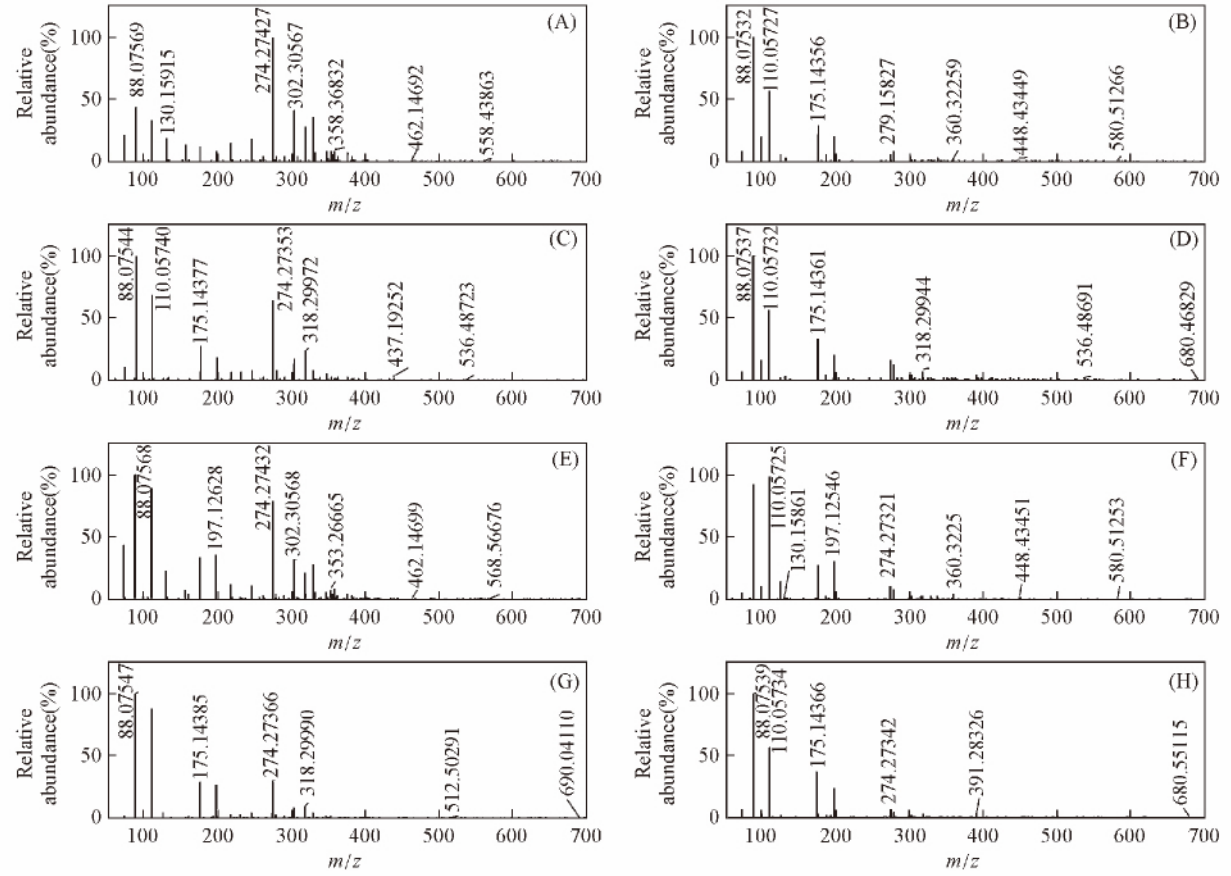

Fig. 2 EESI-MS spectra of exhaled breath from liver failure patients and healthy volunteers

(A)—(D) MS data of exhaled breath from liver failure patients; (E)—(H) MS data of exhaled breath from healthy volunteers. Each row represents a batch.

3 数据分析与讨论

3.1 数据预处理

对于原始谱图数据，设置噪声阈值，将谱图中小于该阈值的变量删除，同时剔除 m/z 88.0754，110.0573，175.1437 和 197.1254 处袋子底物[33]的信号峰，最终可用矩阵 $X^1 = [X^{D_1}_{9\times1035}; X^{C_1}_{4\times1035}]$，$X^2 = [X^{D_2}_{9\times1035}; X^{C_2}_{4\times1035}]$，$X^3 = [X^{D_3}_{8\times1035}; X^{C_3}_{11\times1035}]$ 和 $X^4 = [X^{D_4}_{9\times1035}; X^{C_4}_{16\times1035}]$ 分别描述 4 个批次获取的呼出气体代谢组学数据，其中上标“D”和“C”分别表示疾病组和对照组. 采用概率商归一化法(PQN)和单位方差法(UV)分别对数据矩阵 X^b(b = 1,2,3,4)进行行处理和列处理. 处理后数据矩阵仍用 X^b(b = 1,2,3,4)表示.

3.2　数据统计分析与建模

设置类别矢量 y^b($b=1,2,\cdots\cdots4$) ，其中“1”表示疾病组，“0”表示对照组. 采用 MB-PLS 方法对 4 个批次数据 X^b($b=1,2,3,4$) 进行统计建模. 通过 7-fold 交叉验证法确定 MB-PLS 模型的最优成分数为 2 个. 图 3(A) 为 MB-PLS 模型前 2 个成分对应的得分图. 图中每 1 个点对应 1 个样本，不同批次样本用不同的图形进行区分(如，图形△代表第 1 批次样本) ；蓝色和红色分别表示疾病组与健康组样本. 可见，疾病组与对照组样本之间存在明显的分组趋势，根据公式 X 计算类别矢量的回归值，利用分类准确率(CA) ，即正确分类的样本数除以总样本个数，来描述样本可分性，计算结果 CA = 0.93. 图 3(A) 中，相同类别的不同批次样本相互混叠，无明显分组趋势，表明模型中未提取出各批次数据块之间的差异信息，因此批次间的差异信息并未对该模型产生干扰.

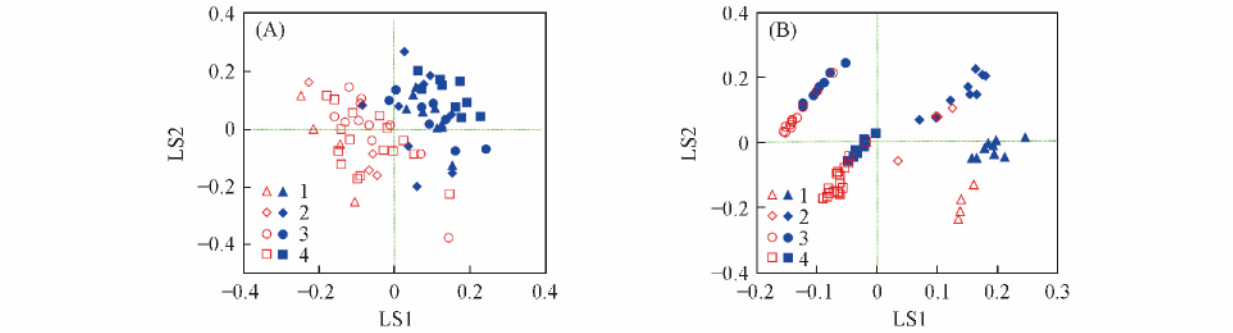

Fig. 3　Scores plot of MB-PLS model(A) and PLS model(B) by the four batches of data respectively

The batches labeled by different graphics, the hollow red graphics and solid blue graphics for the disease group and control group respectively.

为与传统 PLS 方法进行对比，将 4 个批次数据串联成 1 个大的数据矩阵，即 $X=(X^1\ X^2\ X^3\ X^4)^{\mathrm{T}}$，并采用 PLS 方法进行分析，模型前 2 个成分的得分图如图 3(B) 所示. 在图 3(B) 中，虽然同一批次疾病组样本与对照组样本之间表现出一定的分组趋势，但是样本间批次差异信息在模型中占绝对优势，严重干扰了与疾病相关的特征信息的提取. 在不同批次的样本间，疾病组样本难以与对照组样本区分开来，分类准确率 CA 仅为 0.77.

采用蒙特卡洛交叉验证法(MCCV) [34] 对 MB-PLS 模型进行了验证. 在各批次样本中随机选取 70% 的样本作为训练集用于建立分类模型；剩余的样本作为测试集代入模型中，计算测试样本集的预测值. 重复上述过程 100 次，计算测试集的平均分类准确率 CA_{MCCV} 来评估模型的预测性能. 进一步对疾病组和对照组的 MB-PLS 模型进行置换检验(Permutation test) [35] ，样本类别被随机打乱 100 次，每次利用打乱后的类别矢量来建模，并结合 MCCV 计算预测集的分类准确率 CA_{MCCV}，结果见图 4. 图 4 中，横坐标 $|r|$ 为随机打乱后的类别矢量与原类别矢量的相关系数的绝对值，其中 $|r|=1$ 对应的 CA_{MCCV} 值为利用正确类别信息建立模型的分类准确率. 对于一个鲁棒的模型，当类别信息被打乱，模型预测性能应该比正确类别信息建立的模型预测性能要差. 图 4 中，100 次置换检验的结果相对正确类别计算得到的 CA_{MCCV} 要低，表明疾病组与对照组数据存在差异信息，MB-PLS 模型中提取的差异信息是有效的. 此外，PLS 模型由于受到批次差异信息的干扰，模型的预测能力($CA_{MCCV}=0.72\pm0.08$) 显著低于 MB-PLS 模型($CA_{MCCV}=0.84\pm0.06$) . 综上所述，利用 MB-PLS 对多批次数据进行分类建模，能有效避免批次差异对模型的影响，提取出数据中有用的特征信息.

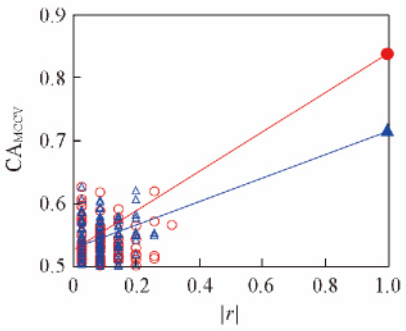

Fig. 4　Model validation results of MB-PLS(○) and PLS(△) respectively

3.3　变量筛选

在上述 MB-PLS 模型中，变量具有很高的维数(1035 个变量) ，数据中只有少部分变量对建立分类模型有贡献. 因此对变量进行了筛选，以有效降低数据的冗余，提高模型的性能.

由模型变量 VIP 的定义可知：变量的 VIP 值越大，该变量在模型中越重要. 利用式(2)计算得到了 MB-PLS 模型中各变量的 VIP 值(见图5). 图5中大部分变量对于该模型并不重要，其对应的 VIP 值非常小(VIP < 1.0)，故可以利用变量的 VIP 值对变量的重要性进行评估，选择 VIP 值大的变量来重新建立分类建模.

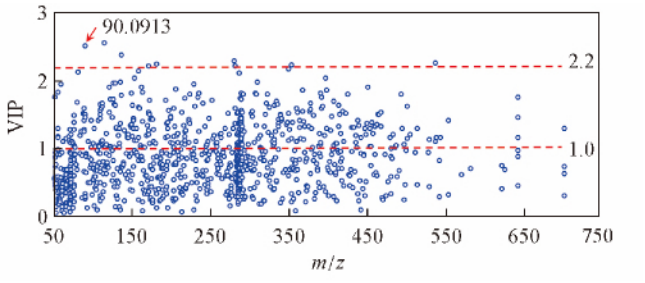

Fig. 5 VIP scores of MB-PLS model

实验中将变量按 VIP 值进行排序，选取其中 VIP 值大于某一阈值的变量构成新的数据子集，进行 MB-PLS 建模. 结合 MCCV 计算 MB-PLS 模型的分类准确率 CA_{MCCV}，结果如图6所示. 图6中模型分类准确率 CA_{MCCV} 随着选取的 VIP 阈值总体表现出先增大后降低的变化趋势. 当 VIP 阈值由0增加到0.8时，大量冗余或者无用的变量被删除，模型的 CA_{MCCV} 急剧增大；当 VIP 阈值由0.8增加到2.2时，由于 VIP 阈值在这一区域变量相对较少，模型 CA_{MCCV} 值变化相对缓慢，尽管某一阶段区域中 CA_{MCCV} 值出现了小幅度的降低，但总体变化趋势仍是不断增大；当 VIP 阈值大于2.2时，可能由于某些有意义的变量被删除，此时模型的 CA_{MCCV} 开始下降. 选取 VIP 阈值为2.2的9个变量用于 MB-PLS 建模，此时模型的分类准确率 CA_{MCCV} 由原来的 0.84 ± 0.06(1035 个变量)提高到 0.96 ± 0.04.

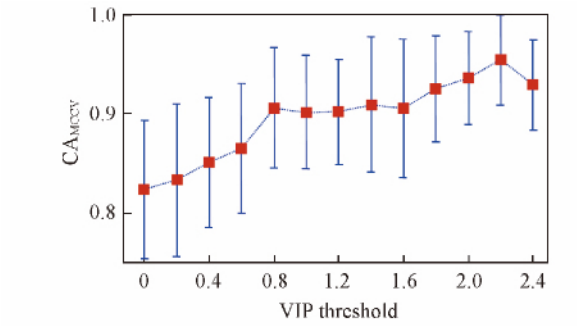

Fig. 6 Variation of the mean CA_{MCCV} of MB-PLS model with the threshold of VIP

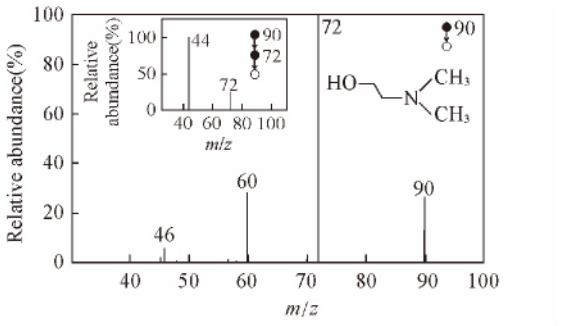

Fig. 7 MS/MS spectrum and proposed structure of *m/z* 90 from the data of liver failure patients and healthy volunteers breath samples

3.4 差异变量分析

为推测差异变量的可能结构，对 VIP 值较大的小分子 *m/z* 90.0913(图5)进行了串联质谱分析. 通过一级质谱结果可知，精确质量数 90.0913 对应的结构式可能为 $C_4H_{12}NO$，与 Xcalibur 数据处理系统给出的理论值误差仅为 −0.45 mg/L. 结合串联质谱结果推测 *m/z* 90 为二甲基乙醇胺($C_4H_{11}NO + H$)，其串联质谱结果及结构式见图7. 在二级质谱中，高峰度的碎片离子 *m/z* 72 $[CH_2{=}CHNH(CH_3)_2]^+$ 可能是母离子 $[HOCH_2CH_2NH(CH_3)_2]^+$(*m/z* 90)丢失 H_2O(*m/z* 18)后形成，而如果丢失 CHOH(*m/z* 30)和 $CHCH_2OH$(*m/z* 44)则分别产生 $[NH(CH_3)_3]^+$(*m/z* 60)和 $[NH_2(CH_3)_2]^+$(*m/z* 46)的碎片离子. 由三级质谱结果可知，母离子丢失 H_2O 后继续丢失 $CH_2{=}CH_2$(*m/z* 28)从而形成 $[CH_2{=}NHCH_3]^+$(*m/z* 44)的离子峰. 碎裂结果与人类代谢组数据库(HMDB)中的结果一致. 组成人体蛋白质的氨基酸-丝氨酸脱羧后形成乙醇胺[36]，而乙醇胺[37]、单甲基乙醇胺和二甲基乙醇胺[38]均是合成胆碱的前体物质. 胆碱是卵磷脂合成的基本原料，卵磷脂是细胞膜和脂蛋白的重要组成成分. 另外，研究[39]发现，肝脏可利用二甲基乙醇胺合成磷脂酰乙醇胺(脑磷脂)，磷脂酰乙醇胺可被 *S*-腺苷甲硫氨酸甲基化合成磷脂酰胆碱，即卵磷脂. 因此，可以确定二甲基乙醇胺是合成卵磷脂的中间体. 人体肝脏内的磷脂合成非常活跃，尤其是卵磷脂的合成. 当肝脏发生病变时，卵磷脂的合成可能受到影响，从而使相应中间产物的含量发生变化. 因此，检测到的肝衰竭患者和健康志愿者呼气中的二甲基乙醇胺差异比较明显. 结果表明，通过 MB-PLS 模型的 VIP 值筛选，可以快速从大量繁杂的数据中发现患者和健康志愿者呼出气体中的差异物质，这些差异物质可能成为区分患者和健康志愿者的生物标

记物.

4　结　　论

采用 EESI-MS 对肝衰竭患者和健康志愿者呼出气体样本进行快速检测，结合 MB-PLS 方法对多批次获取的呼出气体代谢数据进行分析，并与传统的 PLS 方法进行比较. 结果表明，MB-PLS 方法能有效消除批次间差异对统计建模的影响，建立区分疾病组与健康组的分类模型；采用蒙特卡洛交叉验证和排序测试对模型进行验证，发现肝衰竭患者与健康志愿者呼出气体中存在显著的代谢差异. 此外，采用基于 MB-PLS 模型变量 VIP 值的筛选方法，对变量进行筛选，模型交叉验证分类准确率由原来的 0.84 ±0.06 提高到了 0.96 ±0.04，利用该模型能有效区分肝衰竭患者与健康人群. 该项工作有望为不同分析平台、不同样本源获取的多批次代谢组学数据的处理提供一种新的途径和依据. 由于部分差异变量及其与肝衰竭的相关性未能确定，因此仍需要进一步研究.

参　考　文　献

[1] Phillips M., Herrera J., Krishnan S., Zain M., Greenberg J., Cataneo R. N., *J. Chromatogr. B*, **1999**, *729*(1/2), 75—88

[2] Gieger C., Geistlinger L., Altmaier E., de Angelis M. H., Kronenberg F., Meitinger T., Mewes H. W., Wichmann H. E., Weinberger K. M., Adamski J., Illig T., Suhre K., *Plos Genet.*, **2008**, *4*(11), e1000282

[3] Want E. J., Wilson I. D., Gika H., Theodoridis G., Plumb R. S., Shockcor J., Holmes E., Nicholson J. K., *Nat. Protoc.*, **2010**, *5*(6), 1005—1018

[4] Sugimoto M., Wong D. T., Hirayama A., Soga T., Tomita M., *Metabolomics*, **2010**, *6*(1), 78—95

[5] Yuan M., Breitkopf S. B., Yang X. M., Asara J. M., *Nat. Protoc.*, **2012**, *7*(5), 872—881

[6] Carraro S., Rezzi S., Reniero F., Héberger K., Giordano G., Zanconato S., Guillou C., Baraldi E., *Am. J. Respir. Crit. Care. Med.*, **2007**, *175*(10), 986—990

[7] Motta A., Paris D., Melck D., de Laurentiis G., Maniscalco M., Sofia M., Montuschi P., *Eur. Respir. J.*, **2012**, *39*(2), 498—500

[8] Gu H. W., Qi Y. P., Xu N., Ding J. H., An Y. B., Chen H. W., *Chinese J. Anal. Chem.*, **2012**, *40*(12), 1933—1937(顾海威，齐云鹏，许宁，丁健桦，安艳波，陈焕文. 分析化学，**2012**，*40*(12)，1933—1937)

[9] Chen C., Deng L. L., Wei S. W., Gowda G. A. N., Gu H. W., Chiorean E. G., Abu Zaid M., Harrison M. L., Pekny J. F., Loehrer P. J., *J. Proteome Res.*, **2015**, *14*(6), 2492—2499

[10] Gu H. W., Huang Y., Filgueira M., Carr P. W., *J. Chromatogr. A*, **2011**, *1218*(38), 6675—6687

[11] Hanouneh I. A., Zein N. N., Cikach F., Dababneh L., Grove D., Alkhouri N., Lopez R., Dweik R. A., *Clin. Gastroenterol. H.*, **2014**, *12*(3), 516—523

[12] Van Den Velde S., Nevens F., Van Hee P., Van Steenberghe D., Quirynen M., *J. Chromatogr. B. Analyt. Technol. Biomed. Life Sci.*, **2008**, *875*(2), 344—348

[13] Krotoszynski B., Gabriel G., Oneill H., Claudio M. P. A., *J. Chromatogr. Sci.*, **1977**, *15*(7), 239—244

[14] Phillips M., Gleeson K., Hughes J. M. B., Greenberg J., Cataneo R. N., Baker L., McVay W. P., *Lancet*, **1999**, *353*(9168), 1930—1933

[15] Netzer M., Millonig G., Osl M., Pfeifer B., Praun S., Villinger J., Vogel W., Baumgartner C., *Bioinformatics*, **2009**, *25*(7), 941—947

[16] Casalinuovo I. A., Di Pierro D., Coletta M., Di Francesco P., *Sensors*, **2006**, *6*(11), 1428—1439

[17] Roeck F., Barsan N., Weimar U., *Chem. Rev.*, **2008**, *108*(2), 705—725

[18] Skeldon K. D., Mcmillan L. C., Wyse C. A., Monk S. D., Gibson G., Patterson C., France T., Longbottom C., Padgett M. J., *Respir. Med.*, **2006**, *100*(2), 300—306

[19] Storer M., Dummer J., Sturney S., Epton M., *Curr. Anal. Chem.*, **2013**, *9*(4), 576—583

[20] Moser B., Bodrogi F., Eibl G., Lechner M., Rieder J., Lirk P., *Resp. Physiol. Neurobi.*, **2005**, *145*(2/3), 295—300

[21] Pan S. S., Zhao N., Ouyang Y. Z., Huang K. K., Ding J. H., Chen H. W., Yuan L., Wang X. X., *Chem. J. Chinese Universities*, **2013**, *34*(6), 1379—1384(潘素素，赵娜，欧阳永中，黄科科，丁健桦，陈焕文，袁龙，王兴祥. 高等学校化学学报，**2013**，*34*(6)，1379—1384)

[22] Ding J. H., Wang X. X., Zhang H., Pan S. S., Luo M. B., Li J. Q., Chen H. W., *Chem. J. Chinese Universities*, **2011**, *32*(8), 1714—1719(丁健桦，王兴祥，张慧，潘素素，罗明标，李建强，陈焕文. 高等学校化学学报，**2011**，*32*(8)，1714—1719)

[23] Chen H. W., Wortmann A., Zhang W. H., Zenobi R., *Angew. Chem. Int. Ed.*, **2007**, *46*(46), 580—583

[24] Pan S. S., Tian Y., Li M., Zhao J. Y., Zhu L. L., Zhang W., Gu H. W., Wang H. D., Shi J. B., Fang X., Li P. H., Chen

H. W., *Sci. Rep.*, **2015**, *5*, *8725*
[25] Ding J. H., Yang S. P., Liang D. P., Chen H. W., Wu Z. Z., Zhang L. L., Ren Y. L., *Analyst*, **2009**, *134*(10), 2040—2050
[26] Wood C. C., Mccarthy G., *Electroencephalogr. Clin. Neurophysiol.*, **1984**, *59*(3), 249—260
[27] Frank I. E., Kowalski B. R., *Anal. Chim. Acta*, **1984**, *162*, 241—251
[28] Wang W. X., Zhou H. H., Lin H., Roy S., Shaler T. A., Hill L. R., Norton S., Kumar P., Anderle M., Becker C. H., *Anal. Chem.*, **2003**, *75*(18), 4818—4826
[29] Redestig H., Fukushima A., Stenlund H., Moritz T., Arita M., Saito K., Kusano M., *Anal. Chem.*, **2009**, *81*(19), 7974—7960
[30] Jauhiainen A., Basetti M., Narita M., Narita M., Griffiths J., Tavare´ S., *BMC Bioinformatics*, **2014**, *30*(15), 2155—2161
[31] De Livera A. M., Dias D. A., De Souza D., Rupasinghe T., Pyke J., Tull D., Roessner U., McConville M., Speed T. P., *Anal. Chem.*, **2012**, *84*(24), 10768—10776
[32] Wangen L. E., Kowalski B. R., *J. Chemometr.*, **1989**, *3*(1), 3—20
[33] Beauchamp J., Herbig J., Gutmann R., Hansel A., *J. Breath Res.*, **2008**, *2*(4), 046001
[34] Picard R. R., Cook R. D., *J. Am. Stat. Assoc.*, **1984**, *79*(387), 575—583
[35] Lindgren F., Hansen B., Karcher W., Sjostrom M., Eriksson L., *J. Chemometr.*, **1996**, *10*(5/6), 521—532
[36] Levine M., Tarver H., *J. Biol. Chem.*, **1950**, *184*(2), 427—436
[37] Pilgeram L. O., Gal E. M., Sassenrath E. N., Greenberg D. M., *J. Biol. Chem.*, **1953**, *204*(1), 367—377
[38] Duvigneaud V., Chandler J. P., Simmonds S., Moyer A. W., Cohn M., *J. Biol. Chem.*, **1946**, *164*(2), 603—613
[39] Artom C., Crowder M., *Fed. Proc.*, **1949**, *8*(1), 180—181

EESI-MS Detection and Statistical Analysis of Multi-batch of Exhaled Breath Metabolomics Data of Liver Failure Patients†

LI Penghui[1], DENG Lingli[1,2], LUO Jiao[3], LI Wei[3], NING Jing[1], DING Jianhua[1], WU Xiaoping[3*]

(1. *East China University of Technology, Jiangxi Key Laboratory for Mass Spectrometry and Instrumentation, Nanchang* 330013, *China*;
2. *East China University of Technology, Information Engineering College, Nanchang* 330013, *China*;
3. *The First Affiliated Hospital of NanChang University, Nanchang* 330123, *China*)

Abstract In metabolomics studies, the number of samples should be enough to guarantee the reliability of data statistical analysis. The effective storage time of exhaled breath is short, and it is difficult to collect and detect a large number of breath samples in a short time. Combining multi batches of samples may obtain a large data, but usually there is a large variance between batches induced by ambient air varying. In this paper, the exhaled breath data of liver failure patients and healthy volunteers were obtained by high resolution extractive electrospray ionization mass spectrometry(EESI-MS) and then analyzed by multi-block partial least square(MB-PLS). The results were compared with traditional PLS method and showed its strength of removing the variance of batches for modeling. Moreover, we provided a variable selection strategy that based on variable importance in the projection(VIP) of MB-PLS to reduce the redundancy of data and eliminate the effect of non-information variables for modeling, and the performance of MB-PLS model had a great improvement.

Keywords Exhaled breath; Metabolomics; Extractive electrospray ionization mass spectrometry; Multi-block partial least square analysis

(Ed.: D, K)

† Supported by the Jiangxi Major Scientific and Technological Innovation Research Project, China(No. 2012ACB00700), the Program for Changjiang Scholars and Innovative Research Team in University, China(No. IRT13054) and the National Natural Science Foundation of China (No. 21265002).