

基于高光谱成像技术的喷墨打印文件 形成时间鉴别研究

梁 源, 崔 岚*, 杨雪颖

(中国刑事警察学院 刑事科学技术学院, 辽宁 沈阳 110035)

摘要: 利用高光谱成像技术结合机器学习回归, 提出了一种无损确定喷墨打印文件形成时间的方法。采集9台不同品牌、型号的喷墨打印机历时打印样品的高光谱数据, 首先利用主成分分析(PCA)和偏最小二乘回归(PLSR)对高光谱数据进行降维特征提取, 并通过特征提取后各主成分的降维解释率对模型进行初步解释。随后应用套索回归(LASSO)、PLSR、岭回归(RR)和贝叶斯回归(BR)4种模型, 以1:4的比例确定测试集和训练集, 对特征提取后的高光谱数据进行回归。结果表明, 经PLSR特征提取后的LASSO回归、PLAR、RR和BR的决定系数(R^2)均不低于0.99, 且4种回归的平均绝对误差(MAE)、均方误差(MSE)和均方根误差(RMSE)3个指标均接近0。4种经PLSR特征提取后的回归方法的 F 值均足够大且 p 值均小于0.001, 达极显著水平。其中PLSR和BR的回归效果优于LASSO和RR。进一步通过PLSR权重系数解释了主成分与波段的关联, 以明确关键光谱波段的作用。回归效果检测结果的 R^2 均大于0.9, 且显著性检验的 p 值均小于0.05, 说明模型对因变量的解释具有极显著的统计学意义, 同时检测误差在合理范围内。结果显示, 高光谱成像技术能够无损、准确、快速地判断喷墨打印文件的制成时间, 可用于喷墨打印机打印文件形成时间鉴别。

关键词: 高光谱成像; 机器学习; 喷墨打印机; 打印文件形成时间

中图分类号: O657.3; TB9 **文献标识码:** A **文章编号:** 1004-4957(2025)12-2478-08

Hyperspectral Imaging-based Investigation for Identification of the Production Timeline of Inkjet-printed Documents

LIANG Yuan, CUI Lan*, YANG Xue-ying

(College of Criminal Science and Technology, Criminal Investigation Police University of China, Shenyang 110035, China)

Abstract: This study proposes a novel nondestructive methodology for determining the formation time of inkjet-printed documents through the integration of hyperspectral imaging technology with machine learning regression algorithms. Experimental data were collected from chronologically printed samples produced by nine inkjet printers of distinct brands and models. The methodology framework comprises two principal phases: dimensionality reduction of hyperspectral data and subsequent regression modeling. In the initial phase, principal component analysis(PCA) and partial least squares regression(PLSR) were employed to address the high dimensionality inherent in hyperspectral datasets. The interpretation rate of dimensional reduction of principal components derived from these dimensionality reduction techniques were systematically quantified, providing preliminary interpretability for the model architecture. This phase effectively condensed the hyperspectral information while preserving critical temporal features associated with document aging. Four advanced regression methodologies were subsequently implemented on the dimensionality-reduced data: least absolute shrinkage and selection operator regression(LASSO), partial least squares regression(PLSR), ridge regression(RR), and Bayesian regression(BR). The dataset was partitioned into training and testing subsets at a 1:4 ratio to ensure robust model validation. Quantitative evaluation metrics revealed exceptional model performance across all regression approaches. The coefficient of determination(R^2) values for LASSO-PLSR, PLSR, RR, and BR models consistently approached unity($R^2 > 0.99$),

收稿日期: 2025-03-26; 修回日期: 2025-05-19

基金项目: “十三五”国家重点研发计划资助项目(2016YFC08007005); 公安部科技强警基础工作计划项目(2022JC03); 中国刑事警察学院研究生创新能力提升项目(2023YCD07)

* 通讯作者: 崔 岚, 硕士, 教授, 研究方向: 文件检验, E-mail: cuilan0605@126.com

网络首发时间: 2025-08-18

indicating near-perfect explanatory power. Furthermore, all models demonstrated minimal error metrics: mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) values were all proximate to zero, confirming high predictive accuracy. Statistical significance testing yielded F -statistic values of substantial magnitude ($p < 0.001$) for all PLSR-based regression models, confirming the exceptional statistical validity of the results. Comparative analysis revealed superior performance characteristics in PLSR and BR models relative to LASSO and RR implementations. Further explain the relationship between principal components and spectral bands through the PLSR weight coefficients to clarify the role of key spectral bands. External validation using independent hyperspectral datasets confirmed the models' generalizability, with R^2 values exceeding 0.9 ($p < 0.05$) across all test scenarios. Error analysis indicated that residual variations fell within acceptable methodological thresholds, confirming measurement reliability. This investigation establishes that hyperspectral imaging technology exhibits significant potential for forensic document analysis, particularly in temporal authentication of inkjet-printed materials. The implemented machine learning framework successfully decouples temporal signatures from complex spectral data while maintaining complete non-invasiveness. The method proposed in this article provides a new solution for the problem of dating documents, offering substantial improvements over conventional destructive chemical analysis methods.

Key words: hyperspectral imaging; machine learning; inkjet printer; timeline of inkjet-printed documents

文件检验作为刑事科学技术体系的重要组成部分,能够为确定文件与案件事实、文件与当事人或嫌疑人之间的关系提供证据支持。随着科学技术的发展,喷墨打印机以其操作简便、价格低廉的特点被广泛使用,对喷墨打印文件物证的检验在法庭科学实践中越来越常见。其中,喷墨打印文件的物证形成时间鉴别是检验鉴定工作中面临的难点问题之一。当前对于喷墨打印文件形成时间的研究还处于起步阶段,王千羽等^[1]利用气相色谱-质谱联用法对墨迹中的挥发溶剂进行检验,通过测定样品提取液中挥发性溶剂含量判定物证是否为新打印形成,但此方法破坏样品且仅考虑4份喷墨打印样品在半年时间跨度下的预测,预测结果中 R^2 的最大值仅为0.95。鉴于此,当前亟需对喷墨打印文件形成时间鉴别问题开展进一步研究。

高光谱成像技术(HSI)是一种生成光谱变化空间的技术^[2],与普通成像技术相比,具有“图谱合一”的特性。该技术将光谱学和计算机视觉集成在一个系统中,通过扫描样品可获得数百个光谱带,每一个像素点都包含丰富的空间和光谱信息,可同时获取被检测物体的物理特征和化学结构^[3]。高光谱成像技术的核心优势在于通过非接触的方式获取数据,能够实现无损检测,此外还可以获取空间分布的连续光谱信息,捕捉微小的化学变化,扫描得到的数据更有利于进一步结合机器学习进行分析。目前,高光谱成像技术已经被广泛应用于农产品质量检测^[4]、医学组织成像^[5]、文物鉴定^[6]等领域。在文件检验领域,高光谱成像技术得到越来越广泛的应用。目前,有研究通过光谱特征区分不同类型的墨水^[7],或通过光谱差异识别伪造文件中的添改痕迹^[8],以及通过光谱成像技术进行朱墨时序鉴定^[9]、印章油墨类型快速识别^[10]、黑色签字笔种类识别^[11]、鉴别激光打印墨粉种类^[12]等。以上研究成果表明,高光谱成像技术在文件物证检验工作中具有应用价值。

本文利用高光谱成像仪对9台不同品牌型号的喷墨打印机两年内的打印样品进行检测,采用主成分分析(PCA)和偏最小二乘回归(PLSR)对高光谱数据进行降维特征提取。随后使用套索回归(LASSO)、偏最小二乘回归(PLSR)、岭回归(RR)和贝叶斯回归(BR)4种模型对特征提取后的高光谱数据进行回归。对不同品牌型号喷墨打印的样品分别进行回归建模,以得到可以鉴别喷墨打印文件形成时间的模型,为喷墨打印文件时间鉴别提供一种快速、无损的检验方法。

1 实验部分

1.1 仪器和材料

HyScan WZ-HS-VN 中科谱光物证高光谱成像仪,由成像光谱仪、自动扫描平台以及微孔陶瓷真空吸附载物台组成;仪器的控制电脑预装物证高光谱成像仪数据采集软件和物证高光谱仪数据分析软件;

高光谱数据采集由 ENVI 软件完成;使用 python3.13 版本的解释器进行编码和数据分析。

从市面上常见的 3 个品牌(惠普、佳能、爱普生)中,各选取 3 台不同型号喷墨打印机所使用的黑色墨水,打印机型号和原装黑色墨水型号见表 1,使用亚太森博品牌 A4 型号 75 g/cm² 复印纸作为承印纸张,采集 9 台打印机两年内每月制作的样品;每台打印机 22 份样品,共计 198 份样品,所有样品均在相同温度、湿度下避光保存。

表 1 喷墨打印机型号和原装墨水型号
Table 1 Inkjet printer models and genuine ink models

HP		Canon		Epson	
Printer model	Original ink model	Printer model	Original ink model	Printer model	Original ink model
1212	805BK	E478	BK-48	L4266	002BK
2132	803BK	IX6780	PGBK	L8188	014BK
8210	955BK	E500	BK-88	L7188	106PGBK

1.2 数据采集

高光谱成像仪扫描范围设置为 A4 纸,光谱波长范围为 400~1 000 nm,平均间隔为 1.6 nm,共获取 372 个波段的光谱数据。在每台打印机的 22 份样品不同区域分别提取 5 个 ROI 点位,每台打印机获得 110 条数据,共 990 条高光谱数据。依据样品打印时间分类保存。

利用 Python 3.13 版本解释器计算每台打印机每份样品打印日期距离高光谱检测日期的天数,作为每条数据因变量,并记录保存。

1.3 实验设计

利用高光谱成像仪采集光谱数据。采用 PCA 与 PLSR 方法,对所获取的高光谱数据进行降维特征提取,将其转化为 10~20 维的数组,得到可以反映不同样品特征的关键波段。再结合 LASSO 回归、PLSR、RR 和 BR 4 种机器学习回归模型,以 1:4 的比例确定测试集和训练集,数据分组由回归算法随机划分以保证训练集与测试集的样本分布具有一致性,避免因数据划分偏差导致模型评估结果失真。以决定系数(R^2)、平均绝对误差(MAE)、均方误差(MSE)和均方根误差(RMSE)作为判断拟合效果参数,筛选最佳回归模型。

1.4 算法模型介绍

对光谱数据进行降维特征提取的算法包括 PCA 和 PLSR 降维。PCA^[13]降维算法是常用的无监督学习方法。通过数据标准化、计算协方差矩阵、求解特征值与特征向量,并选取前 k 个主成分构成投影矩阵,完成数据降维。PLSR^[14]降维是多变量数据分析方法,对包含多自变量与多因变量的数据进行建模分析,同时考虑自变量之间以及自变量和因变量之间的相关性,进而从变量系统中提取正交成分元素组合。

喷墨打印文件形成后,墨水的表层结构和物质组成均发生变化。高光谱成像技术具有“图谱合一”的特性,喷墨打印样品的高光谱数据既可以反映样品的表层结构,也可以反映样品墨水的物质组成,能够直观反映喷墨打印文件的结构和物质随时间的变化,基于这种由多种因素形成的变化,本文通过多元线性回归模型对降维后得到的数据进行回归分析。

可对特征提取后的光谱数据进行回归的模型包括 LASSO 回归、PLSR、RR 和 BR。LASSO^[15]回归通过在传统线性回归模型的目标函数中添加一个惩罚项,在回归系数的绝对值之和小于一个常数的约束条件下,使残差平方和最小化;PLSR^[16]通过将预测变量和观测变量投影到新空间,寻找线性回归模型,同时兼顾 PCA 和典型相关分析(CCA),解决了自变量共线性问题;RR^[17]通过在普通最小二乘法(OLS)的目标函数中加入一个 L_2 正则化项,即岭惩罚项,使回归系数向零收缩,从而达到克服多重共线性的目的,同时引入岭迹的概念,用于二维展示非正交性的影响;BR^[18]则将线性模型的参数视为随机变量,通过模型参数的先验分布来计算其后验分布,能够更好地处理模型中的不确定性。

使用 R^2 、MAE、MSE、RMSE 评估 4 种回归模型的拟合优度。 R^2 取值范围为 0~1,其值越接近 1,表明模型对因变量变异的解释能力越强,拟合效果越优;MAE、MSE 及 RMSE 的数值越小拟合效果越好,其中 MAE 能稳定反映预测误差的实际水平且不受异常值方向影响;MSE 因对误差进行平方运算而对异常值更敏感,可突出较大误差的影响;RMSE 作为 MSE 的平方根,与原始数据单位一致,更便于直观理解误差大小。通常而言,当 R^2 接近 1 且 MAE、MSE、RMSE 均处于较小水平时,模型的拟合优度更佳。计算公式见式(1)~(4)。以上算法均基于 Python 软件实现。

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (1)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4)$$

式中, SSR 为回归平方和, SSE 为残差平方和, SST 为总平方和; n 为样本量, y_i 为实际值, $\hat{y}_i$ 为预测值。

为了验证回归模型整体的统计显著性, 本文提出以下假设进行显著性检验:

零假设(H_0): 所有自变量回归系数为零(模型无解释能力);

备择假设(H_1): 至少有一个自变量回归系数不为零(模型具有统计显著性)。

模型的显著性通过 F 检验进行验证, 计算公式见式(5):

$$F = \frac{MSR}{MSE} \quad (5)$$

式中, MSR 为回归均方, MSE 为残差均方。显著性水平设为 $\alpha=0.05$ 。

2 结果与讨论

2.1 光谱特征分析

9 台喷墨打印机打印的历时样品的高光谱图如图 1 所示。 X 轴代表不同波段, Y 轴代表文件打印日期距高光谱检测日期的天数, Z 轴代表不同波段的波长。

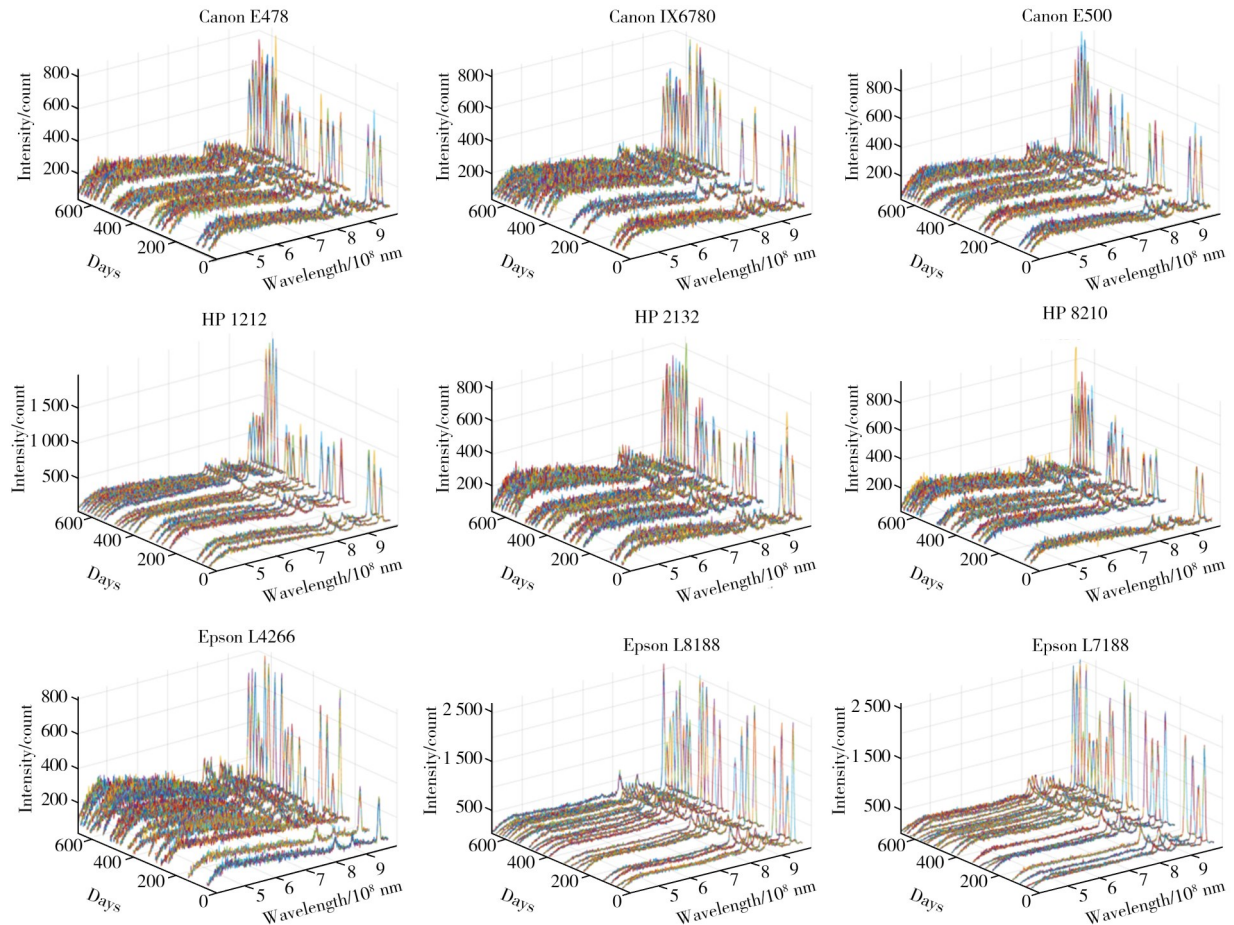


图1 不同品牌型号喷墨打印机历时打印样品的高光谱图

Fig. 1 Hyperspectral images of samples printed by different brands and models of inkjet printers

The band data in the figure have been scaled by multiplying with 10^6 , the original unit is reflectance(unit: parts per million, i. e., 10^{-2}), and the axis is labeled as "Reflectance($\times 10^6$)" after scaling

整体而言, 9 台喷墨打印机打印文件样品的高光谱图像总体趋势相似, 同一台打印机不同时期打印的文件样品的高光谱图走势基本一致。400~500 nm 波段曲折且有缓慢上升, 500~800 nm 波段呈平缓且略下降趋势, 800~950 nm 呈现两次小幅度波峰, 而后在 960 nm 波段陡然上升并在 990 nm 波段附近形成波峰。

由于高光谱数据的冗余性, 直接对其进行回归无法达到理想效果, 因此, 本文在后续回归前先对数据进行降维特征提取处理。

2.2 数据特征提取与效果评价

分别利用 PCA 和 PLSR 算法对高光谱数据进行降维特征提取后得到主成分, 各台打印机打印样品高光谱数据的不同降维算法合计降维解释率见表 2。由表 2 可知, 两种降维方法处理后的数据降维解释率均在 60% 以上, 能够反映原始数据信息。为保证降维解释率在 60% 以上, 经过不同主成分数测试, 将 PCA 降维主成分数设定为 10, PLSR 降维主成分数设定为 15。

2.3 高光谱数据回归结果与评价

使用 LASSO 回归、PLSR、RR 和 BR 四种模型对喷墨打印墨水历时数据进行回归, 基于不同的降维特征提取与回归算法组合, 对不同品牌、型号打印机打印的样品数据进行回归, 为防止回归结果过拟合, 运行回归代码时分别输出训练集和测试集的 R^2 、MAE、MSE、RMSE 四项拟合优度指标, 训练集与测试集 R^2 的差值小于 0.1 且 MAE、MSE 和 RMSE 数值相差在 20% 以内, 即初步判定不存在过拟合情况。同时对模型进行显著性检验, 根据本文所需解释内容, 此处仅报告 F 值和 p 值。9 台喷墨打印机 4 种回归方法的结果见表 3~表 6。

表 2 不同喷墨打印样品的高光谱数据降维解释率

Table 2 Interpretation rate of dimensional reduction of hyperspectral data printed by different inkjet printers

Printer model	PCA/%	PLSR/%
Canon E478	64.28	63.76
Canon IX6780	69.06	68.26
Canon E500	64.91	64.68
HP 1212	75.46	74.77
HP 2132	69.59	68.96
HP 8210	64.10	63.24
Epson L4266	80.39	79.39
Epson L8188	75.57	74.83
Epson L7188	87.22	86.53

表 3 LASSO 回归结果汇总

Table 3 Summary of LASSO regression results

Printer model	Dimensionality reduction method	Whether there is overfitting	R^2	MAE	MSE	RMSE	F	p
Canon E478	PCA	No	0.49	109.64	17 186.40	131.10	7.22	0.00
	PLSR	No	0.99	1.99	6.57	2.56	866 775.35	0.00
Canon IX6780	PCA	Yes	0.18	134.62	27 977.71	167.27	1.15	0.34
	PLSR	No	1.00	0.33	0.17	0.41	849 846.89	0.00
Canon E500	PCA	No	0.76	86.29	12 090.66	109.96	20.68	0.00
	PLSR	No	1.00	0.26	0.10	0.32	1 786 596.89	0.00
HP 1212	PCA	No	0.39	114.11	20 773.46	144.13	7.67	0.00
	PLSR	No	1.00	0.35	0.18	0.43	853 518.73	0.00
HP 2132	PCA	No	0.30	125.17	24 842.32	157.61	2.46	0.02
	PLSR	No	1.00	0.35	0.19	0.44	850 114.27	0.00
HP 8210	PCA	No	0.49	98.19	17 593.69	132.64	12.30	0.00
	PLSR	No	1.00	0.27	0.13	0.35	1 129 398.20	0.00
Epson L4266	PCA	No	0.22	122.64	22 797.64	150.99	2.30	0.04
	PLSR	No	1.00	0.35	0.18	0.43	682 200.34	0.00
Epson L8188	PCA	No	0.47	109.15	19 996.20	141.41	6.77	0.00
	PLSR	No	1.00	0.36	0.19	0.44	956 137.14	0.00
Epson L7188	PCA	Yes	0.32	135.33	26 882.48	163.96	3.95	0.00
	PLSR	No	1.00	0.55	0.57	0.75	335 085.03	0.00

表 4 偏最小二乘回归结果汇总

Table 4 Summary of PLSR regression results

Printer model	Dimensionality reduction method	Whether there is overfitting	R^2	MAE	MSE	RMSE	F	p
Canon E478	PCA	No	0.48	112.72	17 616.97	132.73	79.97	0.00
	PLSR	No	1.00	0.08	0.01	0.10	69 032 733.49	0.00
Canon IX6780	PCA	Yes	0.20	133.67	27 321.52	165.30	19.87	0.00
	PLSR	No	1.00	0.10	0.02	0.12	33 534 964.94	0.00

(续表4)

Printer model	Dimensionality reduction method	Whether there is overfitting	R^2	MAE	MSE	RMSE	F	p
Canon E500	PCA	No	0.78	83.08	11 342.59	106.50	86.27	0.00
	PLSR	No	1.00	0.09	0.02	0.12	31 320 076.82	0.00
HP 1212	PCA	Yes	0.41	115.36	20 185.61	142.08	57.21	0.00
	PLSR	No	1.00	0.11	0.02	0.15	24 650 383.22	0.00
HP 2132	PCA	No	0.33	125.34	23 905.80	154.62	40.27	0.00
	PLSR	No	1.00	0.10	0.02	0.13	34 748 657.85	0.00
HP 8210	PCA	No	0.50	98.13	17 113.41	130.82	12.41	0.00
	PLSR	No	1.00	0.04	0.002	0.05	182 953 573.73	0.00
Epson L4266	PCA	Yes	0.24	120.80	22 143.90	148.81	25.29	0.00
	PLSR	No	1.00	0.11	0.02	0.14	23 048 933.80	0.00
Epson L8188	PCA	No	0.50	109.53	19 182.61	138.50	41.84	0.00
	PLSR	No	1.00	0.15	0.04	0.19	17 618 604.56	0.00
Epson L7188	PCA	Yes	0.34	134.30	26 195.36	161.85	44.67	0.00
	PLSR	No	1.00	0.49	0.42	0.65	1 523 134.32	0.00

表5 岭回归结果汇总

Table 5 Summary of RR regression results

Printer model	Dimensionality reduction method	Whether there is overfitting	R^2	MAE	MSE	RMSE	F	p
Canon E478	PCA	No	0.49	111.17	17 337.61	131.67	6.10	0.00
	PLSR	No	1.00	0.20	0.06	0.25	2 790 639.71	0.00
Canon IX6780	PCA	Yes	0.14	137.02	29 385.64	171.42	0.32	0.97
	PLSR	No	1.00	0.22	0.07	0.28	1 861 091.17	0.00
Canon E500	PCA	No	0.78	82.58	11 299.45	106.30	13.93	0.00
	PLSR	No	1.00	0.41	0.22	0.47	555 543.99	0.00
HP 1212	PCA	Yes	0.40	113.74	20 253.71	142.32	4.09	0.00
	PLSR	No	1.00	0.24	0.08	0.28	1 942 860.30	0.00
HP 2132	PCA	No	0.33	123.77	23 996.52	154.91	2.89	0.00
	PLSR	Yes	1.00	0.22	0.07	0.27	2 201 341.47	0.00
HP 8210	PCA	No	0.49	100.27	17 387.81	131.86	4.30	0.00
	PLSR	No	1.00	0.23	0.08	0.29	1 456 464.94	0.00
Epson L266	PCA	No	0.18	127.38	24 087.38	155.20	0.47	0.90
	PLSR	No	1.00	0.21	0.06	0.25	1 952 392.88	0.00
Epson L8188	PCA	Yes	0.49	108.21	19 388.36	139.24	6.02	0.00
	PLSR	No	1.00	0.25	0.09	0.30	2 010 034.22	0.00
Epson L7188	PCA	Yes	0.34	134.97	26 236.87	161.98	3.27	0.00
	PLSR	No	1.00	0.53	0.48	0.70	394 143.05	0.00

表6 贝叶斯回归结果汇总

Table 6 Summary of BR regression results

Printer model	Dimensionality reduction method	Whether there is overfitting	R^2	MAE	MSE	RMSE	F	p
Canon E478	PCA	No	0.48	112.00	17 511.27	132.33	5.63	0.00
	PLSR	No	1.00	0.08	0.01	0.10	25 044 597.78	0.00
Canon IX6780	PCA	No	0.15	136.60	29 159.28	170.76	0.37	0.96
	PLSR	No	1.00	0.09	0.01	0.12	10 659 111.38	0.00
Canon E500	PCA	No	0.78	83.31	11 473.61	107.11	12.43	0.00
	PLSR	No	1.00	0.01	0.000 1	0.01	56 155 851 558.30	0.00
HP 1212	PCA	No	0.39	115.39	20 962.36	144.78	3.29	0.00
	PLSR	No	1.00	0.11	0.02	0.15	7 237 768.44	0.00
HP 2132	PCA	No	0.27	128.13	26 172.35	161.78	1.86	0.07
	PLSR	No	1.00	0.10	0.02	0.13	10 431 790.09	0.00
HP 8210	PCA	No	0.49	101.46	17 601.56	132.67	3.85	0.00
	PLSR	No	1.00	0.01	0.000 3	0.02	391 836 750.27	0.00
Epson L4266	PCA	No	0.20	125.28	23 368.66	152.87	0.72	0.70
	PLSR	No	1.00	0.11	0.02	0.13	7 117 526.81	0.00
Epson L8188	PCA	Yes	0.49	108.19	19 501.64	139.65	5.64	0.00
	PLSR	No	1.00	0.15	0.04	0.19	5 206 676.47	0.00
Epson L7188	PCA	Yes	0.32	138.21	26 891.17	163.99	2.27	0.02
	PLSR	No	1.00	0.49	0.43	0.65	446 051.44	0.00

结果表明, 9 份喷墨打印历时样品光谱数据在两种降维特征提取算法后, 不仅减少了特征数量、降低了计算成本, 且经 PLSR 特征提取后数据的回归拟合优度明显高于 PCA 特征提取。推测其原因在于 PCA 与 PLSR 方法的原理不同: PCA 作为无监督降维, 未利用因变量信息, 导致主成分与时间关联性弱且易过拟合; 而 PLSR 通过有监督降维直接关联时间变化特征, 优化因子数后可显著提升预测精度。因此本文选择 PLSR 作为降维特征提取方法。

4 种经 PLSR 特征提取后的回归方法中, Canon E478 喷墨打印历时样品的光谱数据通过 LASSO 回归的 R^2 为 0.99, 其余喷墨打印机的 R^2 均为 1.00, MAE、MSE 和 RMSE 三个指标均接近 0; 9 台喷墨打印机样品经 RR 回归的 R^2 均为 1.00, 且 MAE、MSE 和 RMSE 三个指标均接近 0, 但其中超过 74% 的指标大于 0.1; PLSR 和 BR 的 R^2 均为 1, 且 MAE、MSE 和 RMSE 三个指标较岭回归更加接近 0。

根据显著性检验结果可知, 在 72 种“降维特征提取-回归”方法组合中, 超过 95% 的组合 p 值小于 0.05, 达极显著水平, F 检验结果拒绝了零假设(H_0), 表明模型对因变量的解释能力具有极显著的统计学意义。同时经 PLSR 特征提取后回归模型的 F 值较经 PCA 特征提取后更大, 对因变量的解释能力更强。

综上所述, 4 种模型经 PLSR 特征提取后都能达到比较理想的回归效果, 实现了对不同型号喷墨打印文件形成时间的精准鉴别, 4 种回归模型中 PLSR 和 BR 的回归效果优于 LASSO 回归和岭回归。为明确关键光谱波段的作用, 通过 PLSR 权重系数进一步解释主成分与波段的关联, 结果见表 7。

表 7 不同喷墨打印样品关键光谱波段的 PLSR 权重系数

Table 7 PLSR weight coefficients of key spectral bands for different inkjet printing samples

Serial number	Key spectral band (PLSR weight coefficient)								
	Canon E478	Canon IX6780	Canon E500	HP 1212	HP 2132	HP 8210	Epson L4266	Epson L8188	Epson L7188
KSB1	603.58	986.52	793.43	713.92	475.39	871.32	915.13	947.58	975.17
	(-25.38)	(-31.39)	(-14.79)	(-33.35)	(27.91)	(20.81)	(39.10)	(-30.35)	(-60.93)
KSB2	645.77	981.66	936.22	702.56	687.96	853.47	907.01	846.98	400.75
	(-24.18)	(29.32)	(-13.39)	(29.46)	(24.95)	(-20.27)	(30.01)	(-27.55)	(47.52)
KSB3	785.32	846.98	770.71	533.80	855.09	416.97	720.41	795.05	540.29
	(-23.32)	(-29.09)	(13.01)	(-28.12)	(-24.90)	(19.52)	(29.25)	(26.90)	(-44.38)
KSB4	819.39	631.16	809.66	608.45	993.02	694.45	473.77	996.26	991.39
	(-23.08)	(28.26)	(-12.78)	(25.60)	(23.25)	(-17.13)	(28.82)	(25.58)	(42.56)
KSB5	731.77	783.69	412.10	502.97	621.43	869.69	947.58	647.39	756.11
	(-22.62)	(26.88)	(11.79)	(25.31)	(23.13)	(-16.83)	(28.62)	(24.63)	(-42.39)

KSB1-KSB5: the first, second, third, fourth, fifth key spectral band, respectively

2.4 回归模型检测

为检测上述模型的回归效果, 在提取的高光谱样品数据的基础上, 另在每种样品的色料饱满区域随机提取 ROI 点位, 每台打印机选择 20 条数据, 共 180 条高光谱数据。使用前文回归方程对数据进行预测并分析回归效果, 模型预测效果见表 8。

表 8 回归模型预测结果

Table 8 Prediction results of the regression model

Printer model	PLSR						BR					
	R^2	MAE	MSE	RMSE	F	p	R^2	MAE	MSE	RMSE	F	p
Canon E478	0.98	15.14	444.61	21.09	226.11	0.00	0.98	15.15	445.08	21.10	166.66	0.00
Canon IX6780	0.98	18.70	663.55	25.76	59.60	0.00	0.98	18.72	663.65	25.76	33.47	0.00
Canon E500	0.93	48.27	3 315.42	57.58	14.67	0.02	0.93	48.28	3 316.05	57.59	15.69	0.03
HP 1212	0.99	10.62	145.22	12.05	235.44	0.00	0.99	10.62	145.22	12.05	139.20	0.00
HP 2132	0.97	27.52	1 050.57	32.41	24.80	0.00	0.97	27.52	1 050.57	32.41	16.90	0.01
HP 8210	0.99	8.61	219.26	14.81	163.49	0.00	0.99	8.60	219.10	14.80	100.63	0.00
Epson L4266	0.99	11.32	153.28	12.38	148.04	0.00	0.99	11.33	153.67	12.40	77.65	0.00
Epson L8188	0.99	12.70	202.65	14.24	133.16	0.00	0.99	12.69	202.54	14.23	83.42	0.00
Epson L7188	0.91	34.03	1 846.56	42.97	9.58	0.02	0.92	34.03	1 846.91	42.98	8.23	0.03

根据 R^2 可知, 真实值与预测值在大部分区域的拟合程度较好, 表明模型对数据主体模式的拟合能力较强。根据显著性检验结果可知, 经 PLSR 特征提取后, PLSR 和 BR 回归的 p 值均小于 0.05, 达极显著水平, F 检验结果拒绝了零假设(H_0), 表明模型对因变量的解释能力具有极显著的统计学意义。由

于MAE等指标受特殊值影响较大,推测MAE等指标出现的偏差来自少量极端偏离点,反映了模型对特殊情形的敏感性。因变量喷墨打印文件形成时间的取值范围为39~664天,全距为626天,而9台喷墨打印机回归模型预测结果中RMSE的最大取值(57.59)仅占全距的9.2%,且在司法鉴定场景中, ± 60 天的误差范围足以区分相隔半年以上的文件批次,满足多数鉴定需求,在合理误差范围内。回归模型检测结果表明,高光谱成像技术可用于喷墨打印机打印文件形成时间的鉴别。

3 结 论

本文通过使用PCA和PLSR方法对不同品牌、型号的喷墨打印样品的高光谱数据进行降维特征提取处理,并在此基础上,采用LASSO回归、PLSR、RR和BR四种模型对每台打印机样品的高光谱数据进行回归。结果显示,PLSR和BR的回归效果优于LASSO回归和RR。同时通过PLSR权重系数解释了主成分与波段的关联,以明确关键光谱波段的作用。回归效果检测的 R^2 均大于0.9,且 p 值均小于0.05,表明模型对因变量的解释具有极显著的统计学意义,检测中出现的误差在合理范围内。结果显示,高光谱成像技术可以应用于喷墨打印机打印文件形成时间鉴别。本研究的样品时间跨度为两年,且操作简单、无需样品处理、耗时短、结果准确,为法庭科学喷墨打印墨水时间判定提供了一种无损、准确且快捷的鉴别方法。后续将扩大样品量并结合温度、湿度以及光照等环境变化,纸张种类等影响因素,进一步开展相关研究以完善鉴别方法。

参考文献:

- [1] Wang Q Y, Zhao P C, Cui L, Song H, Li Y D, Zhang Z Y. *J. Instrum. Anal.* (王千羽, 赵鹏程, 崔岚, 宋辉, 李扬动, 张祉悦. 分析测试学报), **2024**, 43(6): 858–865, 874.
- [2] Wang C L, Wang Y M. *Chin. J. Inorg. Anal. Chem.* (王彩玲, 王一鸣. 中国无机分析化学), **2023**, 13(9): 986–992.
- [3] Chen Y Y. *China Food Saf. Mag.* (陈燕雨. 食品安全导刊), **2025**, (3): 144–147.
- [4] Yuan J Q, Qiu X C, Gao R, Su Z B. *J. Yunnan Agric. Univ.: Nat. Sci.* (袁建清, 仇逊超, 高睿, 苏中滨. 云南农业大学学报:自然科学), **2025**, 40(1): 167–175.
- [5] Zhang R, Jin W, Mu Y, Yu B W, Bai Y W, Shao Y B, Ping J L, Song P T, He X Y, Liu F, Fu L L. *Spectrosc. Spectral Anal.* (张然, 金伟, 牟颖, 于丙文, 柏怡文, 邵益波, 平金良, 宋鹏涛, 何湘漪, 刘飞, 付琳琳. 光谱学与光谱分析), **2025**, 45(2): 551–557.
- [6] Wang X, Cheng Y, Zhang R Y, Fan Y, Huang J Z, Zhang Y, Yan H B. *Laser Optoelectron. Prog.* (王鑫, 程远, 张若愚, 樊尧, 黄继忠, 张悦, 闫宏彬. 激光与光电子学进展), **2025**, 62(10): 421–430.
- [7] Li S, Cui L, Fu P. *Chin. J. Inorg. Anal. Chem.* (李硕, 崔岚, 付沛. 中国无机分析化学), **2024**, 14(6): 826–835.
- [8] Hao Y L, Zhang B. *Electron. Qual.* (郝艳丽, 张彪. 电子质量), **2022**, (7): 143–147.
- [9] Wei D, Gan D A, Luo F. *Guangdong Police Sci. Technol.* (魏东, 甘德安, 罗飞. 广东公安科技), **2020**, 28(2): 16–19, 38.
- [10] Wang S Y, He H Y, Lv R L, He W W, Li C Y, Cai N B. *J. Forensic Sci.*, **2022**, 67(2): 550–561.
- [11] Wang S Y, Yang Y Z, He W W, Li R K. *J. Instrum. Anal.* (王书越, 杨玉柱, 何伟文, 李润康. 分析测试学报), **2021**, 40(10): 1489–1496.
- [12] Liu M, Shen S, Wang N. *Chin. J. Lumin.* (刘猛, 申思, 王楠. 发光学报), **2017**, 38(5): 663–669.
- [13] Sun P A, Wang B Z. *J. Hunan Univ. Technol.* (孙平安, 王备战. 湖南工业大学学报), **2019**, 33(1): 80–85.
- [14] Gao L, Zhou Y Z, Wang K, Zeng Z Q, Lu Y T, Huang Y J. *Geotecton. Metallog.* (高乐, 周永章, 王琨, 曾志强, 卢宇彤, 黄勇杰. 大地构造与成矿学), **2020**, 44(2): 258–266.
- [15] Tibshirani R. *J. R. Stat. Soc. B*, **1996**, 58(1): 267–288.
- [16] Geladi P, Kowalski B R. *Anal. Chim. Acta*, **1986**, 185: 1–17.
- [17] Hoerl A E, Kennard R W. *Technometrics*, **1970**, 12(1): 55–67.
- [18] Lindley D V, Smith A F M. *J. R. Stat. Soc. B*, **1972**, 34(1): 1–41.

(责任编辑: 盛文彦)