# 联合特征子空间分布对齐的标定迁移方法

赵煜辉，刘晓东，张 磊，刘永宏

东北大学秦皇岛分校，河北 秦皇岛 066000

**摘 要** 近红外光谱分析技术近年来在各种领域的定性、定量分析等方面得到广泛的应用。多元标定技术则是光谱分析领域中最先进的技术，而环境条件、测量仪器或测量物质自身的变化，都可能导致多元标定模型不再适用于新样本的预测。重新标定和重新建模必然会浪费大量时间和资源。一种解决方案是标定迁移，将源域已有的标定模型扩展到目标域中，避免重复建模的代价。在化学计量学的相关文献中，绝大多数迁移方法都需要在两台仪器相同条件下都测量一组迁移标准样品，但在近红外光谱测量技术中，由于标准样品具有挥发等特性，使得构建仪器标定迁移方法的标准样品难以获得和保存。针对这些问题，提出了一种联合特征子空间分布对齐(JSDA)的标定迁移方法，此方法可以在从仪器没有标准样本的情况下建立标定迁移模型。JSDA首先建立源域和目标域数据特征的联合主成分分析(PCA)子空间；然后通过对齐映射在联合特征子空间中的源域特征分布和目标域特征分布来校正标定模型；最后，应用最小二乘模型构建校正后源域上的标定模型，该模型可直接用于目标域的标定。实验结果表明与已有成熟的标定迁移方法相比，JSDA在公开的真实数据集上的预测性能比较有优势，验证了该模型在实际应用中的有效性和优越性。

**关键词** 近红外光谱；标定迁移；PCA子空间；联合子空间分布对齐

**中图分类号**：O657.33 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)11-3411-07

## 引 言

近红外光(NIR)是一种波长在780～2 526 nm之间的电磁波，近红外光谱区与有机分子中含氢基团(O—H，N—H，C—H)振动的合频和各级倍频的吸收区一致，通过扫描样品的近红外光谱，可以得到样品中有机分子含氢基团的特征信息[1-2]。近红外光谱的多元标定方法是利用含有氢基团化学键伸缩振动倍频和合频，在近红外区域的吸收光谱，通过选择适当的化学计量学领域的多元标定方法，找到标定样本的近红外吸收光谱与其相应的成分浓度或性质数据之间的关联，建立两者之间的标定关系模型[3]。主成分回归(principal component regression，PCR)[4]和偏最小二乘(partial least squares，PLS)[5]等标定方法已经被证实是有效的，建立可靠的多元标定模型通常耗时且成本高昂，然而在实际工业生产中，通过对原有近红外光谱数据进行分析建立的模型往往对新的数据集并不适用，从而导致原有模型失效。解决此类问题通常有两种方法：一是重新对新的数据集进行重新标定和重建模型；二是建立标定迁移模型，将已有可靠的源域多元标定模型迁移到目标域中。重新标定和重建模型需要耗费大量的时间和资源[6]，而标定迁移不仅可以有效的避免这一缺点，而且还可以使得目标领域取得可靠的学习效果。显然，选择第二种方法是解决此类问题的最佳策略[7]。

一般来说，标定迁移方法可以分为两类：有标样的标定迁移和无标样的标定迁移。目前比较有代表性的有标样的标定迁移方法有直接标准化(direct standardization，DS)[8]、分段直接标准化(piecewise direct standardization，PDS)[9]、基于典型相关分析的标定迁移(canonical correlation analysis based calibration transfer，CCACT)[10-11]以及斜率和偏差校正算法(slope bias correction，SBC)[12]等，无标样的标定迁移方法有多元散射校正(multiplicative scatter correction，MSC)[13]、迁移成分回归(transfer component regression，TCR)[14]等，其中DS和PDS的前提是假设光谱响应的变异都是测量环境引起的；但是实际上，我们所收集和整理的化学样品也存在着一定的不确定性；SBC为一种单变量方法，因此在测量仪器和测量条件变化引起系统化的光谱差异的情况下，才能取得较好的效果。现实生活中，光谱差异往往比较复杂，此时它的预测能力是不确定的；MSC预处理方法并不能显著提高模型的预测能力；TCR虽然具有较好的泛化能力，但与其他方法相比预测精度较低。

**收稿日期**：2020-10-20，**修订日期**：2021-02-15

**基金项目**：国家自然科学基金项目(61601104)资助

**作者简介**：赵煜辉，1971年生，东北大学秦皇岛分校教授 e-mail：1000272@neuq.edu.cn

大多数能够显著地提高预测性能的迁移方法都属于有标样的标定迁移方法，即需要标准样本来构建标定迁移模型，且标准样本中主仪器与从仪器的样本必须一一对应紧密匹配，具备良好的代表性和适应性，能够很好地解释两种仪器之间的差异。由于这些要求的限制，有标样的模型通常泛化能力较差。而已被提出的少量无标准标定迁移方法虽然不需要标准样本，但其预测性能与有标样的标定迁移方法相比相差较大。因此，结合两者优点，开发一种性能可与有标样的迁移方法媲美的无标准样本的迁移学习方法，将具有很大的意义。因此结合近红外光谱维度高且存在多重共线性的特点，以主成分回归(PCR)作为标定模型，应用迁移学习的思想，提出了一种无标准样本的基于联合特征子空间分布对齐(joint feature subspace distribution alignment，JSDA)的标定迁移方法，在不需要标准样本的情况下，取得相同甚至优于已有经典有标样的标定迁移方法的预测性能。

## 1 理 论

### 1.1 符号定义

在本文中，源域和目标域将各用下标“s”和“t”表示，$X_s=[x_s^1,\cdots,x_s^{n_s}]\in R^{n_s\times m}$表示源域数据集，$X_t=[x_t^1,\cdots,x_t^{n_t}]\in R^{n_t\times m}$表示目标域数据集。其中 $m$ 表示域中数据的维数，$n_s$ 和 $n_t$ 各表示源域和目标域样本个数。$U$ 表示主成分分析(principal component analysis，PCA)子空间的基变换，$\|\cdot\|_2$表示 2 范数，$\|\cdot\|_F$表示矩阵的 Frobenius 范数。

### 1.2 模型建立

下面我们将具体说明如何建立基于近红外光谱特征预测物质成分浓度的无标准样本的标定迁移模型。用均值和协方差来描述光谱数据分布。由于均值在数据预处理(如中心化)后通常为零，不受子空间投影的影响，因此不需要对它们进行处理。协方差反映着多维空间基向量之间的相关关系，源域和目标域的协方差矩阵存在差异，且向子空间投影会对其产生影响，因此我们需要消除投影后两者特征光谱协方差矩阵之间的差异，进而使得两者数据分布对齐[15]。

下面我们从理论上详细阐述 JSDA 模型的建立过程：

第一步：构建联合公共特征子空间

对于光谱数据，一般均为高维小样本 $X\in R^{n\times m}$，$m>n$，即光谱属性维度远大于样本个数。针对此类数据，为了便于构建预测模型，一般要对其进行降维，如 PCA，将高维样本空间 $X$ 投影转换为低维子空间 $\Phi\in R^{n\times d}$，$n>d$。标定迁移问题中，我们可以利用一种特殊的方式直接构建一个公共特征子空间，将源域和目标域数据均投影到该子空间中，如此，两者之间不存在子空间漂移的问题。对于属性维度相同的光谱数据 $X_s\in R^{n_s\times m}$，$X_t\in R^{n_t\times m}$，我们直接利用分别中心化后的光谱矩阵 $\hat{X}_s$ 和 $\hat{X}_t$ 构造一个联合光谱矩阵 $X_{com}=\begin{bmatrix}\hat{X}_s\\ \hat{X}_t\end{bmatrix}\in R^{(n_s+n_t)\times m}$，然后采用 PCA 求解联合光谱矩阵的特征向量矩阵 $U\in R^{d\times(n_s+n_t)}$，联合公共特征子空间由特征向量矩阵 $U$ 组成，如式(1)所示

$$X_{com}^{T}X_{com}=U\Lambda U^{T} \tag{1}$$

利用 PCA 构建子空间的过程中涉及到特征值个数 d 的选取，可以通过碎石图准则、累计方差贡献率法等方法进行确定，即 $U=[U_1,\cdots,U_d,\cdots,U_{n_s+n_t}]$，$U_d\in R^{n\times d}$。直观地说，通过特征向量矩阵 $U_d$，将主从域原始光谱数据投影到低维的子空间中，既保留了源域和目标域共有的关键结构，同时可以去除有噪声的维度，从而更容易找到后续的映射关系。通过特征向量矩阵 $U_d$，可以得到投影至公共特征子空间中的特征光谱矩阵 $\Phi_s=\hat{X}_sU_d$ 和 $\Phi_t=\hat{X}_tU_d$。

对于传统的子空间对齐方法，源域与目标域数据分别构建低维特征子空间时，存在一个问题，由于投影矩阵 $U_s$ 和 $U_t$ 的不同，造成转换后两者特征子空间基存在差异；通过计算线性映射矩阵来对齐子空间，从而最小化它们之间分布差异，这种方法称为子空间对齐。而我们提出的构建源域和目标域的联合特征子空间，使得源域和目标域的特征光谱不仅具有相同的子空间基，并且能够尽可能的保证原始数据在投影到该子空间上的时候不会失真，达到最优状态，因此不需要进一步对齐子空间，又有很好的优越性。

第二步：特征分布对齐

公共特征子空间中，源域和目标域具有相同的子空间基，但这并不能解决两者数据特征分布之间的差异，不能满足预测模型应用的独立同分布条件。如上所述，我们用均值和方差描述一个分布。前面提到，均值在数据中心化处理后不受子空间投影的影响，因此我们只需消除投影后两者特征光谱的协方差差异。为了最小化源域特征和目标域特征的二阶统计量(协方差：$\Sigma_s$ 和 $\Sigma_t\in R^{d\times d}$)之间的距离，我们对源域特征进行线性变换 $A\in R^{d\times d}$，使用 Frobenius 范数作为矩阵距离度量，从而最小化它们之间差异，如式(2)所示

$$\min_A\|\hat{\Sigma}_s-\Sigma_t\|_F^2=\|A^T\Sigma_sA-\Sigma_t\|_F^2 \tag{2}$$

其中，$\hat{\Sigma}_s$ 为线性变化后的源域特征光谱 $\Phi_sA$ 的协方差矩阵。

进一步对式(2)推导可得

$$A^T\Sigma_sA=\Sigma_t \tag{3}$$

对于半正定的协方差矩阵 $\Sigma=\Phi^T\Phi$，式(3)可写做 $(\Sigma_s^{1/2}A)^T(\Sigma_s^{1/2}A)=(\Sigma_t^{1/2})^T(\Sigma_t^{1/2})$，即 $\Sigma_s^{1/2}A=\Sigma_t^{1/2}$。因此，我们可以得到一个最优化结果，如式(4)所示

$$A^*=\Sigma_s^{-1/2}\Sigma_t^{1/2} \tag{4}$$

而实际应用中根据已有样本估计的光谱数据协方差矩阵常是不可逆的，因为样本数据集的特征数总大于样本数，但一般样本可以集中于一个低维子空间中，构建子空间中的特征光谱，此时一般可逆。对于协方差矩阵不可逆的情况，我们将结果修正如式(5)所示

$$\hat{A}^*=(\Sigma_s^{1/2})^+\Sigma_t^{1/2} \tag{5}$$

其中“+”表示 Moore-Penrose 伪逆，$\hat{A}^*$ 为修正后的线性变换矩阵。

为了便于理解，我们给出联合特征子空间下的特征分布对齐示意图如图 1，红色表示源域特征样本，蓝色表示目标域特征样本。其中图 1(a)表示中心化后的两域原始数据投影到联合特征子空间上的分布差异，图 1(b)表示对源域特征进行线性变换后差异。可以看到经过均值和协方差校正后，两域的特征分布基本相同。

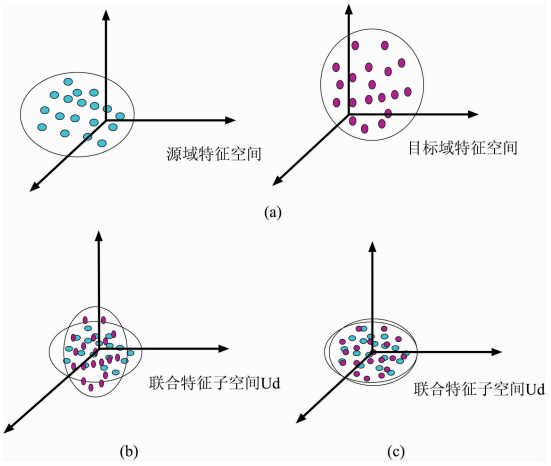


**图 1　特征分布对齐示意图**

(a)：原数据；(b)均值校正；(c)：协方差校正

**Fig. 1　Feature distribution alignment diagram**

(a)：Original；(b)：Mean correction；(c)：Covariance correction

第三步：构建目标函数

本工作所解决的标定迁移问题是一个预测问题，根据上述步骤的结果，我们可以应用最小二乘法构建校正分布差异后的源域回归预测模型的目标函数，其形式化如式(6)所示

$$\min_{\beta} \| \hat{\Sigma}_s \beta + b - y_s \|_2 = \| \Sigma_s \boldsymbol{A} \beta + b - y_s \|_2 \quad (6)$$

其中，$\Phi_s$ 为源域投影至公共特征子空间中的特征光谱，$\boldsymbol{A}$ 为源域与目标域分布对齐的线性变换矩阵。$\beta$ 和 $b$ 为源域最小二乘回归模型参数的向量形式，其中 $\beta$ 为回归系数向量，$b$ 表示由截距常数 $b=\bar{y}_s=\mu(y_s)$ 组成的列向量，表达形式如式(7)

$$\beta = (\boldsymbol{A}^{\mathrm{T}} \Phi_s^{\mathrm{T}} \Phi_s \boldsymbol{A})^{-1} (\Phi_s \boldsymbol{A})^{\mathrm{T}} (y_s - b)$$
$$b = \bar{y}_s = \mu(y_s) \quad (7)$$

经过上述步骤，源域和目标域具有相同的子空间基，且实现数据分布对齐，因而源域上构建的回归模型在两域之前满足数据独立同分布条件。显然，上述目标函数求解得到的源域回归模型，可以直接用于目标域上的回归预测。

第四步：得到目标域标定模型

上一步中，源域上得到的最小二乘回归模型参数 $\beta$ 和 $b$ 可以直接用于目标域上的回归预测，如式(8)所示

$$\hat{y}_t = \Phi_t \beta + b \quad (8)$$

### 1.3　算法流程

算法：JSDA 算法

输入：主仪器光谱矩阵 $X_s$；主仪器样本物质浓度矩阵 $y_s$；从仪器光谱矩阵 $X_t$。

输出：标定迁移模型 $f(\beta, b, A)$。

开始：

(1)数据中心化处理

$\hat{X}_s \leftarrow X_s - \mu(X_s)$，$\hat{X}_t \leftarrow X_t - \mu(X_t)$，

$\hat{y}_s \leftarrow y_s - \mu(y_s)$

(2)构造联合光谱矩阵 $X_{com} \leftarrow \begin{bmatrix} \hat{X}_s \\ \hat{X}_t \end{bmatrix}$；

(3)利用式(1)找到公共特征子空间 $U_d$；

(4)求得到两域的特征光谱 $\Phi_s \leftarrow \hat{X}_s U_d$，$\Phi_t \leftarrow \hat{X}_t U_d$；

(5)利用式(2)求解线性映射矩阵 $A \leftarrow \Sigma_s^{-1/2} \Sigma_t^{1/2}$；

(6)利用式(6)建立源域标定模型，得到模型参数 $\beta$ 和 $b$，返回标定迁移模型。

## 2　实验部分

为了验证算法的准确性和实用性，使用玉米数据集和小麦数据集作为实验对象，对数据集进行了数据分析，来检验 JSDA 方法的性能。

### 2.1　数据集

第一个数据集是玉米数据集，包含三个 NIR 光谱仪(M5，MP5 和 MP6)测得的 80 个样品的光谱数据。这三台不同的红外光谱仪因其工作原理不同，所以得到的近红外光谱略有差异，但对绝大多数谷物而言，仪器的工作原理不同所产生的误差并不会影响试验结果，所以我们采用这三台仪器测量的 80 个玉米的近红外光谱做分析。玉米数据集中每个样品含有四种成分：水分，油，蛋白质和淀粉。波长范围为 1 100～2 498 nm(700 通道)，间隔为 2 nm。该数据集可以从 http://www.eigenvector.com/Data/Corn/下载。仪器 M5 和仪器 MP5 之间的光谱差异如图 2(a)所示；仪器 M5 和仪器 MP6 之间的光谱差异如图 2(b)所示；仪器 MP5 和仪器 MP6 之间的光谱差异如图 2(c)所示。其中横轴表示波长，纵轴表示吸光度差异(即两种仪器的吸光度差值)，每条曲线代表一个光谱样本。

第二个数据集是小麦数据集，它被用作 2016 年国际漫反射会议(IDRC)上发布的“Shootout”数据集，选择蛋白质含量作为属性。小麦数据集的相关信息访问网址 http://www.idrc-chambersburg.org/content.aspx?page_id=22&club_id=409746&module_id=191116。它分析了来自三个不同 NIR 仪器制造商(A1，A2 和 A3)的 248 份小麦数据集的样本。仪器 A1 和仪器 A2 之间的光谱差异如图 2(d)所示；仪器 A1 和仪器 A3 之间的光谱差异如图 2(e)所示；仪器 A2 和仪器 A3 之间的光谱差异如图 2(f)所示。

### 2.2　数据处理

通过 Kennard-Stone(KS)算法将玉米数据集的 80 个样本分成两组：80%用做标定集的样本，20%用做测试集的样本；将小麦数据集的 248 个样本分成两组：80%用作标定集的样本，20%用作测试集的样本。对于有迁移标准的迁移方法，使用 Kennard-Stone(KS)算法在标定样本上选择若干个标准样品。

### 2.3　模型评估指标

在该实验中，均方根误差(root mean squard error，RMSE)被用作参数选择和模型评估的指标。RMSE 是预测值与真实值偏差的平方与观测次数 n 比值的平方根，可表示数据偏离真实值的程度，其计算方法如式(9)所示

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \quad (9)$$

式(9)中，$y_i$ 为第 $i$ 个样本的真实值，$\hat{y}_i$ 为第 $i$ 个样本的预测值，$n$ 为观测样本数。RMSE越小，表示模型的预测精度越

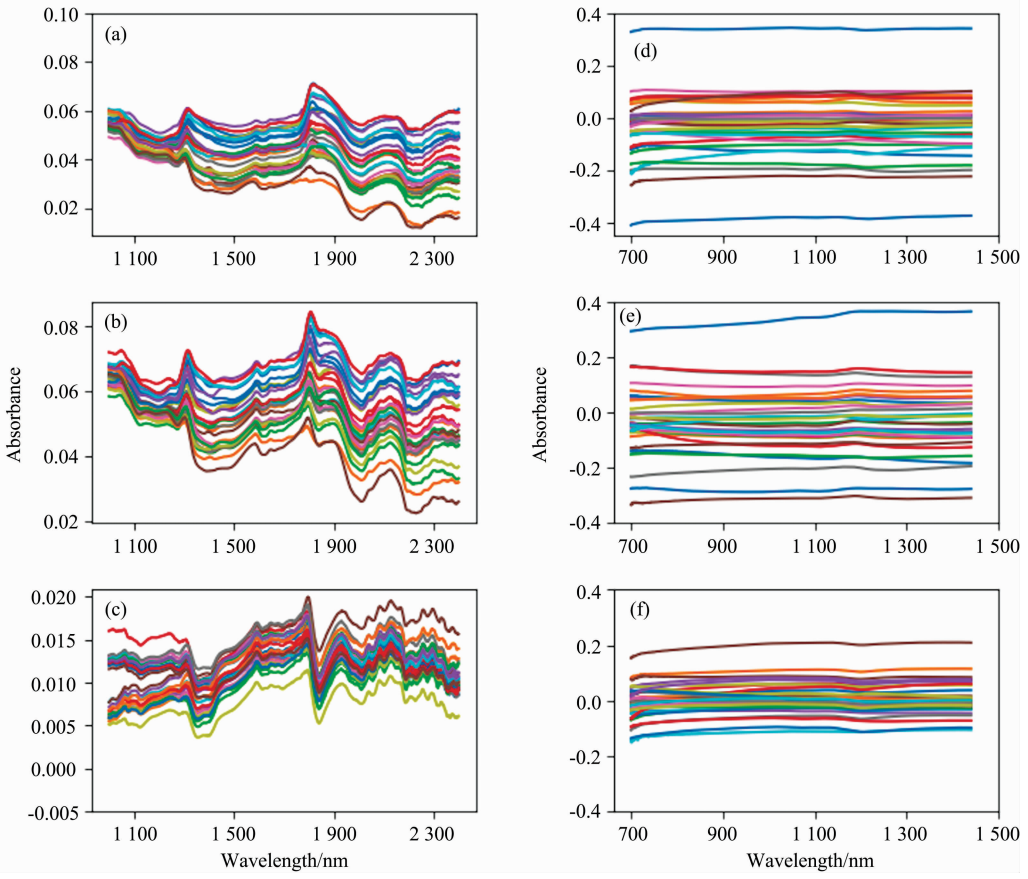


**图 2 不用仪器之间的光谱差异**

**Fig. 2 Spectral differences between different instruments**

高。RMSEP 表示最终测试集上的预测均方根误差。

## 3 结果与讨论

玉米数据集包含各仪器样本各 80 个，以 M5 为主仪器，MP5 和 MP6 分别为从仪器以及 MP5 为主仪器，MP6 为从仪器的实验预测误差 RMSEP 如表 1 所示。小麦数据集包含各仪器样本各 248 个，以 A1 为主仪器、A2 和 A3 分别为从仪器以及 A3 为主仪器、A2 为从仪器的实验预测误差 RMSEP如表 2 所示。其中表中有标样的迁移学习模型(SBC，

**表 1 SBC，PDS，CCACT，MSC，TCR 和 JSDA 六种迁移方法在玉米数据集下的 RMSEP**

**Table 1 RMSEP of corn datasets with SBC，PDS，CCACT，MSC，TCR and JSDA**

| $N_{std}$ | SBC | PDS | CCACT | MSC | TCR | JSDA |
|---|---|---|---|---|---|---|
| 玉米数据集的 RMSEP（M5 作为主仪器，MP5 作为从仪器） | | | | | | |
| $N=15$ | 0.317 49 | 0.240 54($15^a$) | 0.252 59 | | | |
| $N=25$ | 0.251 58 | 0.218 47($15^a$) | 0.250 05 | 0.899 59 | 0.355 51($6^b$) | **0.118 64** |
| $N=35$ | 0.267 10 | 0.227 79($15^a$) | 0.246 27 | | | |
| 玉米数据集的 RMSEP（M5 作为主仪器，MP6 作为从仪器） | | | | | | |
| $N=15$ | 0.436 55 | 0.433 16($7^a$) | 0.546 91 | | | |
| $N=25$ | 0.388 60 | 0.467 49($9^a$) | 0.420 87 | 1.921 09 | 0.474 99($4^b$) | **0.146 09** |
| $N=35$ | 0.355 98 | 0.403 44($5^a$) | 0.406 74 | | | |
| 玉米数据集的 RMSEP（MP5 作为主仪器，MP6 作为从仪器） | | | | | | |
| $N=15$ | 0.225 77 | 0.240 54($15^a$) | 0.252 59 | | | |
| $N=25$ | 0.227 63 | 0.218 47($15^a$) | 0.250 05 | 0.899 59 | 0.355 51($10^b$) | **0.172 82** |
| $N=35$ | 0.238 45 | 0.227 79($15^a$) | 0.246 27 | | | |

注：$N_{std}$：需要标准样本的迁移方法中，标准样本的数目；a：PDS 中最优的窗口大小；b：TCR 中对应的最优子空间的维度(下同)

Note：$N_{std}$ is the number of standard samples required by the migration method；a represents the optimal window size of PDS and b represents the optimal subspace dimension of TCR (the same below)

**表 2　SBC，PDS，CCACT，MSC，TCR 和 JSDA 六种迁移方法在小麦数据集下的 RMSEP**

**Table 2　RMSEP of wheat datasets with SBC，PDS，CCACT，MSC，TCR and JSDA**

| $N_{std}$ | SBC | PDS | CCACT | MSC | TCR | JSDA |
|---|---|---|---|---|---|---|
| 小麦数据集的 RMSEP（A1 作为主仪器，A2 作为从仪器） | | | | | | |
| $N=15$ | 2.170 63 | 3.947 74($15^a$) | 2.322 63 | | | |
| $N=25$ | 1.284 26 | 3.793 84($15^a$) | 2.046 23 | 1.605 83 | 2.035 40($10^b$) | **0.283 44** |
| $N=35$ | 0.905 77 | 3.504 91($15^a$) | 1.988 81 | | | |
| 小麦数据集的 RMSEP（A1 作为主仪器，A3 作为从仪器） | | | | | | |
| $N=15$ | 0.477 07 | 1.524 29($7^a$) | 2.030 12 | | | |
| $N=25$ | 0.477 39 | 1.408 40($15^a$) | 2.071 13 | 1.215 13 | 1.964 12($19^b$) | **0.252 32** |
| $N=35$ | 0.478 35 | 1.346 55($15^a$) | 1.902 66 | | | |
| 小麦数据集的 RMSEP（A3 作为主仪器，A2 作为从仪器） | | | | | | |
| $N=15$ | 8.660 48 | 3.593 11($15^a$) | 2.249 79 | | | |
| $N=25$ | 7.082 65 | 2.229 61($3^a$) | 2.250 17 | 1.255 70 | 2.120 74($10^b$) | **0.277 81** |
| $N=35$ | 5.996 79 | 2.069 79($3^a$) | 2.062 74 | | | |

PDS，CCACT）需要迁移标准样本的个数$N_{std}$不能过少也不能过多，因此，在[15，35]的范围内选取标准样本，以 10 为增量，获取不同数量标准样本对模型预测误差的影响。观察表中的预测误差结果，总体来说，本文提出的 JSDA 方法在六组对比实验中具有最小的预测误差，最好的预测精度。在其他五种有标样和无标样标定迁移方法中，可以发现三种有标样标定迁移方法（SBC，PDS，CCACT）的预测误差都小于无标样标定迁移方法（MSC，TCR）。有标样方法虽然需要获取标准样本，增加了模型的应用代价，但相应的预测精度也得到了提升，而无标样方法不需要标准样本，提高了模型的泛化能力和适用性，但相应的预测精度也受到了影响。本文提出的 JSDA 方法，很好的解决了无标样标定迁移方法预测精度较低的问题，在具备与标定迁移方法相同甚至更加优异的预测精度的同时，还具备良好的适用性，应用代价较低。

为了直观地观测六种标定迁移方法的性能，实验中，以从仪器测试集的物质浓度数据测量值为横坐标，以标定迁移方法的预测值为纵坐标，描绘玉米数据集三组实验和小麦数据集三组实验的观测浓度与预测浓度关系图，如图 3—图 8 所示。图中的无差异直线表示，若观测浓度与预测浓度之间误差为零，则对应的样本点会落在此直线上。对比观察图 3—图 8 中的预测结果可知，六种模型中 MSC 模型在两组实

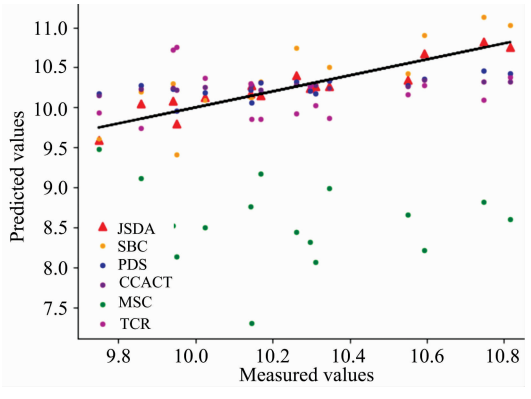


**图 3　JSDA，SBC，PDS，CCACT，MSC，TCR 六种方法在仪器 M5 和仪器 MP5 之间预测结果的散点图**

**Fig. 3　Scatter plots of prediction comparison between instruments M5 and MP5 using JSDA，SBC，PDS，CCACT，MSC，TCR**

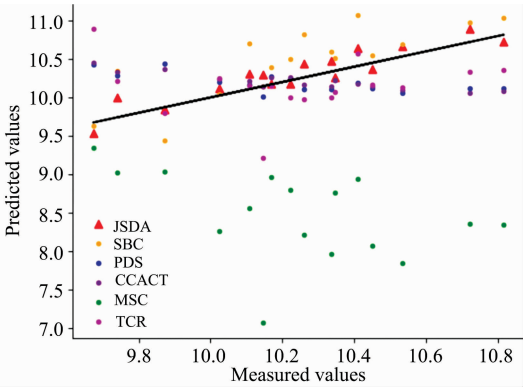


**图 4　JSDA，SBC，PDS，CCACT，MSC，TCR 六种方法在仪器 M5 和仪器 MP6 之间预测结果的散点图**

**Fig. 4　Scatter plots of prediction comparsion between instruments M5 and MP6 using JSDA，SBC，PDS，CCACT，MSC，TCR**

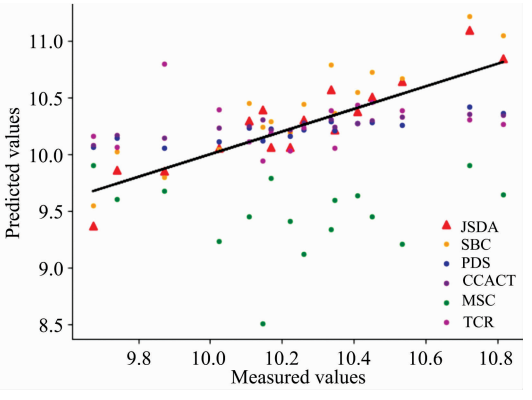


**图 5　JSDA，SBC，PDS，CCACT，MSC，TCR 六种方法在仪器 MP5 和仪器 MP6 之间预测结果的散点图**

**Fig. 5　Scatter plots of prediction comparison between instruments MP5 and MP6 using JSDA，SBC，PDS，CCACT，MSC，TCR**

验四种物质上的预测结果基本都聚集在无差异直线的某一侧，这与表 1 和表 2 中展示的结果相呼应，表明 MSC 模型的性能较差，无法准确的标定从仪器的物质浓度。而 CCACT，PDS，SBC，TCR 以及本文提出的 JSDA 模型在两组实验上的预测结果基本都聚集在无差异直线的两侧，分布都较为均匀，但相对来说，SBC 模型的预测结果分布较为散乱，表明模型鲁棒性较差。对比所有模型的预测结果，以 JSDA 模型的预测结果最为贴近无差异直线，拟合效果最好，结合表 1 和表 2 中的结果，可以得知，本文提出的 JSDA 方法具备最佳的预测性能，同时具有更好的泛化能力。

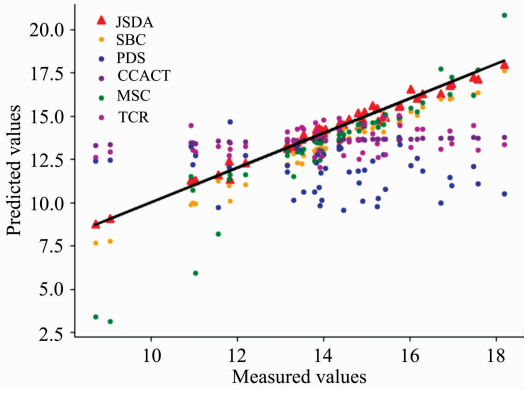


**图 6 JSDA，SBC，PDS，CCACT，MSC，TCR 六种方法在仪器 A1 和仪器 A2 之间预测结果的散点图**

**Fig. 6 Scatter plots of prediction comparison between instruments A1 and A2 using JSDA，SBC，PDS，CCACT，MSC，TCR**

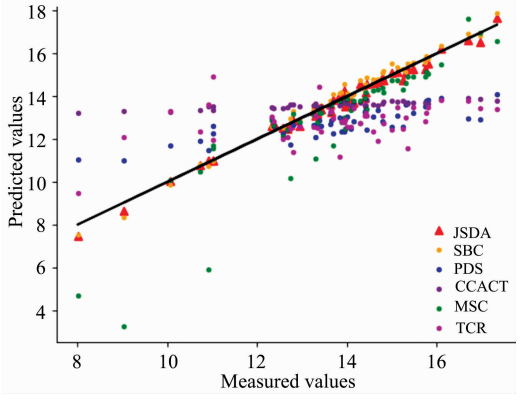


**图 7 JSDA，SBC，PDS，CCACT，MSC，TCR 六种方法在仪器 A1 和仪器 A3 之间预测结果的散点图**

**Fig. 7 Scatter plots of prediction comparison between instruments A1 and A3 using JSDA，SBC，PDS，CCACT，MSC，TCR**

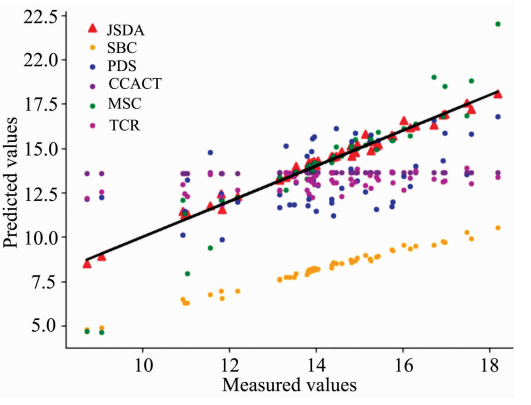


**图 8 JSDA，SBC，PDS，CCACT，MSC，TCR 六种方法在仪器 A3 和仪器 A2 之间预测结果的散点图**

**Fig. 8 Scatter plots of prediction comparison between instruments A3 and A2 using JSDA，SBC，PDS，CCACT，MSC，TCR**

## 4 结 论

通过在玉米和小麦的近红外光谱数据集上，在 JSDA 与 SBC，PDS，CCACT，MSC，TCR 五种对比标定迁移方法之间，进行的两组对比实验，验证了本文方法的性能。总体来说，实验结果中，本文提出的 JSDA 方法的预测误差都是最低的，表明在实验的两个数据集上，JSDA 方法的性能最优异，其次是 PDS 和 CCACT，SBC 虽然预测的 RMSE 较小，但预测结果不稳定，然后是 TCR，而 MSC 方法的预测性能最差。实验结果充分验证了本文所提 JSDA 方法在实际应用中的优越性，JSDA 方法在解决传统标定迁移方法大多需要标准样本这一缺点的同时，具备与有标样的标定迁移方法相同甚至更优异的性能。

## References

[1] CHU Xiao-li，LU Wan-zhen(褚小立，陆婉珍). Spectroscopy and Spectral Analysis(光谱学与光谱分析)，2014，34(10)：2595.
[2] Wiesner K，Fuchs K，Gigler A M，et al. Procedia Engineering，2014，87：867.
[3] López Ainara，Arazuri S，García I，et al. Journal of Agricultural and Food Chemistry，2013，61(23)：5413.
[4] Liu W X，Xu J K，Jiang H Y，et al. Advances in Energy Science and Technology Part 4，2013. 2381.
[5] Wu N，Xu C S，Yang R J，et al. IOP Conference Series：Earth and Environmental Science，2018，113(1)：012004.
[6] Abdelkader M F，Cooper J B，Larkin C M. Chemometrics and Intelligent Laboratory Systems，2012，110(1)：64.
[7] Malli B，Birlutiu A，Natschläger T. Chemometrics and Intelligent Laboratory Systems，2017，161：49.
[8] Fonollosa J，Fernández L，Gutiérrez-Gálvez A，et al. Sensors and Actuators B：Chemical，2016，236：1044.
[9] Alves J C L，Poppi R J. Spectrochimica Acta Part A：Molecular and Biomolecular Spectroscopy，2013，103：311.

[10] Fan W, Liang Y, Yuan D, et al. Analytica Chimica Acta, 2008, 623(1): 22.
[11] Zheng K, Zhang X, Iqbal J, et al. J. Chemom., 2014, 28: 773.
[12] Brown S D. Transfer of Multivariate Calibration Models. Chemometrics, 2013. 10.1016/B978-0-12-409547-2.00644-2.
[13] Chen Huazhou, Song Qiqing, Tang Guoqiang, et al. ISRN Spectroscopy, 2013. 10.1155/2013/642190.
[14] Pan S J, Tsang I W, Kwok J T, et al. IEEE Transactions on Neural Networks, 2011, 22(2): 199.
[15] Fernando B, Habrard A, Sebban M, et al. IEEE International Conference on Computer Vision, 2013, 1: 2960.

# Research on Calibration Transfer Method Based on Joint Feature Subspace Distribution Alignment

ZHAO Yu-hui, LIU Xiao-dong, ZHANG Lei, LIU Yong-hong
Northeastern University Qinhuangdao Campus, Qinhuangdao 066000, China

**Abstract** Near-infrared spectroscopy analysis technology has the advantages of low cost, high efficiency, and pollution-free. In recent years, it has been widely used in qualitative and quantitative analysis in various fields. Multivariate calibration technology is the most advanced technology in the field of spectroscopy. Changes in conditions, instruments, or substances may cause the multivariate calibration model to no longer be suitable for the prediction purposes of newly measured samples. Re-calibration and re-modeling will inevitably waste a lot of time and resources; another option is calibration transfer, which extends the existing calibration model in the source domain to the target domain to avoid the cost of repeated modeling. In the related chemometrics literature, most transfer methods need to measure a set of transfer standard samples under the same conditions of two instruments. However, in the near-infrared spectroscopy measurement technology, due to the characteristics of volatilization of the standard samples, It is not easy to obtain and save the standard samples for constructing the transfer method for instrument calibration. This paper proposes a joint feature subspace distribution alignment (JSDA) calibration transfer method in response to these problems. This method can establish a calibration transfer model without a standard sample from the instrument. JSDA first establishes the joint PCA subspace (Principal component analysis) of the data features of the source and target domains; then corrects the calibration model by aligning the source domain feature distribution and target domain feature distribution mapped in the joint feature subspace; Finally, the least squares model is used to build a calibration model on the corrected source domain, which can be directly used for the calibration of the target domain. The experimental results show that compared with the existing mature calibration transfer methods, JSDA has more advantages in predicting performance on public real data sets, which verifies the effectiveness and superiority of the model in practical applications.

**Keywords** Near infrared spectroscopy; Calibration transfer; PCA subspace; Joint subspace distribution alignment

(Received Oct. 20, 2020; accepted Feb. 15, 2021)