Raman Spectral Analysis of Low-content Benzene Concentration in Gasoline with Partial Least Squares Based on Interference Peak Subtraction

Wei LIU and Lian-kui DAI†

State Key Lab of Industrial Control Technology, Zhejiang University, Hangzhou 310027, China

Raman spectroscopy is adopted to detect the low-content benzene concentrations in gasoline products. Due to the peak overlap of benzene and other species in the gasoline spectrum, the associated statistical regression methods cannot make stable predictions unless there are enough training samples. To extend their extrapolation to small-size training sets, we propose the method of partial least squares based on a spectral pretreatment of interference peak subtraction (IPS-PLS). During the analysis, after spectral interpolation and baseline removal, we extract the benzene peak by interference peak subtraction (IPS), and then partial least squares (PLS) is applied to make a prediction. The experimental results demonstrate that, IPS can extract benzene information effectively, and help to decrease principal components needed by PLS, thus IPS-PLS is superior to direct PLS with small-size training sets, and depends less on the training sample distribution. Meanwhile, IPS-PLS can reach the standard of ASTM 3606-10 with the least of 9 training samples, while keeping its max predictive error less than 0.1254% (v/v), which shows promising prospects in gasoline quality test.

Keywords Gasoline, Raman spectroscopy, spectral interference peak subtraction, partial least squares

(Received February 26, 2016; Accepted April 26, 2016; Published August 10, 2016)

Introduction

With the development of the automotive industry, vehicle exhaust emission has become the primary source of urban pollution. Although the amount of benzene in gasoline is small, long-term breathing in vehicle exhaust with excessive benzene will gradually lead to a decreased immune system, childhood leukemia and even cancer. During the analysis, after spectral interpolation and baseline removal, we extract the benzene peak by interference peak subtraction (IPS). Meanwhile, IPS-PLS is superior to direct PLS with small-size training sets, and depends less on the training sample distribution. Consequently, more training samples are needed for regression methods to make accurate predictions. However, in the actual production process, it often fails to collect sufficient training samples beforehand due to volatile producing conditions as well as detecting noises; as a result, the lack of training samples may lead to unsatisfactory performance of those conventional statistical methods.

Motivated by guaranteeing the accuracy of statistical methods with a small-size training set, we propose an analysis strategy involving the PLS method based on a spectral pretreatment of interference peak subtraction (IPS-PLS). The principle of the interference peak subtraction (IPS) method is to extract the actual benzene peak by fitting the interference peak with the line shape of the corresponding pure spectrum. To determine the benzene concentration, we first pretreat spectra with conventional preprocessing of linear interpolation, spectral baseline removal and normalization. We then extract the useful spectral signal with IPS method, and finally PLS is used to build calibration model and make predictions on gasoline samples. In the experimental part, to verify the effect of the new method, we first compared the linear correlation coefficients of the benzene concentration against its peak intensity at 997 cm⁻¹ before and after the IPS pretreatment; then we validated the predictive

† To whom correspondence should be addressed.
E-mail: lkdai@iipc.zju.edu.cn
performance of IPS-PLS and direct PLS method with different numbers of principal components as well as different numbers of training samples. The result shows that, the spectral signal extracted by the IPS method succeeds in reflecting the actual benzene information, so that under the premise of keeping prediction accuracy, the IPS pretreatment can help to decrease the principal components needed by PLS, and to extend its usage to a small-size training set. Finally, the IPS-PLS method was evaluated with small-size training sets. The result demonstrates that it can reach the standard of ASTM 3606-10 with the least of 9 training samples, while keeping its max predictive error to be less than 0.1254% (v/v), which shows a promising prospect for gasoline quality tests.

**Experimental**

**Samples and instruments**

We collected 65 gasoline samples from an oil refinery in NingBo, China, with the benzene concentration ranging from 0.4 to 1% (v/v). Their Raman spectra were recorded with an accumulation time of 10 s using a fiber-probe-based portable Raman system. The excitation beam was generated by a temperature-stabilized diode laser operating at 785 nm with power of 450 mW (B&W Tek, USA). The Raman signal was detected by a spectrometer (QE65000, Ocean Optics, USA), with a resolution of 5 cm$^{-1}$ at a 1000 cm$^{-1}$ Raman shift. The benzene concentrations were measured by ASTM 3606-10, represented with volume fraction (v/v). Original spectra of the samples are shown in Fig. 1.

**Data preprocessing**

The main purpose of data preprocessing is to reduce the noise influence to the Raman spectra, including fluorescence and changes of the detecting conditions. The pretreated spectrum ranges from 400 to 1800 cm$^{-1}$. First of all, we smoothed them with linear interpolation, and then subtracted baseline drifting by iterative polynomial fitting to ensure that the spectra in the region without Raman emission can reach an intensity of nearly 0. Taking the spectrum of sample No. 1 as an example, selecting the baseline polynomial order as 2, the baseline subtraction result is shown in Fig. 2. Since the spectra vary according to the changes of the detecting conditions such as the laser power and temperature, we normalize the spectral intensity according to the saturated hydrocarbon peak at 1450 cm$^{-1}$, which is relatively stable under different conditions. The preprocessed spectra of all the samples are shown in Fig. 3.

**Interference peak subtraction**

Since the characteristic peak for benzene at 997 cm$^{-1}$ is more or less overlapped with others, its actual signal can be easily masked by other useless information. In order to obtain the actual benzene peak intensity, we need to subtract the influence of the other components in gasoline. The principle of interference subtraction is to fit the mixture spectrum with spectral line profiles of pure components so as to resolve the overlapping peaks in mixture. Since gasoline is such a complicated mixture, it is impossible to obtain all spectra of its constituents, so it is more convenient to consider the local spectrum from 970 to 1020 cm$^{-1}$, where there are only 2 peaks. The first one is the unique characteristic peak for benzene,
which locates at 997 cm\(^{-1}\), and the second one at around 1007 cm\(^{-1}\) belongs to the \(-C=C–\) symmetric stretching vibrations in aromatic substances with a single or meta substituent. To simplify our analysis, the spectrum of methylbenzene, which is one of the main ingredients in gasoline, showing correlation to the interference peak as high as 0.996, is taken to approximate the peak profile at 1007 cm\(^{-1}\). The detailed spectrum of gasoline at 970 – 1020 cm\(^{-1}\) is shown in Fig. 4.

The following Eq. (1) shows the basic idea of IPS algorithm, in which \(I_{gas}(v)\) represents the spectral intensity of gasoline at \(v\), \(I_{MB}(v)\) is the pure spectral intensity of methylbenzene, and \(\omega_{MB}\) is its fitting weight in the gasoline spectrum. The optimization objective is to minimize the square sum of the residual error \(I_{gas}(v) - \omega_{MB}I_{MB}(v + \Delta v)\), in the range from 1007 to 1020 cm\(^{-1}\) by adjusting the weight, \(\omega_{MB}\), and the peak position bias, \(\Delta v\), caused by frequency shifts.\(^{22}\) Taking No. 1 gasoline spectrum for example, the fitting and residual lines are shown in Fig. 5, where the residual represents the actual intensity of benzene. The gasoline samples’ spectra subtracted by interference peaks are shown in Fig. 6.

\[
\min \sum_{v=997}^{1020} [I_{gas}(v) - \omega_{MB}I_{MB}(v + \Delta v)]^2 \\
\text{s.t.} \quad [I_{gas}(v) - \omega_{MB}I_{MB}(v + \Delta v) \geq 0] \quad (1) \\
\quad \Delta v \in [-2.2] 
\]

**Calibration model and comparative linearity**

For all of the following analysis, we first stochastically divide the samples into 3 sets, which include a training set containing 32 samples, a validation set with 15 samples and a test set with the rest 18 samples.

The basic idea for calibration is according to the principle of Raman spectroscopy that the Raman signal intensity of a component is mathematically proportional to its concentration. Thus the spectral intensity of benzene, \(I_{gas,Ben}\) at 997 cm\(^{-1}\) and saturated hydrocarbon, \(I_{gas,Sat}\) at 1450 cm\(^{-1}\) can be represented as follows:

\[
I_{gas,Ben} = c_{Ben}K_{Ben}A, \\
I_{gas,Sat} = c_{Sat}K_{Sat}A, \\
(2)
\]

where \(K_{Ben}\) and \(K_{Sat}\) are the cross section coefficient of benzene and saturated hydrocarbon, respectively, and are affected by the molecule structure. \(A\) is the total Raman intensity influenced by the measurement conditions, such as the effective laser power and the focal distance of the Raman probe; \(c_{Ben}\) and \(c_{Sat}\) indicate the concentration of benzene and saturated hydrocarbon respectively. The relationship of \(I_{gas,Ben}\) and \(c_{Ben}\) can be written as

\[
c_{Ben} = \lambda \frac{I_{gas,Ben}}{I_{gas,Sat}}. \\
(3)
\]

where \(\lambda = \frac{K_{Sat}}{K_{Ben}}\). Because the concentration of saturated hydrocarbon in gasoline is relatively stable and \(K_{Sat} , K_{Ben}\) both stay basically invariant for different samples, the coefficient \(\lambda\) can be regarded as an invariant. Also, since the intensity of saturated hydrocarbon, \(I_{gas,Sat}\) in the pretreated spectra equals 1, Eq. (3) can be shortened as

\[
c_{Ben} = \lambda I_{gas,Ben}. \\
(4)
\]

We compare the linear correlation coefficients for \(I_{gas,Ben}\) and \(c_{Ben}\) for the training set before and after IPS pretreatments. The result is like Fig. 7, showing that linear correlation coefficient after IPS is 0.934, while the other one is 0.784. The increase of the linear correlation coefficient indicates that the spectral signal extracted by the IPS method can effectively reflect the actual benzene information.

However, the linearity for \(c_{Ben}\) against \(I_{gas,Ben}\) is still
unsatisfactory because of other interferences, such as the spectral deformation and the frequency shift,\textsuperscript{23,24} caused by multi-components in gasoline. In order to be more robust against such fluctuations, the PLS method is applied so that component variables with more spectral information can be adopted.

Two analysis strategies are compared, including the PLS method with an IPS pretreatment (IPS-PLS) and PLS method directly without IPS; both approaches are applied within the same spectral range of 970 – 1020 cm\textsuperscript{-1}. The predictive ability of the methods is assessed by the root-mean-square-error of prediction (RMSEP) and the coefficient of determination ($R^2$):

\begin{equation}
\text{RMSEP} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{x}_i - x_i)^2},
\end{equation}

\begin{equation}
R^2 = 1 - \frac{\sum_{i=1}^{N} (\hat{x}_i - x_i)^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2},
\end{equation}

where $N$ is the number of samples, $\hat{x}_i$ is the predictive value, and $\bar{x}$ represents the average of the real value, $x_i$.

**Results and Discussion**

In this section, we verify the performance of IPS-PLS. In section “Validation with comparison of PLS”, its predictive performance is validated in comparison to the PLS method on the validation set from 2 different perspectives, containing different numbers of principal components and different numbers of training samples. Both methods were conducted based on the same data preprocessing in section “Different numbers of training samples”, as well as in the same spectral region of 970 - 1020 cm\textsuperscript{-1}. Section “Predictive ability of IPS-PLS with small-size training set” verifies the IPS-PLS method on the test set with small-size training sets to find the least number of training samples needed to reach the actual analysis standard.

**Validation with comparison of PLS**

**Different numbers of principal component.** The main parameter for the IPS-PLS and PLS methods is the number of principal components considered. We compare their predictive performance with different numbers of principal components, ranging from 2 to 11. The comparison result is shown in Fig. 8, which demonstrates that IPS-PLS method can perform more efficiently with fewer principal components; in this case, 4 principal components are enough for IPS-PLS to reach a satisfactory result. On the other hand, with the increasing of the principal component number, more information can be contained, so that direct PLS method can perform as well as IPS-PLS when there are enough components, as shown in Fig. 8, that 6 components are needed for PLS to reach a stable and accurate performance.

**Different numbers of training samples.** As a further study on the approach, the size of the training set is changed. We predicted the same set of validation samples while limiting the training set
size to a range from 6 to 24 samples. During the experiment, we formed subsets with certain amounts of samples randomly selected from the original training set, and then models of IPS-PLS with 4 principal components and PLS with 6 principal components were evaluated with the validation set samples. The above steps were repeated 10 times to obtain a more credible result.

Prediction results of the two methods are shown in Fig. 9, while Table 1 shows the detailed result when the sample number was 6, 12 and 20. On one hand, since fewer principal components are needed by IPS-PLS, it shows higher accuracy when the size of sample set is quite small, detailedly speaking, the IPS-PLS method can obtain an ideal result when the size of the training set is as small as 6; on the contrary, the direct PLS method needs at least 18 samples to reach a stable performance. On the other hand, even with the same amount of training samples, the direct PLS method is more dependent on their distribution. Taking the case with 20 samples for example, although the mean values of $R^2$ are both more than 0.93, $R^2$ for direct PLS distributes over a wide range from 0.8998 to 0.9527, as in some cases samples in the training set may not be able to represent those in the validation set; in contrast, that of the IPS-PLS method, which is based on more effective spectral information focuses within a smaller range of 0.9373 to 0.9611.

**Predictive ability of IPS-PLS with small-size training set**

To evaluate whether the predictive accuracy of IPS-PLS is suitable for actual application, we made predictions on the test set with subsets of 8 to 11 samples randomly selected from the original training set. IPS-PLS with 4 principal components is applied, and to get a more credible result, the experiment was repeated 20 times. The prediction results are shown in Fig. 10, in which the mean and max values of RMSEP for different numbers of training sample are plotted. According to the standard in ASTM 3606-10, which requires that the predictive error of benzene detection should be less than 0.1254 (%, v/v) in this case, the corresponding RMSEP of the analysis method should be less than 0.045 (%, v/v). As is shown in Fig. 10, with the decrease of sample number, the analysis result of IPS-PLS can keep high accuracy until there are as few as 8 training samples, showing the RMSEP with a mean value of 0.0356 and a max value of 0.0718. To keep its effectiveness in actual analysis, the test result shows that the IPS-PLS method with 4 principal components can remain effective with at least 9 training samples.

### Table 1 Influence of training sample number to the predict results

<table>
<thead>
<tr>
<th>Method</th>
<th>$R^2$</th>
<th>RMSEP, % (v/v)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Max</td>
<td>Min</td>
</tr>
<tr>
<td>IPS-PLSa</td>
<td>0.9287</td>
<td>0.8037</td>
</tr>
<tr>
<td>PLSa</td>
<td>0.8748</td>
<td>0.3334</td>
</tr>
<tr>
<td>IPS-PLSb</td>
<td>0.9558</td>
<td>0.9349</td>
</tr>
<tr>
<td>PLSb</td>
<td>0.9545</td>
<td>0.6839</td>
</tr>
<tr>
<td>IPS-PLSc</td>
<td>0.9611</td>
<td>0.9373</td>
</tr>
<tr>
<td>PLSc</td>
<td>0.9527</td>
<td>0.8998</td>
</tr>
</tbody>
</table>

a. Number of training samples = 6.  
   b. Number of training samples = 12.  
   c. Number of training samples = 20.

As described in this work, combined with Raman spectroscopy, we applied the PLS method to determine low-content benzene in gasoline based on our proposed pretreatment of IPS, and validated the result with comparisons of the direct PLS method. The experiment results demonstrate that the IPS pretreatment can extract the actual information of benzene by subtracting overlapping peaks; as a result, it can not only guarantee the accuracy of the PLS method, but also help decrease the principal components needed so as to extend its application to a training set with at least 9 samples.

Since in the actual industrial field, the condition is not ideal with insufficient training samples beforehand, the IPS-PLS method dealing with a small-size training set for benzene concentration prediction can have a promising application in the gasoline quality test.

### References

2005, 547, 188.