Chapter 6.
Classification and Quantification of Iron Phosphate Glass Samples

Abstract

The simulated nuclear waste in glass matrix has been studied using LIBS technique for the classification and quantification. Five samples of iron-phosphate glasses were synthesized by doping the three analytes viz., Cr, Sr and Ti with various concentrations. The spectra recorded using LIBS system were analyzed using principal component analysis (PCA), partial least square-discriminant analysis (PLS-DA) and support vector classification (SVC). Further, the quantitative analysis of the analytes was also carried out using advanced statistical methods such as partial least square regression (PLSR) and support vector regression (SVR).
6.1 Introduction

As compared to the other materials like ceramic and synroc, glass is a promising material to act as a host for storing the nuclear waste. Currently the borosilicate glass is widely used for this purpose. In spite of this, interest in iron phosphate glasses has been grown in the recent decades due to its advantages over borosilicate glasses. Although borosilicate glasses possess an excellent chemical durability, they require high melting temperature and have lower retention for some of the constituents of the nuclear waste [1]. High level waste (HLW) comprises of P₂O₅, chlorides, fluorides, nitrates, sulfates, and the heavy metal oxides of Sr, Cr, Cs, Zr, Zn, etc [2]. All of these components in the HLW are highly incompatible for the vitrification of borosilicate glass [3]. However, appropriate pre-treatment procedure helps in diluting the incompatible constituents in borosilicate glasses [4]. Yet, the procedure of pre-treatment of the waste is very expensive and also hindered due to larger volume of the final waste formed. Moreover the HLW has many actinides contained in it and such actinides are not soluble in high concentrations of borosilicate glass matrix [5]. The high amount of sulfates in HLW is also insoluble in borosilicate glass and causes the phase separation [6]. Phase separation reduces the durability of glass in the long run, and to prevent these complications borosilicate glasses are synthesized with limited amount of HLW as compared to the iron-phosphate glasses.

Thus the need for an alternative glass matrix is essential and iron-phosphate glasses are found to be promising due to their characteristics like low glass transition temperatures, and high thermal expansion coefficient [7,8]. Further, the features such as encapsulating large amount of waste, to melt at lower temperature, better chemical durability and reduced size of the final waste discerns the importance of iron-phosphate glasses [5,9,10]. The advantages of iron-phosphate glasses pave the way to bring down the cost, ease in transportation and can reduce the necessity of huge space required for the disposal of nuclear waste.

Investigating the glass composition followed by the vitrification of HLW in the glass matrix is an important step which ensures the quality of the product so that it is in the acceptable form [11-13]. Though the elemental analysis of these glasses can be carried out by many standard laboratory methods like ICP-MS, ICP-AES, etc. [14], they are time consuming; require rigorous sample preparation and most importantly the analysis of such samples using these methods is restricted to laboratory environments. Handling of the samples in nuclear power plants requires special arrangements which increases the complexity of the instrument.
Further, the aforementioned methods (ICP-AES and ICP-MS) are not suitable for the field applications. However, LIBS is a viable tool which can overcome the problems faced almost all sophisticated methods [15,16]. LIBS technique is fast in data collection, require no/minimal sample preparation and can perform effectively in hazardous/off-laboratory conditions [17,18].

Elemental analysis of glass materials using LIBS is of great interest [19] and also challenging due to the presence of several elements. As compared to the LIBS study on other samples, very few reports are observed in the literature relevant to the work on vitrified nuclear waste glass. Stratis et al. performed the Double Pulse (DP)-LIBS experiments on prototype samples used for nuclear waste immobilization like borosilicate glass matrix [11]. They have reported 11-20 times enhancement for Al, Ti and Fe using DP-LIBS and also claimed this method is more sensitive than single pulse-LIBS. Jung et al. have used the LIBS system to quantitatively study uranium (U) and europium (Eu) in glass matrix [20]. Sarkar et al. have carried out the quantitative analysis of U in barium borosilicate glass under air and Ar atmosphere [12]. The authors have observed enhancement in the emission intensity for the spectrum recorded under Ar conditions than the same recorded in air. This enhancement was mainly attributed to higher plasma temperature in Ar than air. It’s worth noting that all these studies used univariate methods for quantitative analysis.

Exploring the conventional univariate methods along with the multivariate approaches for the investigation of borosilicate glasses have been reported in the literature [21,22]. Yun et al. have carried out a detailed study on borosilicate glasses at elevated temperatures (1200 °C) [21,22]. The authors have discussed about the theoretical and experimental aspects of LIBS by verifying the local thermodynamic equilibrium (LTE) and also compared the performance of echelle spectrograph with Czerny-turner spectrograph. They have evaluated the LIBS spectra of HLLW in borosilicate glass matrix by univariate and multivariate regression (PLSR) methods. The comparison of LIBS with ICP-AES and XRF reveals that results obtained with LIBS agree with predetermined values. Similarly, Tripathi et al. have carried out quantitative analysis of various elements in lanthanide borosilicate glass matrix using multivariate (PCR and PLSR) methods [23]. Further, in a recent work, Singh et al. have analyzed the nuclear waste in barium borosilicate glass material and quantitatively studied F, Fe, Th, Cr, Al, Ni, and Sr using LIBS system [13]. The LIBS data was analyzed with PLSR method and they reported an average accuracy of 3-8% for the elements studied.
Chapter 6

From all the above reports, it is clear that there is an extensive interest to study different analytes in borosilicate glass matrix and there are no reports available on iron-phosphate glass matrix using LIBS. Therefore, qualitative and quantitative analysis of iron-phosphate glasses is studied here using LIBS combined with chemometric approaches.

6.2 Sample Compositions

The samples used for the analysis have been prepared using the melt-quench method and details of the sample preparation are discussed in the section 2.2.5 of chapter 2. Additionally, the iron phosphate glass samples are denoted as IP glass where the blank (undoped) sample is named as IP 0. The rest of the samples with increasing concentration of dopant are mentioned as IP 1, IP 2, IP 3 and IP 4. The Table 6.1 gives the amount of analyte present in each of the iron phosphate glass sample.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Cr (wt.%</th>
<th>Sr (wt.%</th>
<th>Ti (wt.%</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP 0</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>IP 1</td>
<td>0.426</td>
<td>0.359</td>
<td>0.196</td>
</tr>
<tr>
<td>IP 2</td>
<td>0.850</td>
<td>0.716</td>
<td>0.391</td>
</tr>
<tr>
<td>IP 3</td>
<td>1.273</td>
<td>1.073</td>
<td>0.586</td>
</tr>
<tr>
<td>IP 4</td>
<td>1.695</td>
<td>1.428</td>
<td>0.780</td>
</tr>
</tbody>
</table>

6.3 Methods of Analysis

In this Chapter, PCA, PLS-DA and SVC are used for the classification of the glass samples, whereas quantitative analysis of the doped elements is carried out using the PLSR and SVR. The theory behind the methods such as PCA and PLSR are already discussed in the Chapter 5. Hence, only the PLS-DA and SVMs (classification and quantification) are discussed in this section.

6.3.1 Partial Least square-Discriminant Analysis (PLS-DA)

It is essential to know the concepts of Linear Discriminant Analysis (LDA) before implementing the partial least square-discriminant analysis (PSL-DA). Hence, in this section, a brief overview of LDA is introduced prior to the discussion on PLS-DA. In case of discriminant analysis, appropriate decision boundaries are determined where the number of regions is equal to the number of classes. For example if there are three classes of samples, then the hyperspace of the variables are divided into same number of regions. LDA is one of the supervised methods, which can classify more than two classes of samples. LDA is
described with basic ideas of the concept of probabilities. The first use of this method dates back to 1936 when Fisher published his work on classification of biological samples [24]. Although LDA is a simple discriminant methods, its usage for highly collinear data does not always ensure promising results. To be more precise, LDA fails when the number of variables exceeds the number of samples/objects. In spite of these drawbacks, the method has been explored to investigate the LIBS data. Various groups have demonstrated the use of LDA for the discrimination of samples using LIBS data. Jantzi and Almirall have studied the discrimination of two soil samples for forensic application [25]. The use of LDA on LIBS data enabled the authors to obtain 99.4% correct classification. In another study, LDA as well as principal component analysis (PCA) has been used to classify the brick samples for archaeological applications [26]. The samples collected from seven different locations have been recorded using both table-top and stand-off LIBS.

Partial least square (PLS) is another popular methods in chemometrics and it has been in use from several years for the quantitative analysis [27]. Recently, the method was extended to address the classification problems of different samples [28-32]. The main advantage of PLS-DA is that it can be implemented in model building even when the number of variables exceeds the number of samples/spectra [33]. The variables here are defined as the intensity values of the emission lines at specific wavelengths/pixels. In PLS-DA, the maximum covariance that exists between the independent \( (X) \) and dependent \( (Y) \) variables helps in classification of the samples. The \( Y \) variables now act as qualitative values or in other words, they define the individual classes. The PLS-DA, initially generates new variables known as latent variables. There are two approaches to PLS-DA namely, PLS1 and PLS2. The former is used for the grouping of two objects and the latter is employed for the classification of more than two groups. In PLS1, the model evaluates the samples of two classes where the independent variables \( (X) \) contain the information of intensities of emission lines at each wavelength. The vector \( (y) \) consists of the information of the class it belongs to and is usually represented by either 0 or 1. However, the same vector \( (y) \) in PLSR contains the information of analyte concentration. Mathematically, PLS-DA can be explained using the following equations for the classification of only two objects.

\[
X = TP + E \quad \ldots (1)
\]

\[
y = Tq + f \quad \ldots \ldots (2)
\]

Where \( T \) is known as the score matrix and \( P \) and \( q \) are matrices of loadings. \( E \) and \( f \) are known as residual or error matrices. Scores are the projection coordinates of the objects in
which any two score vectors are uncorrelated. With the help of scores one can visualize the grouping of different samples. Loadings also known as component coefficients represent the weights of the original variable. Initially, the weight vector (\(w\)) and scores (\(t\)) of the matrix \(X\) are calculated. Later, loadings of the matrix \(X\) and \(y\) helps in computing the residual matrices. The class of an unknown sample is finally determined by the estimation of regression coefficients. The brief mathematical formulation explained above can be extended to determine the classification of more than two samples. In other words, if there are more than two classes of samples then the vector (\(y\)) in PLS1 has to be replaced by matrix \(Y\). The columns of matrix \(Y\) represent the individual class of the samples and each column is evaluated step-by-step using PLS1 model. In other words, if there are 4 types of samples, the PLS-DA model is constructed using four individual PLS1-DA models. The detailed algorithm for PLS-DA are available in the literature [34-36]. Though the mathematical formulation is similar to the one in PLS regression, the computations in PLS-DA differ slightly in such a way that the class membership is estimated instead of finding the analyte concentration.

### 6.3.2 Support Vector Machines (SVM)

Corinna Cortes and Vladimir Vapnik developed SVMs and demonstrated the method for classification studies [37]. Later, Smola and Scholkopf implemented the SVMs for regression problems [38]. In this section, the support vector classification (SVC) is introduced first followed by a discussion on support vector regression (SVR). SVM has an additional advantage that it can deal with noisy and non-linear spectroscopic information. The results obtained by SVM are deterministic in nature and prevents the randomness. The randomness is often encountered in case of ANNs which provide different solutions on repeating the procedure/computations [39]. Further, the manifestation of improved classification and quantification in SVM is mainly attributed to its important feature, viz. structural risk minimization (SRM) principle, avoiding the empirical risk. SRM has an important role in SVM which balances the empirical error and model complexity. In other words, SVMs maintain generalization ability circumventing the optimization of given limited calibration data set. SVMs can efficiently transform the linearly inseparable data into a linearly separable one through increasing the dimensions. Suppose if the data is not grouped in the 2 dimensions then it can be separated by adding one more dimension so that the classification is visualized in 3 dimensions. However, this is achieved through a mathematical operation known as kernels using which the dimensions can be increased. There are four types of kernels namely,
linear, polynomial, radial basis function (RBF) and sigmoid using which the inner product (measure of similarity) of the feature space (higher linear separable dimensions) is determined. Mathematically, the general forms of these four kernels are given by the following equations.

\[ k(x_i, x_j) = (ax_i^t x_j + b) \quad \ldots \ldots \ldots (3) \]  (Linear kernel)

\[ k(x_i, x_j) = (ax_i^t x_j + b)^n \quad \ldots \ldots \ldots (4) \]  (Polynomial kernel)

\[ k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad \ldots \ldots (5) \]  (Radial Basis Function)

\[ k(x_i, x_j) = \tanh(ax_i^t x_j + b) \quad \ldots \ldots \ldots (6) \]  (Sigmoid kernel)

Where \( x_i \) and \( x_j \) are the two vectors of the recorded spectra for \( i^{th} \) and \( j^{th} \) samples, \( n \) is the degree of polynomial, \( \gamma \) is the tuning parameter and \( a \) and \( b \) are constants.

With this basic introduction to SVMs, the SVC can be further explained as follows. If the dataset contains two types of samples then the \( y_i \in [-1, 1] \) where \( i = 1, 2, 3 \ldots N \). In other words, if the sample belongs to positive class then \( y_i \) is 1 and if it belongs to negative class then \( y_i \) is -1. However, the classification problems are not always trivial with only two class of samples but becomes complex when confronted by many classes which may be of overlapping nature. Under such situations, the model must be flexible enough to consider the samples (support vectors) that are present not only on the margin but also the ones near to it. The reduced set of such samples ubiquitous near the hyperplane are known as support vectors (SV) \[39\]. However, the fact that considering the SVs near the hyperplane results in poor classification of the samples, yet this poor classification can be determined by slack variables (\( \xi \)). The slack variables provide the information of distant location of samples from the hyperplane. These variables comprises of cost function, which optimize the SVM algorithm in such a way that the model has small error and the large margin. The \( \xi \) may have values of 0, 1 or >1 in which; \( \xi = 0 \) represents the correct classification and \( \xi = 1 \) disseminates that the sample is on the margin. Further, the incorrect classification is represented by \( \xi > 1 \).

Briefly, the procedure to perform the SVC can be summarized in the following two steps; first, the data is transformed using the kernel function to map original data from lower dimensional space to higher dimensional space. Secondly, the margin between the two classes of the samples is maximized by optimally separating hyperplane.

Support vector machines can be also employed to solve the regression problems and the method is generally known as support vector regression (SVR) machines \[40\]. The
classification problem has been taken forward to solve the regression problem by making use of $\varepsilon$-insensitive loss function. The loss is zero if the predicted value is within the $\varepsilon$ tube otherwise the loss is equal to the difference between the predicted value and the radius $\varepsilon$ of the tube [41]. In case of binary classification, the $y_i$ can take either -1 or +1 but in case of SVR the $y_i$ corresponds to the analyte concentrations instead of class label. In SVR, the optimal hyperplane/boundary is determined, in which the distance from all the samples is minimum.

However, the errors associated with the measurements of the calibration and validation cause the samples to fall outside the hyperplane. Hence, to deal with such samples the flexibility is introduced in the model, in the form of slack variables, similar to SVC. In SVR, the sample vectors are denoted as $x_i$ (measured spectra) and the respective response (analyte concentration) is represented by $y_i$. Suppose if the dataset ($D$) has $N$ number of samples then $D = \{(x_i, y_i)\}_{i=1}^N$. Similar to SVC, the concept of SVM is used to perform the regression in which the data is transformed from lower dimensional space into higher dimensional space followed by establishing the relation between $x_i$ and $y_i$.

6.4 Selection of Wavelength Range for the Data Analysis

To investigate the predictive capabilities of the chemometric methods, sparse and broad region of the LIBS spectrum are used to carry out the classification. The sparse region of the LIBS spectrum is chosen in such a way that the region must have stronger emission lines of the dopants. The intensity value at each wavelength is referred as variable and 1330 such variables of Cr, Sr and Ti (shown in the Table 6.2) are merged together to use it for the classification studies.

Although, the LIBS spectrum recorded in the present experiment consist of emissions from 250 nm to 850 nm, all this information is not used. The broad region of the LIBS spectrum chosen for the current analysis is 300 nm to 800 nm in which the number of variables is 16037. The quantum efficiency of the echelle spectrograph is high in 300 nm - 800 nm region and this in turn capable to give stronger emission lines. Hence it is anticipated that better predictions can be achieved for broad region. The emission lines appear outside this region are comparatively weak and are not considered in the analysis. The selected regions of Cr, Sr and Ti independently have 625, 264 and 441 variables respectively. The spectral regions of these elements shown in Table 6.2 are used separately to carry out the quantitative analysis through PLSR and SVR methods.
### Table 6.2: LIBS spectral regions selected for the multivariate analysis

<table>
<thead>
<tr>
<th>Doped elements</th>
<th>Truncated regions of LIBS spectra</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cr</td>
<td>283.00-287.00 + 357.01-361.01 + 425.01-430.05</td>
</tr>
<tr>
<td>Sr</td>
<td>406.50-408.51 + 421.00-422.00 + 460.01-463.01 + 545.02-552.00</td>
</tr>
<tr>
<td>Ti</td>
<td>322.51-324.50 + 334.01-340.00 + 365.00-366.00</td>
</tr>
</tbody>
</table>

### 6.5 Data Pre-processing

The data was mean-centered prior to the analysis so that all the independent variables (intensity values at each wavelength) in the matrix $X$ have mean zero. However, the distances between the variables are not affected in the mean-centered data. Before performing the classification, the data was further subjected to Savitzky-Golay smoothing and results demonstrated no marginal improvement with this pre-processing step. No baseline correction was required since the LIBS spectra had been recorded by the echelle spectrograph with ICCD detector. In nutshell, apart from mean centering, no other pre-processing steps (baseline correction, smoothing, etc) were used.

### 6.6 Cross Validation

A total of thirty spectra recorded from each sample are split into calibration set (25) and validation set (5). Both internal and external cross-validation methods were employed. The full cross-validation also known as leave one out-cross-validation (LOO-CV) method was employed for the quantitative analysis using PLSR. However, SVR is carried out with the method of 10-segments cross-validation. In 10-segments cross-validation method the original data is divided equally into 10 subsets, in which the 9 subsets are used for the calibration and the remaining one subset is used for the validation of the data. The LOO-CV and the 10-segments cross-validation are methods of internal cross-validation. The analysis is also carried out through the method of external cross-validation where the performance of the calibration model is evaluated on a validation set.

### 6.7 Results and Discussion

#### 6.7.1 Qualitative Analysis

LIBS spectra of iron phosphate glasses were recorded by keeping the detector delay and gate width as 0.7 μs and 5.0 μs respectively. Laser beam diameter of 6 mm having the average pulse energy of 24 mJ is used in the experiment. The laser beam is focused using 200 mm focal length lens and focus spot is maintained to 3 mm inside the samples whose thickness varying from 6-8 mm. LIBS spectra of all the samples were recorded using second harmonic
wavelength (532 nm) of Nd-YAG laser, where, each spectrum is an average of 230 laser pulses. The emission lines of LIBS spectra are identified by comparing the wavelengths of respective emission lines with NIST atomic spectral database. The characteristic spectral lines of various doped elements in iron phosphate glass samples have been successfully identified. Figure 6.1 represents the LIBS spectrum of iron phosphate (IP) glasses which includes the (a) blank (undoped) sample and (b) glass sample doped with Cr, Sr and Ti. The LIBS spectrum includes neutral and singly ionized emissions of several elements present in iron phosphate glass. The spectrum looks complex due to the presence of elements like Fe and Ti. These elements can be found in between the region of 250 to 850 nm, where, thousands of NIST listed emission lines for Fe (10071) and Ti (3498) are available. Besides this, the emissions of Cr, Sr, K, O, P and Na of are shown in Table 6.3.

Figure 6.1: LIBS spectrum of: (a) blank and (b) Cr, Sr and Ti doped iron phosphate glass
Table 6.3: Major atomic and ionic emission lines of the elements in the iron phosphate glass sample

<table>
<thead>
<tr>
<th>Element</th>
<th>Atomic Line</th>
<th>Emission lines (nm)</th>
<th>Ionic Line</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>255.32</td>
<td>645.99, 650.34</td>
<td></td>
</tr>
<tr>
<td>Fe</td>
<td>404.58, 406.35, 438.35, 373.71, 374.94, 375.82, 382.04</td>
<td>258.58, 259.93</td>
<td></td>
</tr>
<tr>
<td>Na</td>
<td>330.23</td>
<td>309.27, 588.99, 589.59</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>777.41, 777.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>321.71, 321.76, 404.41, 494.20, 766.48, 769.89</td>
<td>430.53, 430.91</td>
<td></td>
</tr>
<tr>
<td>Cr</td>
<td>359.35, 360.53, 425.43, 427.48, 428.97, 429.61</td>
<td>283.04, 283.56, 283.87, 284.00, 284.98, 285.56, 286.25, 286.76, 407.77, 421.55</td>
<td></td>
</tr>
<tr>
<td>Sr</td>
<td>406.09, 407.61, 460.73, 548.61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ti</td>
<td>365.34, 365.80</td>
<td>323.45, 323.61, 323.90, 324.19, 334.18, 334.94, 336.12, 337.27, 338.03, 338.37, 338.78</td>
<td></td>
</tr>
</tbody>
</table>

6.7.2 Classification Studies

The classification studies of glass samples are performed by unsupervised (PCA) and supervised (PLS-DA and SVC) methods. The classification using PCA enables to visualize the structure of the data or in other words, explore the clustering/grouping of the samples. They are also known as exploratory methods and are usually employed before the use of supervised methods to examine the hidden information about the clustering of a given data set. PCA is sometimes used for the identification of spurious data points/outliers in order to obtain improved classification. The identification of such data sets is demonstrated in Chapter 4 through the concept of Mahalanobis distance. In this work, the performance of the chemometric methods has been explored without removing any outliers. The sample-to-sample variations with respect to the changes in minor concentration of the doped elements are studied. Chemometric methods employed for the same are capable enough to identify these minor changes. PCA being an unsupervised classification method, needs the complete data without a validation set. On the contrary, PLS-DA and SVC are supervised methods and hence the data is divided as calibration and validation sets.

Figure 6.2 gives the overview of classification using PCA and PLS-DA in which the grouping of each sample can be visualized. The score plots of PCA and PLS-DA for broad region of the LIBS spectra are depicted in Figure 6.2 (a) and (b) respectively. Similarly, the PCA and PLS-DA plots of sparse region are shown in respective plots of Figure 6.2 (c) and (d). The PCA score plots gave better classification of blank sample and the rest of the
samples are found close to each other. Since PCA was not successful in separating all the samples, the supervised methods are employed for the classification.

![PCA and PLS-DA Score Plots](image)

**Figure 6.2:** Classification of iron-phosphate glass using PCA (a and c) and PLS-DA (b and d) methods

The PLS-DA score plot shown in Figure 6.2 (b) and (d) demonstrates the successful classification of IP 0 and IP 4 samples. The PLS-DA score plots contain both calibration and validation sets in which the open and closed symbols represent the calibration and validation set respectively. The performance of PLS-DA is further evaluated through the figures of merits such as sensitivity, specificity and accuracy. The formulae to calculate the figures of merit are given in the section 5.3.1 of Chapter 5 for the PLS-DA predictions and the values obtained are presented in Table 6.4.

**Table 6.4:** The figures of merit for the PLS-DA classification

<table>
<thead>
<tr>
<th>Approach</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broad region</td>
<td>68</td>
<td>97</td>
<td>91</td>
</tr>
<tr>
<td>Sparse region</td>
<td>77</td>
<td>95</td>
<td>89</td>
</tr>
</tbody>
</table>
In spite of achieving close to 90% accurate classification by PLS-DA, the LIBS data is further investigated using SVC. The linear kernel of SVC is implemented to analyze the LIBS data, in which the grid search is given to find the optimum value of the parameter C (regularization constant). Tuning the hyper-parameters is an important step before obtaining the SVM models and it is performed using the grid search operation. In case of RBF kernel, the search of hyper-parameters of both C and γ is necessary, however, in case of linear kernel only C is required. In the present work, the software itself generates the hyper-parameters however, the optimum parameters are selected based on the lowest RMSECV. Over fitting of the model is avoided by finding the value of C for broad and sparse regions of the LIBS spectrum and they are found to be 1.00 and 0.01 respectively. The linear kernel has an advantage, it is not prone to the over fitting of the model, meanwhile, it has the ability to handle large number of variables. Similar to PLS-DA, the SVC is used for the classification of samples on both broad and sparse regions where the results are presented in the form of confusion matrix. In machine learning, the confusion matrix (error matrix) is the matrix used to represent the performance of the classification. The rows and columns of the matrix represent the actual and predicted number of spectra/samples respectively. In case of broad region (Table 6.5), both the calibration and validation sets of all the samples are correctly grouped together. Additionally, in sparse region (Table 6.6), the classification is successful for all the samples of calibration set. Similarly, the validation set is also successful in classifying all the samples except the one in IP 2.

Table 6.5: Confusion matrix for the calibration and validation set analyzed on broad LIBS spectra

<table>
<thead>
<tr>
<th>Calibration set</th>
<th>IP 0</th>
<th>IP 1</th>
<th>IP 2</th>
<th>IP 3</th>
<th>IP 4</th>
<th>Validation set</th>
<th>IP 0</th>
<th>IP 1</th>
<th>IP 2</th>
<th>IP 3</th>
<th>IP 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP 0</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>IP 0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IP 1</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>IP 1</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IP 2</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>IP 2</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IP 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>IP 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>IP 4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>IP 4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 6.6: Confusion matrix for the calibration and validation set analyzed on sparse LIBS spectra

<table>
<thead>
<tr>
<th>Calibration set</th>
<th>IP 0</th>
<th>IP 1</th>
<th>IP 2</th>
<th>IP 3</th>
<th>IP 4</th>
<th>Validation set</th>
<th>IP 0</th>
<th>IP 1</th>
<th>IP 2</th>
<th>IP 3</th>
<th>IP 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP 0</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>IP 0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IP 1</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>IP 1</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IP 2</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>IP 2</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IP 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>IP 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>IP 4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>IP 4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>
Chapter 6

The performance of the SVC is explained through the error rate and accuracy, which are given by the following formula.

\[
\text{Error rate (\%)} = \left( \frac{\text{sum of misclassified spectra}}{\text{total number of spectra}} \right) \times 100 \quad \text{(7)}
\]

\[
\text{Accuracy (\%)} = (1 - \text{error rate}) \times 100 \quad \text{(8)}
\]

Using equations (7) and (8), the error rate and the accuracy were calculated for both broad and sparse regions. These equations are used to evaluate the confusion matrix of validation sets. The lowest error rate of 0\% is obtained for the broad region, whereas it is 4\% for the sparse region. Further, the accuracy of the broad region is found to be excellent (100\%) and for the sparse region, it is found to be 96\%. The discrepancy in the accuracy of the methods is attributed to the difference in the number of characteristic emission lines available for the analysis. In other words, the better accuracy is anticipated in the broad region where the emission features of major elements (Fe, P and O) contributed for the classification. Table 6.7 summarizes the figures of merit of SVC.

**Table 6.7:** Figures of merit for the SVC of iron phosphate glass samples

<table>
<thead>
<tr>
<th>Approach</th>
<th>Error rate (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broad region</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Sparse region</td>
<td>4</td>
<td>96</td>
</tr>
</tbody>
</table>

### 6.7.3 Quantitative Analysis

Exploring large number of variables (p) and limited number of samples (n) is detrimental to the performance of the chemometric methods. In most of the cases, the number of samples available for the analysis is limited. However, analysis with chemometric methods is often robust when n>p. Nevertheless, in some cases, such situations are far from reality, where, the analysis needs to be carried out with small n and large p. Additionally, the data recorded using sophisticated modern instruments such as echelle spectrograph coupled with ICCD provides the LIBS spectrum with larger number of variables. Consequently, one needs to select the variables pertinent to the analytes of interest and hence the LIBS spectra are analyzed carefully to identify such regions.

The quantitative analysis of Cr, Sr and Ti is carried out using the kernel PLS to obtain the calibration curves. An overview of this PLSR analysis is described in this section. The regression coefficients obtained from the PLS algorithm identify the significance of the individual variable in the X matrix. In other words, magnitude of the characteristic emission
peak in regression coefficient plot demonstrates its contribution to construct the calibration curve. For clarity, selected regions of the spectrum are shown in the regression coefficient plots with respect to the number of factors. A careful observation of these plots illustrates that for a particular factor the strength of the regression coefficient is maximum, whereas, for rest of the factors they are either weak or found to be noisy.

**Figure 6.3:** Regression coefficients versus wavelength (nm) for (a) Cr, (b) Sr and (c) Ti
The regression coefficient plots are shown in Figure 6.3(a) (b) and (c) for Cr, Sr and Ti respectively.

As a rule, the PLSR method is employed only after the careful investigation of root mean square error (RMSE) with respect to the number of factors. The interpretation of the RMSE versus factors plot reveals whether the calibration curve is underfit or overfit. The calibration curve is said to be underfit when the number of factors selected is less than the optimal number of factors. On the contrary, adding more factors higher than the optimal number of factors overfits the model. Therefore, one should look for the optimal number factors for obtaining the better prediction capabilities of the calibration model. To know the optimal number of factors, the following procedure is followed (i) the particular factor must has lowest RMSE and (ii) the factor must have nearly same RMSEC and RMSECV. As anticipated earlier in Chapter 4, the accuracy of the multivariate calibration models is given in terms of RMSEs. The RMSEs disseminate the closeness of predicted and reference values and have the units of analyte concentration (for example wt.%).

Figure 6.4: RMSE versus number of factors for (a) Cr, (b) Sr and (c) Ti
Chapter 6

The Figure 6.4(a), (b) and (c) shows the RMSE versus number of factor plots for Cr, Sr and Ti respectively. These plots reveal that Cr, Sr and Ti have the RMSEC of 0.167, 0.135 and 0.077 wt.% and they are very close to RMSECV of the respective elements.

The spectral information required to construct the calibration curve is hidden within the factors of PLS and is shown in the plot of variance versus the number of factors. To be specific, the initial few factors represent maximum information about analytes of interest and the remaining factors are of least importance to construct the model. Further, the total variance can be 100% on adding its contributions by all the individual factors. However, on using the selected number of factors one can utilize the contribution arising only from the analytes, ignoring the factors that can add the noise. In most of the cases, the variance (the information related to the analytes) varies from 70% to 90%. The plots for the variance versus number of factors for Cr, Sr and Ti are shown in Figure 6.5(a), (b) and (c) respectively.

![Figure 6.5: Variance versus number of factors for (a) Cr, (b) Sr and (c) Ti](image-url)
In short, from the Figure 6.3, 6.4 and 6.5, it was inferred that the calibration curves for Cr, Sr and Ti can be constructed by selecting the optimum number of factors of 2, 3 and 1 respectively.

In addition to the PLSR, SVR is further investigated to perform the quantitative analysis of the doped elements in iron-phosphate glass. The analysis is carried out with two different types of SVM namely, nu (υ)-SVR and epsilon (ε)-SVR, where, in the kernel function of ε-SVR is further modified to develop the υ-SVR [42,43]. The parameter ε is a positive number and it is always predefined, whereas, the υ is known as sparsity parameter [40]. The initial investigation of the LIBS data to quantify Ti using υ-SVR provided poor predictions on cross validated data. Hence, the quantification is performed using ε-SVR. Additionally, the type of the kernel is also examined before performing the quantitative analysis. To investigate this Ti is analyzed using both linear and RBF kernel of SVMs. The prediction performance investigated on validation set using linear and RBF kernels resulted in RMSECV of 0.094 and 0.100 wt.% respectively. Thus, analysis carried out using linear kernel was found to be almost similar to that of RBF kernel. The non-linear kernels such as polynomial and RBF kernels can be implemented only if the linear kernels become inefficient to perform the quantitative analysis. In the present work, non-linear kernels (polynomial and RBF) tend to overfit the model and hence the linear kernel is employed to carry out the SVR. Further, SVR provided better values of $R^2$ for the calibration data set as compared to the PLSR. However, the internal cross-validation data has almost same $R^2$. The LIBS data is analyzed with linear kernel and hence C is the only tuning parameter to obtain the best calibration models. The values of C used for obtaining the calibration curves of Cr, Sr and Ti are 0.01, 1.00 and 1.00 respectively. Table 6.8 given below summarizes the correlation coefficient ($R^2$) obtained from both PLSR and SVR methods.

<table>
<thead>
<tr>
<th>Element</th>
<th>Correlation coefficient ($R^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PLSR</td>
</tr>
<tr>
<td></td>
<td>Calibration</td>
</tr>
<tr>
<td>Cr</td>
<td>0.921</td>
</tr>
<tr>
<td>Sr</td>
<td>0.927</td>
</tr>
<tr>
<td>Ti</td>
<td>0.921</td>
</tr>
</tbody>
</table>

Similar to PLSR, the results of SVR are presented through the RMSEC and RMSECV. From these results, it is found that the prediction accuracy of SVR is same as that of PLSR. The
accuracy of the PLSR and SVR presented in Table 6.9 in terms of RMSEC and RMSECV are plotted and the same is depicted in Figure 6.6.

**Table 6.9:** Accuracy of PLSR and SVR for the quantitative analysis of the doped elements

<table>
<thead>
<tr>
<th>Element</th>
<th>PLSR</th>
<th></th>
<th></th>
<th></th>
<th>SVR</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Calibration set</td>
<td>Validation set</td>
<td>Calibration set</td>
<td>Validation set</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RMSEC</td>
<td>RMSECV</td>
<td>RMSEC</td>
<td>RMSECV</td>
<td>RMSEC</td>
<td>RMSECV</td>
<td>RMSECV</td>
<td></td>
</tr>
<tr>
<td>Cr</td>
<td>0.167</td>
<td>0.209</td>
<td>0.163</td>
<td>0.076</td>
<td>0.198</td>
<td>0.220</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sr</td>
<td>0.135</td>
<td>0.140</td>
<td>0.209</td>
<td>0.058</td>
<td>0.147</td>
<td>0.164</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ti</td>
<td>0.077</td>
<td>0.078</td>
<td>0.088</td>
<td>0.033</td>
<td>0.069</td>
<td>0.094</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 6.6: The RMSECV of the calibration curves for (a) internal and (b) external-cross validated data

6.8 Conclusion

Advanced chemometric methods such as PCA, PLS-DA and SVC are employed to classify the LIBS data of iron phosphate glass samples. A comparison of classification methods reveals the efficacies of these methods in determining the classes of glass samples based on their composition. The classification results of PCA and PLS-DA are presented through the score plots, whereas, classification using SVC is given in terms of confusion matrix. The performance evaluation of the classification methods are done through the accuracy of the validation set. Among all the classification methods, better results were obtained from SVC. The quantitative analysis carried out on Cr, Sr and Ti using PLSR and SVR provided similar results. Nevertheless, evaluation of the performance of these methods through RMSECV demonstrates that the LIBS combined with chemometrics methods is a potential approach to analyze glass samples.
6.9 References


[40] H. Li, Y. Liang, and Q. Xu, Support vector machines and its applications in chemistry, Chemometrics and Intelligent Laboratory Systems 95, 188 (2009).

