Predicting pH from SERS Data

Zhang Yichi

August 2014

1 Introduction

Blood tests are often used in health care to determine physiological and biochemical states of patients. But using the traditional way, we need to draw blood from patients first and it does take some time to have some tests on it.

Recently, the chemists found the gold nanoshells can be used as intracellular sensors based on surface-enhanced Raman scattering (SERS) and these materials exhibit low toxicity to the cells of interest. This is a very good property since we can just put the nanoshell into the blood vessel of patients and measure redox potential or pH values of patients' blood instantly.

So when given the spectrum, the problem is how to predict the pH value. We have tried 4 regression methods, that is principal component regression (PCR), partial least squared regression (PLSR), lasso regression and kernel regression, on the data.

2 Data

There are 120 samples in the dataset by 2 chips. For each chip, there are 5 replications for 12 pH values, that is 60 samples. It's a 1044 dimension vector for each sample, and each dimension represents a Raman intensity for a Raman shift.

There are 2 datasets we have got for experiments. The first dataset we used was produced with the order that pH value is increasing. We have found in the dataset that the intensity is lower and lower when the pH value is greater than 7 as show in figure 1.

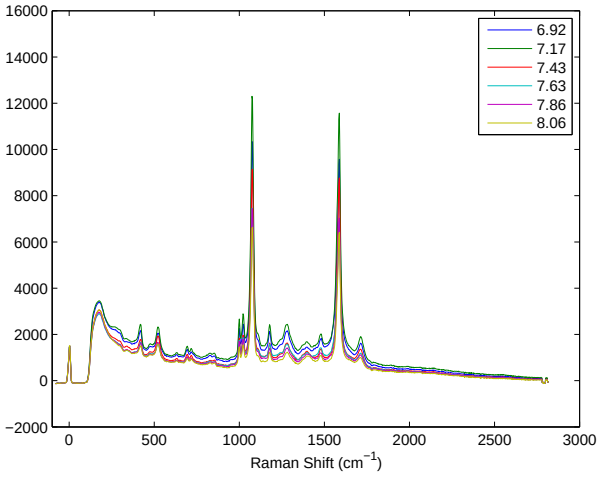

Figure 1: The mean spectrum for different pH values

However, the chemists have told us that the intensity is lower and lower maybe due to the systematic loss of nanoparticles through time.

Thus, we got a another set of data measured with randomized order reproduced by chemists. And the experiments and analyses below are based on the randomized dataset.

2.1 Raw Data

As for the raw data, I plot 5 replications in 1 plot for each pH values as shown in figure 2.

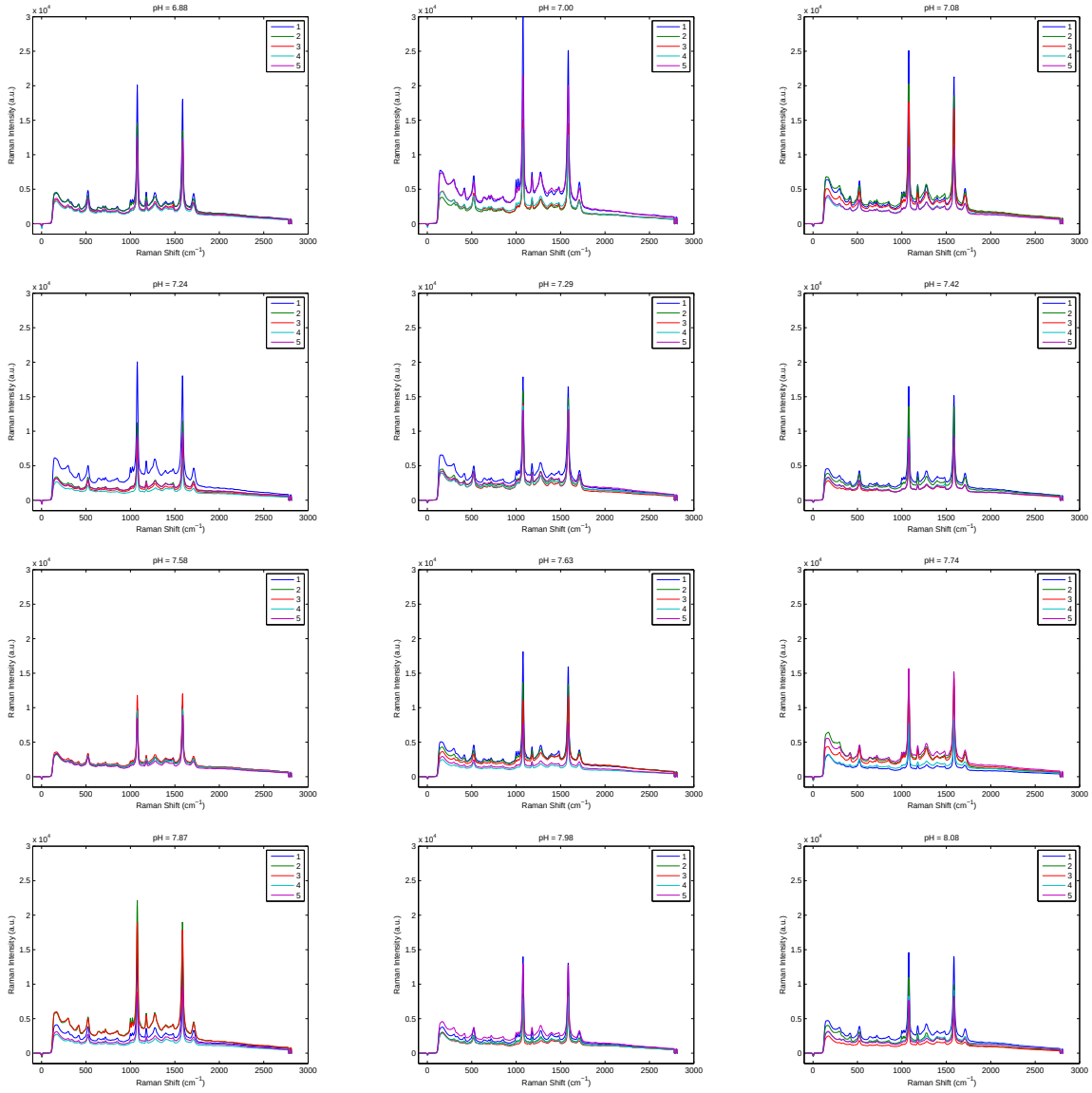

Figure 2: Plots for each pH values of raw data

We can find that when the pH value is increasing, not like in the previous dataset, the peak isn't lower and lower as the pH value is increasing all the time.

Meanwhile, we can find that for some pH values, the curves look dramatically different for the same pH value. Suggested by chemists, normalization is a good way to avoid such system error. And the method we used is normalizing the total area under the curve.

2.2 Normalization

As mentioned in previous section, I have normalized the total area under the curve for every spectrum as shown in figure 3.

As we can see from the figure that the normalization does eliminate the differences among samples of the same pH values.

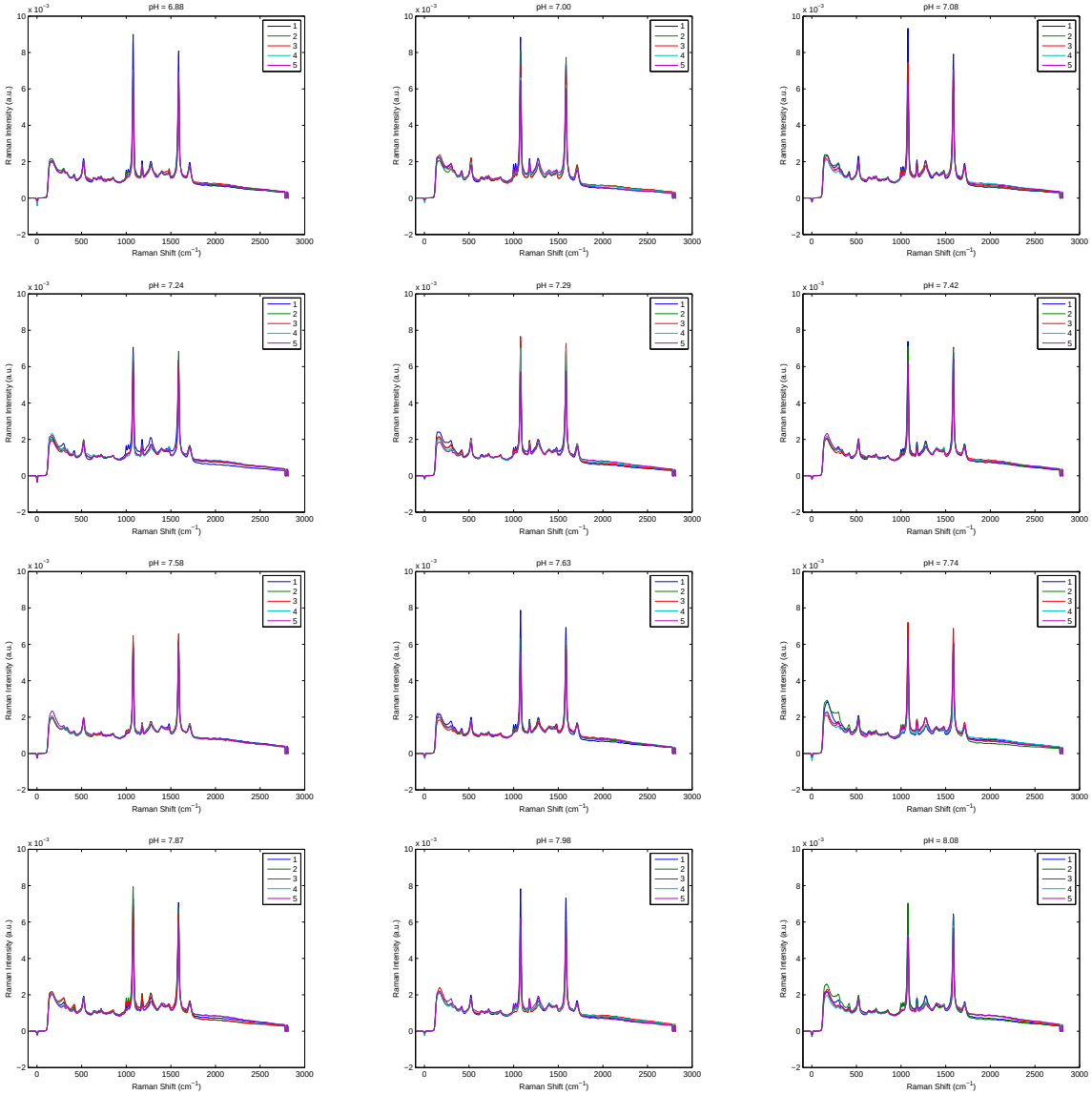

Figure 3: Plots for each pH values of data after normalization

3 Methods

There are only 60 samples. Since the number of sample is very small, we'll use cross validation to judge which method is better.

For cross validation, we divide samples into 5 folds. For each fold, there are not two samples with the same pH value. Every time, we use 4 folds for training and 1 fold for testing, and use standardized mean squared error (SMSE) for evaluation.

$$\text{SMSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{1}$$

3.1 Linear Regression

The basic method of regression is linear regression. The simplest linear model is one that involves a linear combination of the spectrum

$$y(\mathbf{x}, \mathbf{v}) = v_0 + v_1 x_1 \ldots + v_D x_D \tag{2}$$

where $\mathbf{x} = (x_1, \ldots, x_D)^T$ and here D is 1044 in our dataset. The key property of this model is that it is a linear function of the parameter $v_0, v_1, \ldots, v_D$.

However, it may be not possible for all the points representing all the spectra in the training data to be all on the same plane. So what we're going to do is to minimize

$$J(\mathbf{v}) = \sum_{i=1}^{m} (y(\mathbf{x}, \mathbf{v})^{(i)} - \mathrm{pH}^{(i)})^2 \tag{3}$$

which minimize the total difference between the predict pH value and the observed pH value. This leads to a closed-form expression for the estimated value as shown below.

$$\hat{\mathbf{v}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathrm{pH} \tag{4}$$

As the dimension of the spectrum is 1044 dimensions and the number of samples is only 60. So it's not possible to directly use the method mentioned above and there are 2 methods mentioned below which can handle this situation.

3.1.1 Principal Component Regression

Principal component regression (PCR) is a regression analysis technique that is based on principal component analysis (PCA).

PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components we used is less than the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.

Using the PCA first can make the dimension of spectrum less than 60, and then we can use traditional linear regression on the data.

3.1.2 Partial Least Squares Regression

Partial least squares regression (PLSR) is a statistical method that bears some relation to principal components regression. Instead of finding hyperplanes of minimum variance between the response and independent variables, it finds a linear regression model by projecting the predicted variables and the observable variables to a new space.

For PLS considers not only the spectra but also the pH values corresponding to them, PLSR has better performance in a lot of cases than PCR.

3.2 Lasso Regression

Lasso regression is a regularized version of linear regression which can avoid over-fitting. It minimizes

$$J(\mathbf{v}) = \sum_{i=1}^{m} (y(\mathbf{x}, \mathbf{v})^{(i)} - \mathrm{pH}^{(i)})^2 + \alpha ||\mathbf{v}|| \tag{5}$$

Here, α is an important parameter to control the intensity of regularization. Large α is, more numbers of values in $\mathbf{v}$ will be equal to 0 or nearly 0.

3.3 Kernel Regression

Kernel regression is quite a different method from the methods mentioned above.

Before introducing it, we'll introduce a method for classification called k-NN. In this method, for every new sample to be classified, we choose first k nearest samples for it and count which class most of the samples belong to. Normally, we choose Euclid distance to calculate nearest samples.

And for regression, we cannot only count. We should combine the pH values of its neighbours together. And here, we use Gaussian kernel for the weight of each pH values, and we can then predict the value of data in testing set.

4 Results

4.1 Principal Component Regression

As for PCR, we only consider the result on the data after normalization.

We can see the result predicted by PCR in figure 4. And the SMSE is 0.008401.

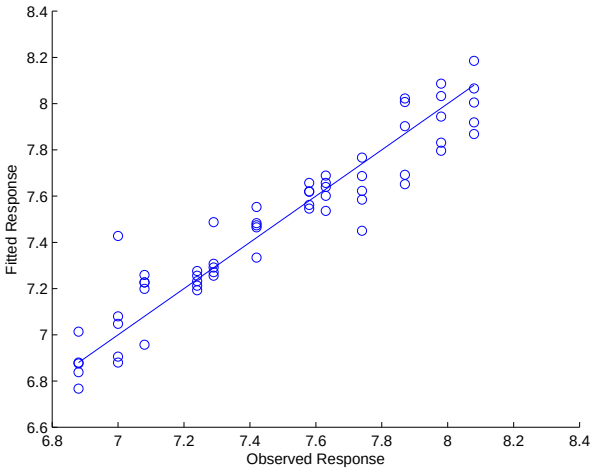

Figure 4: PCR with 10 principal components

The virtualization of linear regression parameter $\mathbf{v}$ is shown in figure 5.

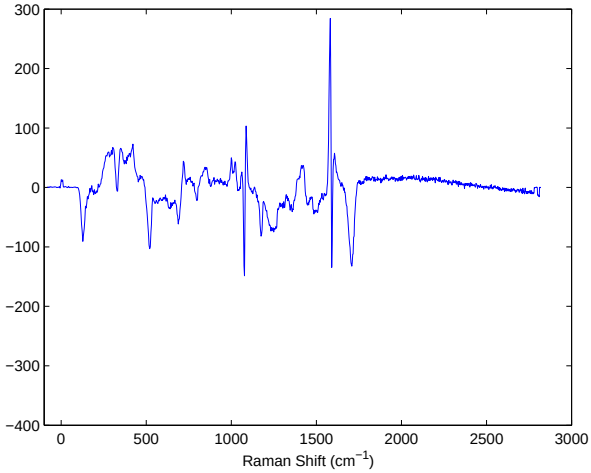

Figure 5: Plot of mean of $\mathbf{v}$ for PCR with 10 components

4.2 Partial Least Squares Regression

As we expected, this method has better performance than PCR, with SMSE 0.007519.

The result predicted by PLSR is shown in figure 6.

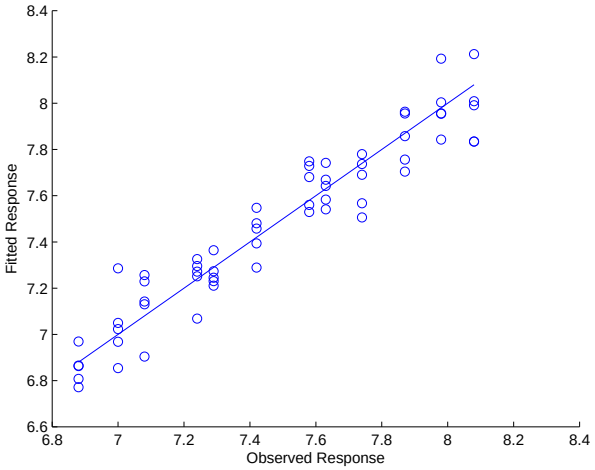

Figure 6: PLSR with 14 principal components

The virtualization of linear regression parameter $\mathbf{v}$ is shown in figure 7.

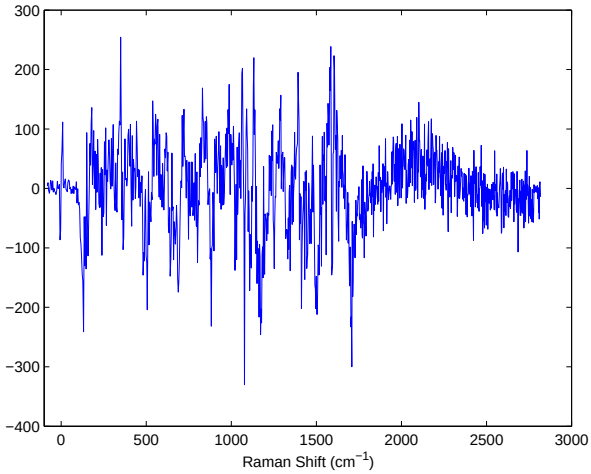

Figure 7: Plot of mean of $\mathbf{v}$ for PLSR with 14 components

As we can see, the curve is not that smooth like that of $\mathbf{v}$ for PCR and thus it's not easy for further analysis on it. So we try lasso regression in the next section to make the curve smoother.

4.3 Lasso Regression

5 Conclusions

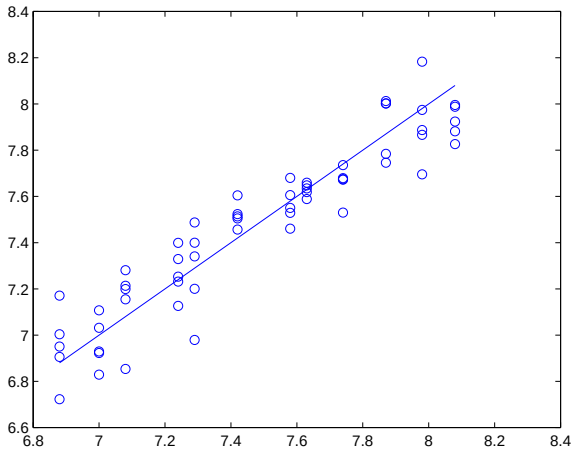

Figure 8: Plot of mean of $\mathbf{v}$ for PLSR with 14 components