The Institute for Integrating Statistics in Decision Sciences

Technical Report TR-2013-8
April 2, 2013

Analysis of SWORrD Data

Joshua Landon
Department of Statistics
The George Washington University, USA

Analysis of SWORrD Data

Joshua Landon

The George Washington University

November, 2012

Abstract

This report is on the analysis of the SWORrD (Swept Wavelength Optical Resonant Raman Detection) data. The goal of the analysis is to determine which chemicals are present in an unknown substance, by analyzing its Raman spectrum. In this report we present three approaches, all of which give excellent results, and can correctly determine the chemcials that make up the unknown substance.

Keywords: Raman Spectrum, Chemical Mixtures, Optimization, Gauss-Newton method, Bayesian Methods, Markov Chain Monte Carlo.

1 Introduction

The Raman spectrum is used to identify what chemical substances or biological agents are present in a substance. A sample of the substance is illuminated with a specific laser wavelength and this generates a resonance Raman spectrum, illustrated in Figure 1.

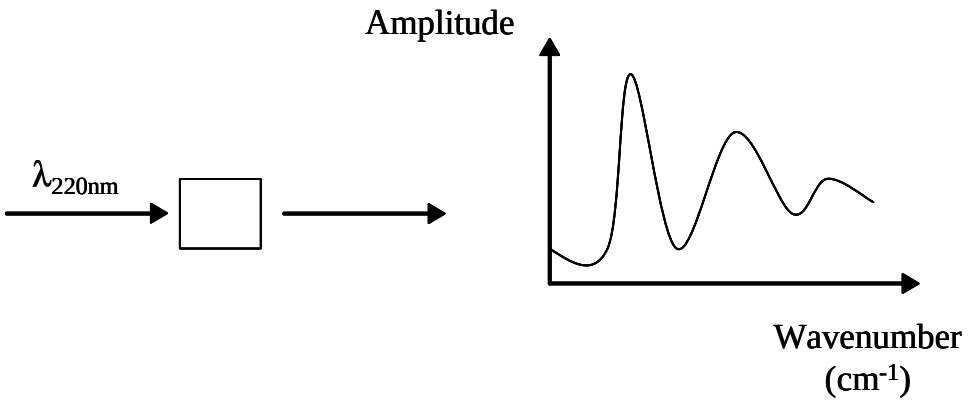

Figure 1: Raman Spectrum

The resulting spectrum constitutes a unique signature of the illuminated substance for this specific laser wavelength. The process is then repeated for different laser wavelengths, and the subsequent set of resonance Raman signatures, one at each laser wavelength, forms a single two-dimensional signature of the substance. One axis of the two-dimensional signature is the input laser wavelength and the other axis is the wavenumber of the Raman spectrum. Figure 2 shows an example Raman spectrum for a sample of Ammonium Nitrate.

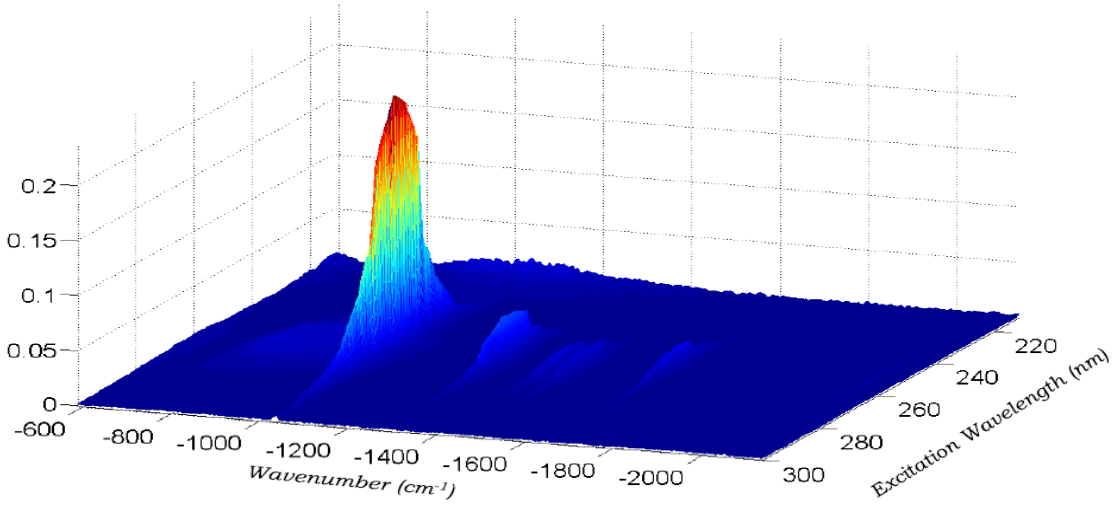

Figure 2: Example Spectrum of Ammonium Nitrate

The purpose of the Raman spectrum is to identify what chemicals or biological agents are present in the illuminated substance. The NRL has a library of chemical substances, say $c_1, \ldots, c_k$, where k is about 15-20. They can use this library to predict what is in a sample using a linear combination of the c_i's. So for example, sample s might be identified as

$$s \approx \alpha c_1 + (1-\alpha)c_2, \qquad 0 < \alpha < 1,$$

meaning that s consists of two chemical substances: c_1 and c_2, with proportions α and $(1-\alpha)$, respectively. In general, we write substance $s = \sum \alpha_i c_i$, with $\alpha_i \geq 0$, for all i, and $\sum \alpha_i = 1$; so if $\alpha_i > 0$, then c_i is present in the substance.

1.1 The Data

The NRL have provided us with data, consisting of the Raman spectrums of several combinations of chemical substances. There are five chemical substances in the data set: Water, Ethanol, Methanol, Acetonitrile, and Ethylene Glycole. We have several runs of the Raman spectrum for each of these five chemicals (each run of a particular chemical gives very similar results but, naturally, there is some difference). We also

have several runs of all of the possible combinations of these five chemicals, i.e. the 10 two-chemical combinations, the 10 three-chemical combinations, the 5 four-chemical combinations, and the mixture of all five chemicals. Our goal is to develop a method which can correctly predict what is in each of these substances. Note that all the combinations of chemicals in the data set consists of equal amounts of the chemicals in it, so for example, the mixture of water and ethanol contains 50% each of the two chemicals, and the mixture of all five chemicals consists of 20% of each. Unfortunately we do not have data on differet proportions for the mixtures, but nevertheless, we have come up with a method to predict the proportions of chemicals in each substance.

In this report we outline three methods: in Section 3 we will use a method in which we do not know what chemicals are in the substance, but we do know that it consists of equal parts of whichever chemicals are in it, and in Sections 4 and 5 we outline two methods for the more realistic scenario in which we have no idea what is in the substance, and nor do we know the proportions of the chemicals involved. The method of Section 4 involves taking a least squares approach, and the Section 5 details a Bayesian method. We will next discuss, in Section 2, the Raman spectrums of the five chemicals to help better understand the data.

2 The Raman Spectrums of the Five Chemicals

We will begin by simply showing the Raman spectrum for water at a wavelength of 220nm. Figure 3 shows the Raman spectrums for six different runs, indicated by the dotted lines, as well as the average of these six runs, indicated by the solid line. It is this average that we will be using as the Raman spectrum for water.

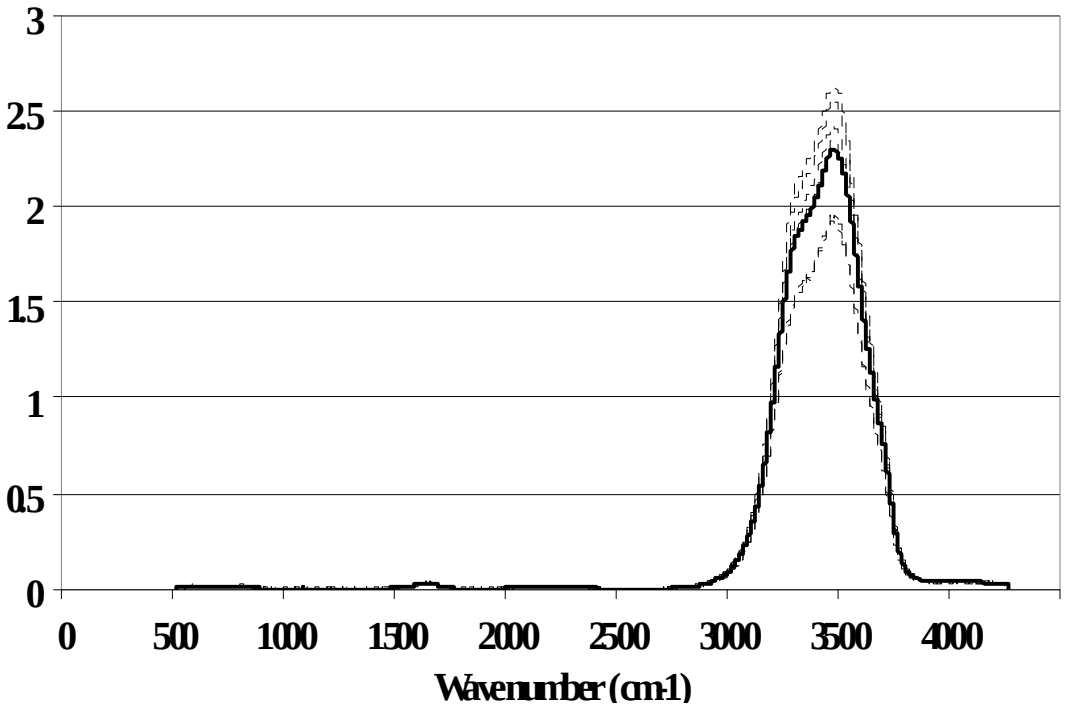

Figure 3: Raman spectrum for water at 220 nm

We also have the Raman spectrums for water taken at other wavelengths, in fact we have the spectrum for every even number wavelength from 220-260nm (21 different wavelengths in total). When we look at the spectrum for water at other wavelengths, it has the same shape as that shown in Figure 3 for 220nm, but the altitude is different, and in fact decreases as the wavenumber increases, as shown in Figure 4.

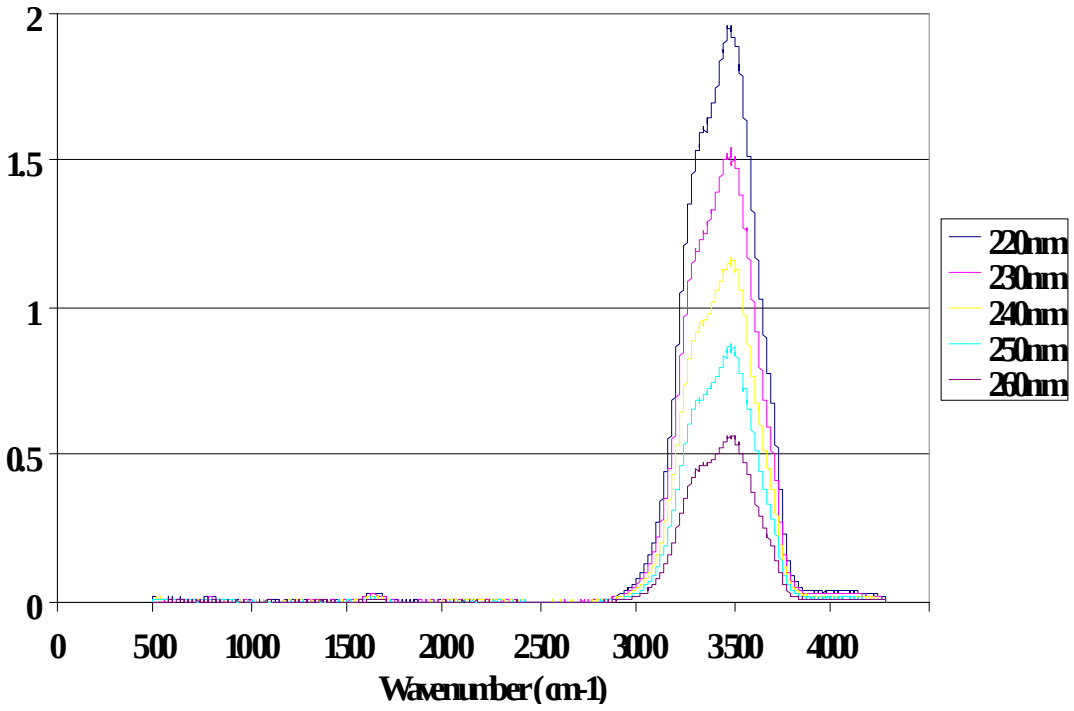

Figure 4: Raman spectrum of water for different wavelengths

As for the other four chemicals, they also have the same pattern over different wavelengths, and they all seem to have the greatest altitude at the 220nm wavelength. Thus, for the rest of this report, we will focus only on the Raman spectrums at 220nm. Now, other substances outside of the five that we are considering might have a very small Raman spectrum at 220nm compared to other wavelengths, in which case considering just the 220nm in not appropriate. In cases such as these we will show that we can simply extend our method to consider all wavelengths; the mathematics will be the same, but there'll just be more data.

Figure 5 shows the Raman spectrum for all five chemicals at the 220nm wavelengths. Note that these curves represent the average of all the runs for the respective chemical.

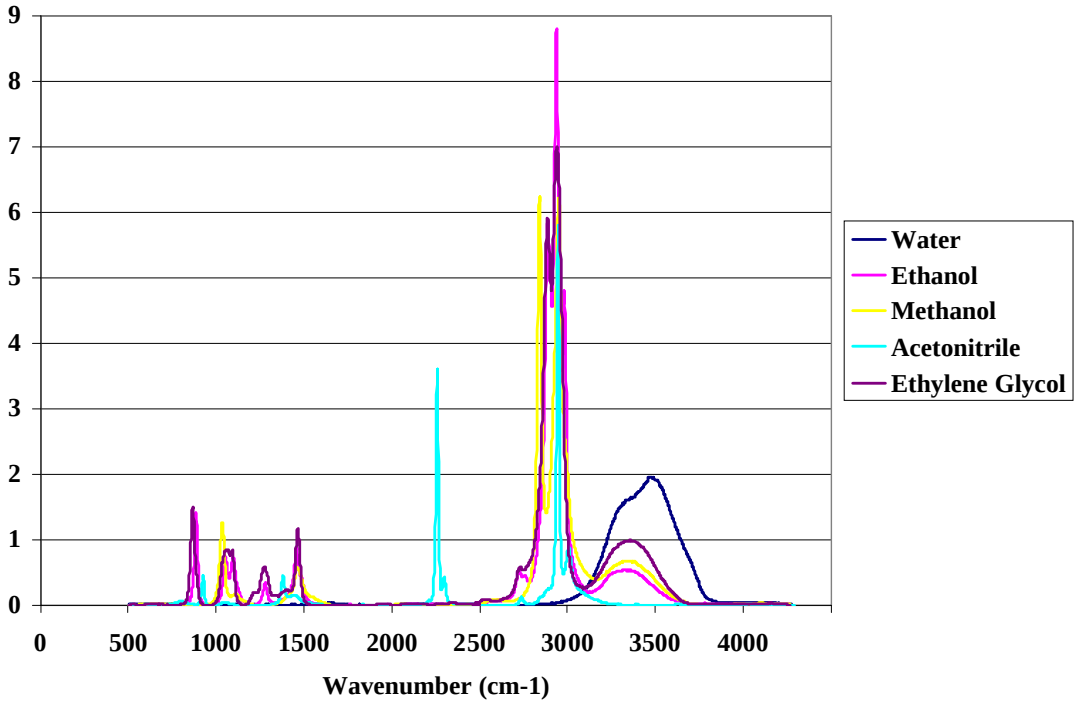

Figure 5: Raman spectrum of the five chemicals at the 220nm wavelength

In Figure 6, we show the Raman spectrum from the mixture of water and ethanol; note that this is the average of the runs for this mixture. This mixture is 50% water and 50% ethanol and, as we can see from Figure 6, its spectrum has peaks at the same place as water and ethanol, only now those peaks are halved. So a Raman spectrum of a mixture of chemicals is simply a linear combination of those Raman spectrum of the chemicals of which the mixture is composed. The challenge is to determine which of these five chemicals are in an unknown substance.

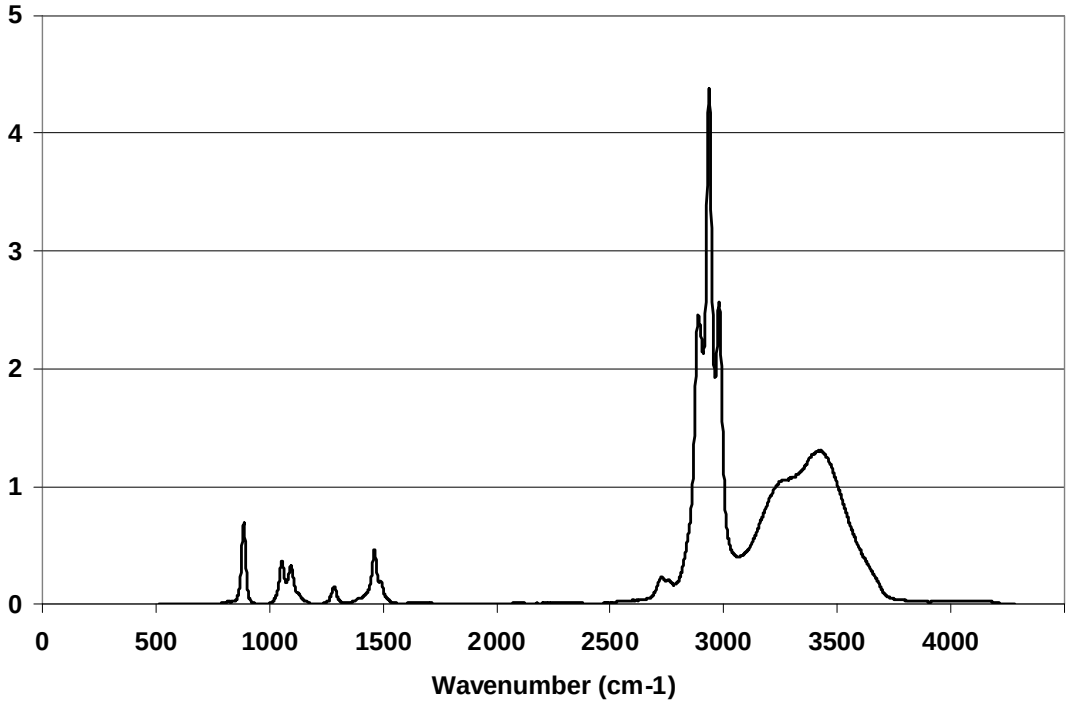

Figure 6: Raman spectrum of water and ethanol combined

In Section 3 we will outline a method for determining which chemicals are present, assuming that the chemicals present are of the same proportion (as is the case with our data). Then in Sections 4 and 5 we outline methods for when we do not assume that the unknown substance is divided up equally among its composition chemicals.

3 Method 1: Assuming Equal Proportions of Chemicals

In this section we assume that any unknown substance is made up of k different chemicals equally, so each of these chemicals account for $1/k$ of the substance. Our goal is to determine what these k chemicals are. We know that an unknown substance consists of one or more of the five chemicals, which we will denote: c_1 = Water, c_2 = Ethanol, c_3 = Methanol, c_4 = Acetonitrile, and c_5 = Ethylene Glycole. This means that there are only 31 possible combinations of the five chemicals: five when $k = 1$ or 4, ten when $k = 2$ or 3, and one when $k = 5$. Also, let z_{ij} be the amplitude of the Raman spectrum of chemical i at wavenumber j, and let z_j be the amplitude of the Raman spectrum of the unknown substance at wavenumber j.

The problem now is simply an optimization problem, in which we want to pick the "best" of the 31 combinations. To determine which combination is the best, i.e. the most likely to be the unknown compound, we will define this as the one whose linear combination of Raman spectrums has the smallest sum of squared errors from the actual Raman spectrum observed from the unknown substance. For example, the expected Raman spectrum of water and ethanol would be $(z_{1j} + z_{2j})/2$ for every wavenumber j, and so the sum of squared errors would be

$$\sum_{j=1}^{4500} \left[z_j - (z_{1j} + z_{2j})/2\right]^2 .$$

So we can simple calculate the sum of squared errors for each of the 31 possible combinations, and then choose the one with the smallest value.

As an example, let's suppose that we do not know which substance forms the Raman spectrum given in Figure 6. Calculating the 31 sets of sum of squared errors leads to the results given in Table 1.

Combination	SSE
c_1	1125.338
c_2	1290.228
c_3	1125.972
c_4	1145.372
c_5	1334.234
c_1 & c_2	37.5336
c_1 & c_3	327.1768
c_1 & c_4	689.6659
c_1 & c_5	118.2748
c_2 & c_3	864.4228
c_2 & c_4	508.7014
c_2 & c_5	1232.545
c_3 & c_4	682.3562
c_3 & c_5	913.9646
c_4 & c_5	465.9163
c_1 & c_2 & c_3	152.777
c_1 & c_2 & c_4	153.6294
c_1 & c_2 & c_5	200.4013
c_1 & c_3 & c_4	377.7902
c_1 & c_3 & c_5	205.7911
c_1 & c_4 & c_5	165.6093
c_2 & c_3 & c_4	517.8165
c_2 & c_3 & c_5	921.477
c_2 & c_4 & c_5	562.0905
c_3 & c_4 & c_5	515.93
c_1 & c_2 & c_3 & c_4	191.6002
c_1 & c_2 & c_3 & c_5	264.0076
c_1 & c_2 & c_4 & c_5	151.2625
c_1 & c_3 & c_4 & c_5	207.974
c_2 & c_3 & c_4 & c_5	555.0006
c_1 & c_2 & c_3 & c_4 & c_5	211.9517

Table 1: Sum of Squared Errors for the 31 possible combinations when estimating water and ethanol

As Table 1 shows, the smallest sum of squares is 37.53 which occurs when using the combination of water and ethanol, so we were able to correctly identify the substance. Similarly, Table 2 shows that the method works when considering the mixture of all five chemicals, with the smallest sum of squared errors (52.98) occuring for the last model. Indeed, this method works for every substance we were given, so we were always able to correctly identify its component chemicals.

Combination	SSE
c_1	1913.863
c_2	970.335
c_3	613.3352
c_4	743.9356
c_5	984.8309
c_1 & c_2	271.8492
c_1 & c_3	465.1204
c_1 & c_4	883.2098
c_1 & c_5	337.8357
c_2 & c_3	448.1575
c_2 & c_4	148.0364
c_2 & c_5	897.8964
c_3 & c_4	225.3192
c_3 & c_5	482.9446
c_4 & c_5	90.49652
c_1 & c_2 & c_3	138.1084
c_1 & c_2 & c_4	176.0276
c_1 & c_2 & c_5	240.144
c_1 & c_3 & c_4	335.9404
c_1 & c_3 & c_5	181.2859
c_1 & c_4 & c_5	178.1709
c_2 & c_3 & c_4	106.494
c_2 & c_3 & c_5	527.4992
c_2 & c_4 & c_5	205.1795
c_3 & c_4 & c_5	94.77103
c_1 & c_2 & c_3 & c_4	80.23951
c_1 & c_2 & c_3 & c_5	165.6554
c_1 & c_2 & c_4 & c_5	80.71044
c_1 & c_3 & c_4 & c_5	89.23594
c_2 & c_3 & c_4 & c_5	159.158
c_1 & c_2 & c_3 & c_4 & c_5	52.98253

Table 2: Sum of Squared Errors for the 31 possible combinations when estimating the combination of all five chemicals

4 Method 2: Unknown Proportion of Chemicals – Least Squares Approach

In this section we outline a method for the more realistic scenario in which we do not assume that the unknown substance consists of equal proportions of its component chemcals. Let the unknown substance consist of proportion α_i of chemical c_i, so $s = \sum \alpha_i c_i$; our goal is to estimate the α_i's. To do this, and using the same notation as the previous section, we want to minimize the sum of squared errors between the estimated Raman spectrum, given by $\sum \alpha_i z_{ij}$, and the actual Raman spectrum observed from our unknown substance. So we want to minimize:

$$\sum_{j=1}^{4500} \left[z_j - (\alpha_1 z_{1j} + \alpha_2 z_{2j} + \alpha_3 z_{3j} + \alpha_4 z_{4j} + \alpha_5 z_{5j})/2 \right]^2,$$

where the constraints are that $\sum \alpha_i = 1$ and $\alpha_i \geq 0$, for $i = 1, \ldots, 5$. So this boils down to a non-linear regression of the form

$$z_j = \alpha_1 z_{1j} + \alpha_2 z_{2j} + \alpha_3 z_{3j} + \alpha_4 z_{4j} + (1 - (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4)) z_{5j}.$$

This cannot be solved explicitly, so we will need to do this numerically. We shall use the Gauss-Newton numerical procedure [cf. Deuflhard (2005)], to calculate the least squares estimates of $\alpha_1, \ldots, \alpha_4$, and thus α_5. To use this method we specify initial values for the four parameters, $\alpha_1, \ldots, \alpha_4$, and then linearlizes z_j near these initial values. The new esimates of the parameters are then the linear least squares solution, using the linearized z_j. This process is then repeated until convergence. As an example, when trying to estimate components of the mixture of water and ethanol, with the Raman spectrum given in Figure 6, we obtain the estimates given in Table 3. Any initial values of the four parameters can be used in the interval $(0, 1)$ and the solution given in Table 3 will always be reached.

Chemical	Parameter Estimate
Water	0.4687
Ethanol	0.4539
Methanol	0.0367
Acetonitrile	0.0106
Ethylene Glycol	0.0299

Table 3: Estimates of the α_i's when predicting the components of a mixture of water and ethanol

These results are very close to the true values of $(0.5, 0.5, 0, 0, 0)$. The very small proportions for methanol, acetonitrile and ethylene glycol correctly indicate that it's unlikely that they are present in the substance. As another example, Table 4 shows the results from this method when analyzing the Raman spectrum of a mixture of all five chemicals.

Chemical	Parameter Estimate
Water	0.1287
Ethanol	0.1458
Methanol	0.1679
Acetonitrile	0.3365
Ethylene Glycol	0.2211

Table 4: Estimates of the α_i's when predicting the components of a mixture of all five chemicals

These results correctly indicate that all five chemicals are present in the substance. We are assuming that the actual proportions are all 0.2, but we do not know this for sure.

Indeed the results of Table 4 suggest that perhaps there is slightly more acetonitrile present in the substance that the other chemicals.

5 Method 3: Unknown Proportion of Chemicals – Bayesian Approach

In this section we outline our Bayesian approach. Using the same notation as the previous sections, we will use the following model:

$$z_j = \alpha_1 z_{1j} + \alpha_2 z_{2j} + \alpha_3 z_{3j} + \alpha_4 z_{4j} + (1 - (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4))z_{5j} + \epsilon_j$$

where $\epsilon_j \sim N(0, \sigma^2)$. Define $f(\boldsymbol{\alpha}, \mathbf{z}) = \alpha_1 z_{1j} + \alpha_2 z_{2j} + \alpha_3 z_{3j} + \alpha_4 z_{4j} + (1 - (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4))z_{5j}$, so now we have

$$z_j \sim N(f(\boldsymbol{\alpha}, \mathbf{z}), \sigma^2).$$

In Section 4 we calculated the least squares estimates of the α_i's, so now for our Bayesian approach we first need to write our likelihood, which is:

$$L \propto \frac{1}{\sigma^n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{4500}\left(z_j - f(\boldsymbol{\alpha}, \mathbf{z})\right)^2\right].$$

We have 5 parameters α_1, $\alpha_2, \alpha_3, \alpha_4$ and σ^2, and we now need to specify prior distributions for them. For the α_i's the conjugate priors would be normal, but normal priors would be inappropriate for these parameters as they are all in the interval $(0, 1)$, so instead we shall use truncated beta priors, with point masses at 0 and 1. So let the prior distribution of α_i be

$$\pi(\alpha_i) \propto \begin{cases} m_{i0}, & \alpha_i = 0 \\ \alpha_i^{a_i - 1}(1 - \alpha_i)^{b_i - 1}, & 0 < \alpha_i < 1 \\ m_{i1}, & \alpha_i = 1 \end{cases}.$$

For the prior distribution of σ^2 we will use the conjugate prior, which is an inverse gamma distribution, which gives (with parameters a_s and b_s):

$$\pi(\sigma^2) \propto \frac{1}{(\sigma^2)^{1+a_s}} \exp\left[-\frac{b_s}{\sigma^2}\right].$$

Now we can write out the full posterior conditional distributions:

$$\pi(\alpha_i|\alpha_{j\neq i},\sigma) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{4500}(z_j - f(\boldsymbol{\alpha},\mathbf{z}))^2\right] a_i^{a_i-1}(1-\alpha_i)^{b_i-1}$$

and

$$\pi(\sigma|\alpha_i) \propto IG\left(a_s + n, b_s + \frac{1}{2}\sum_{i=1}^{4500}(z_j - f(\boldsymbol{\alpha},\mathbf{z}))^2\right).$$

We can then draw a posterior sample from the joint distribution of α_1, $\alpha_2, \alpha_3, \alpha_4$ and σ^2 by iteratively drawing from the given full conditional posterior distributions.

As an example, we will once again use the Raman spectrum of water and ethanol, shown in Figure 6. Figures 7-10 show the posterior distributions of α_1, α_2, α_3,and α_4, after 1,000 simulations. The results are very good as it's clear from these figures that the unknown substance contains just water and ethanol, with about 50% of each.

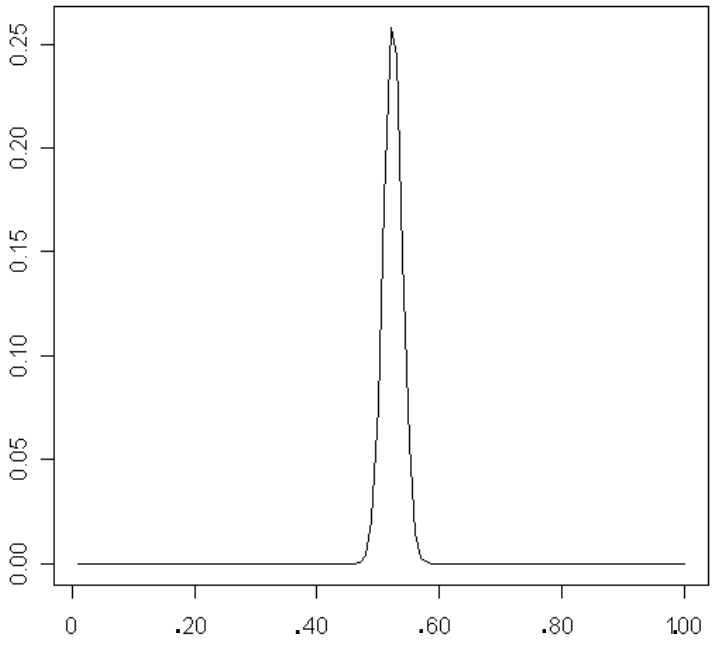

Figure 7: Posterior distribution of β_1

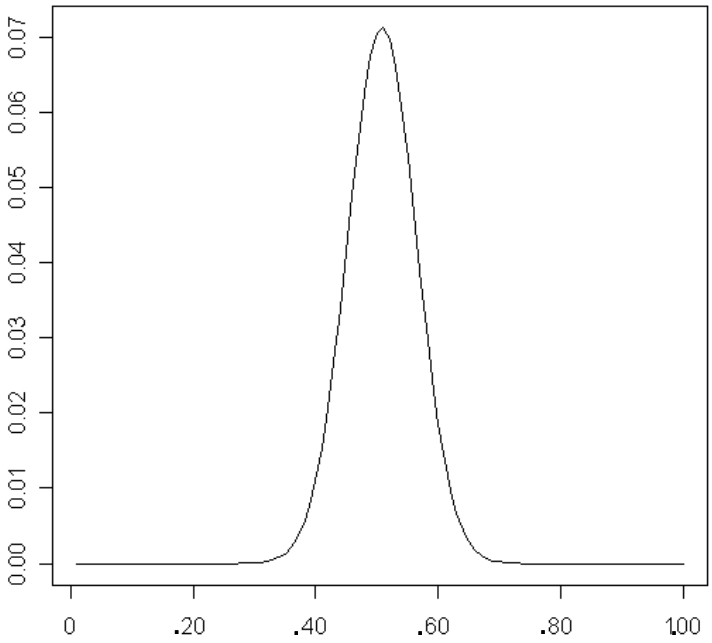

Figure 8: Posterior distribution of β_2

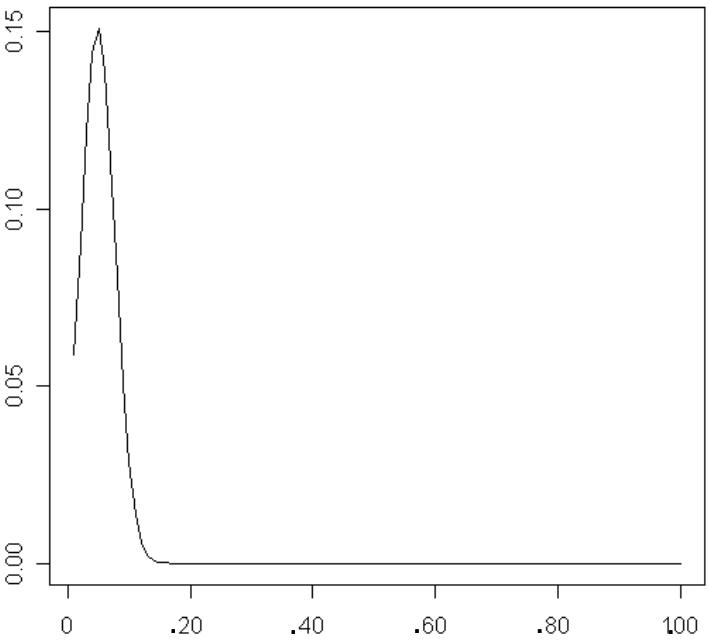

Figure 9: Posterior distribution of β_3

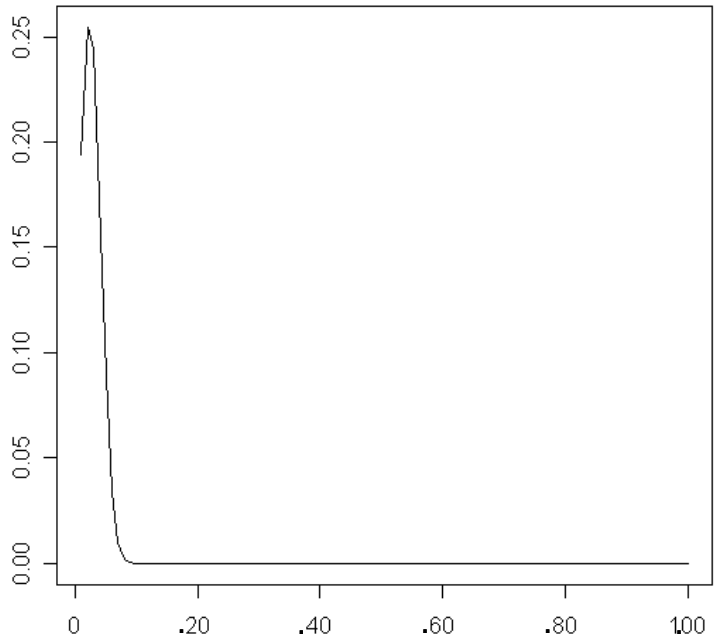

Figure 10: Posterior distribution of β_4

Figures 11-14 show the posterior distributions for α_1, α_2, α_3,and α_4 when the unknown substance is the mixture of all five chemicals. The results are similat to those obtained in Section 4.

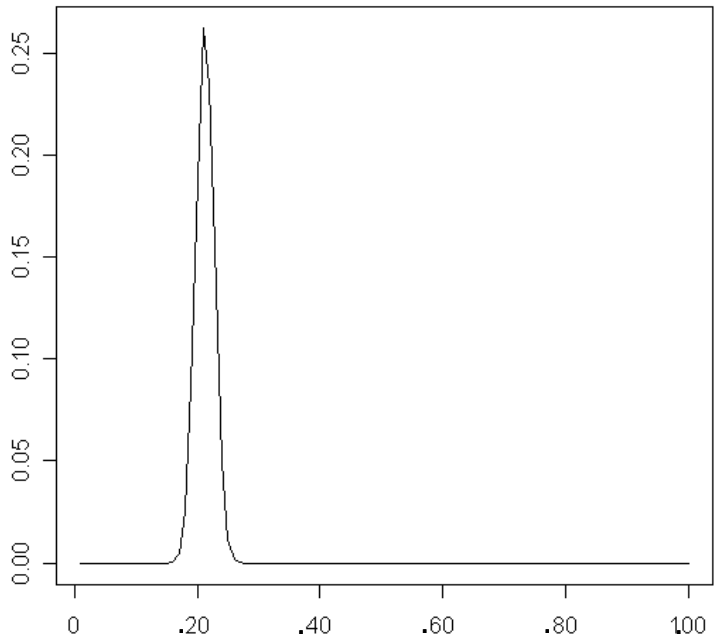

Figure 11: Posterior distribution of β_1

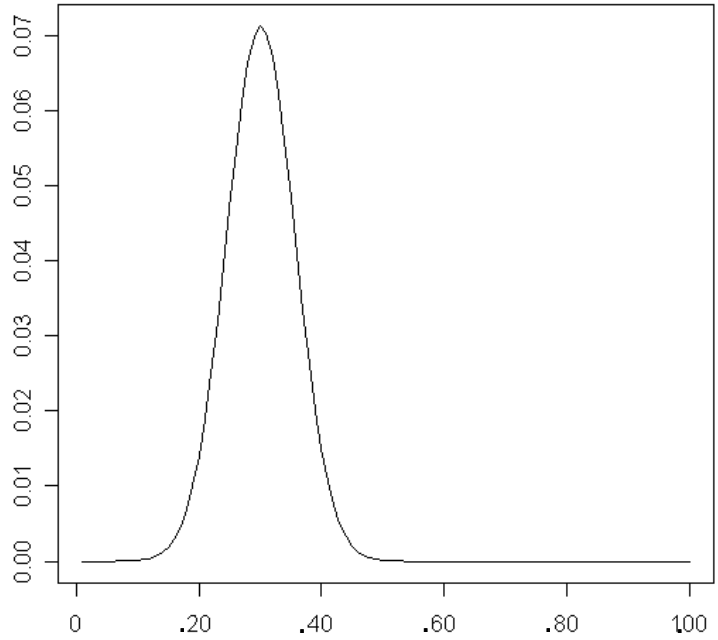

Figure 12: Posterior distribution of β_2

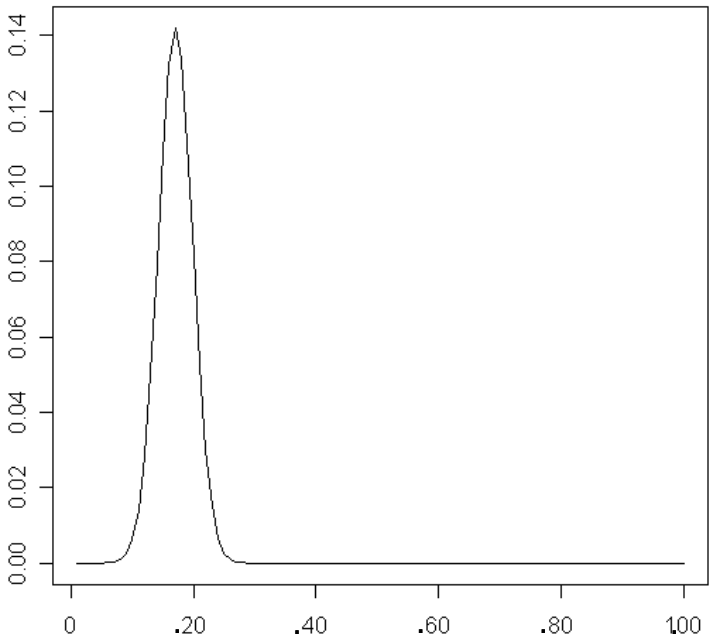

Figure 13: Posterior distribution of β_3

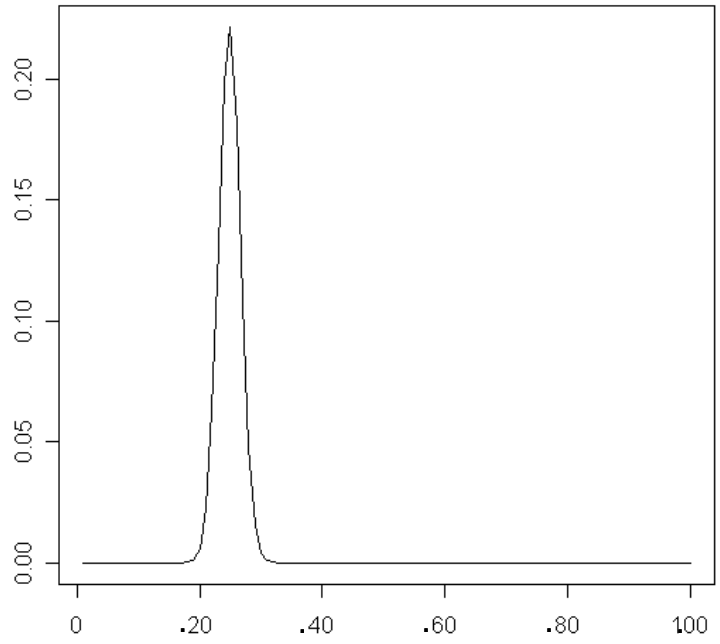

Figure 14: Posterior distribution of β_4

6 Conclusions and Future Work

All three methods given in this report proved to be very accurate at predicting the chemicals that make up an unknown substance. Of course, the methods of Sections 4 and 5 should be used when analyzing a real life unknown substance as we cannot realistically assume equal proportions of chemicals in the substance. This report only focused on the 220nm wavelength of the Raman spectrum, but the method can of course be easily extended to minimizing the sum of squared errors over all wavelengths.

Whilst our method gave very good results when predicting the chemicals present in an uknown substance, the real test will come when applied to an unknown substance made up of biological agents. Biological agents tend to have very similar Raman spectrums that have peaks at the same wavenumbers, making them harder to distinguish. We feel confident, however, that the methods outlined in this report can offer decent predictions for these types of substances, and we hope to test this when we are given this data.

Acknowlegements

The author would like to thank Jacob Grun of the Naval Research Laboratory for providing us with the data, and for helping us to understand it. The research was supported by the National Science Foundation Grant DMS-0915156, with the George Washington University.

References

[1] Deuflhard, P. (2011). *Newton Methods for Nonlinear Problems.* Springer-Verlag, New York.