A Sparse Regularized Model for Raman Spectral Analysis

Di Wu[1], Mehrdad Yaghoobi[1], Shaun Kelly[1], Mike Davies[1] and Rhea Clewes[2]

[1]The University of Edinburgh
[2]DSTL

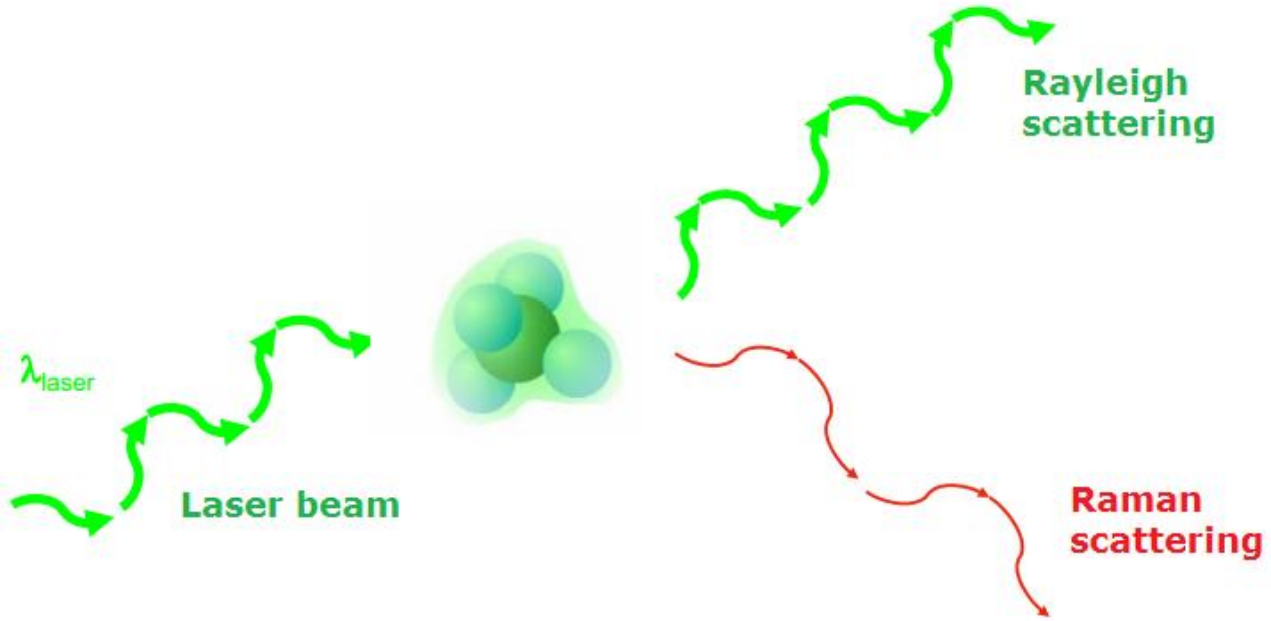
Outline

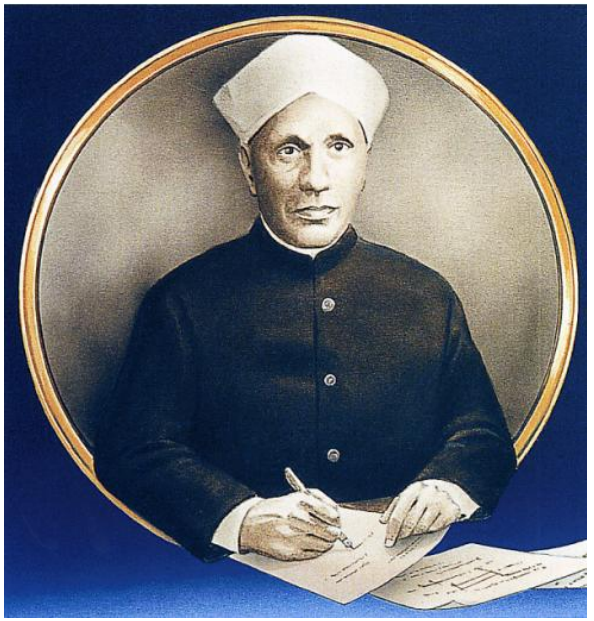
- Introduction
- Objectives
- Methodology
- Experiment and Results
- Conclusions

Introduction

- Raman Spectroscopy

Raman Spectroscopy is a vibrational spectroscopy technique which collects light radiation scattered from an illuminated sample.
A Raman spectrum provides the unique fingerprint to identify molecules.



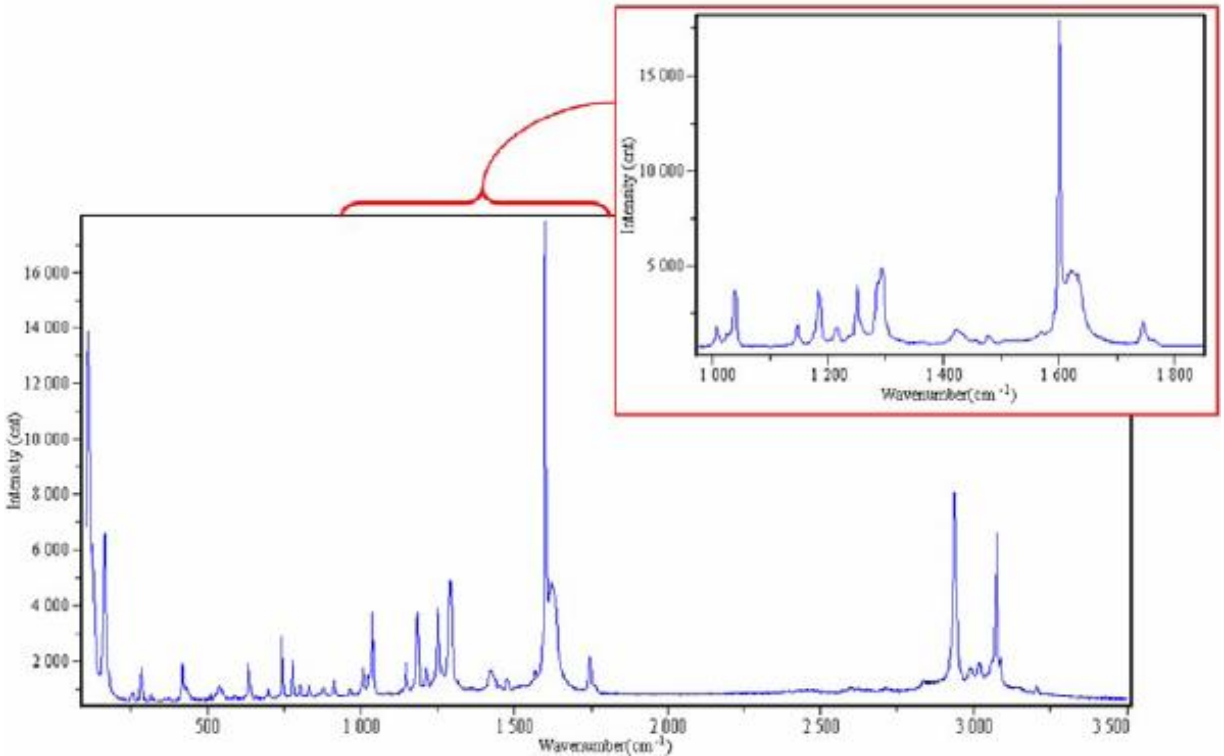


Sir C. V. Raman
Nobel Prize 1930 Physics

Introduction

- Raman Spectroscopy

1. Each Raman spectrum features a number of peaks which correspond to specific molecular bond vibrations.
2. Raman is both qualitative and quantitative.
3. Quantitative Raman methods are frequently based on either peak area or peak height.

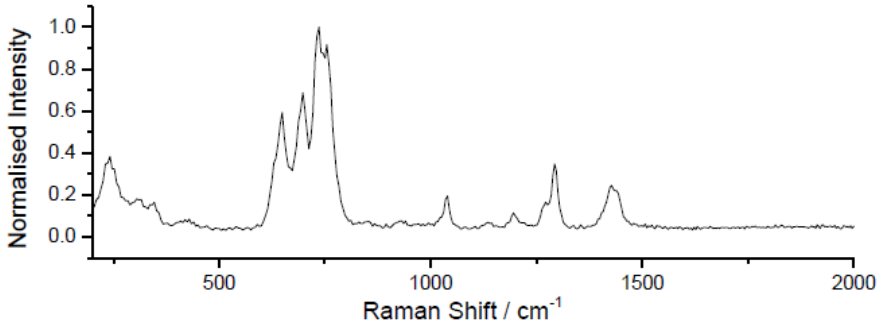


A typical Raman spectrum of aspirin

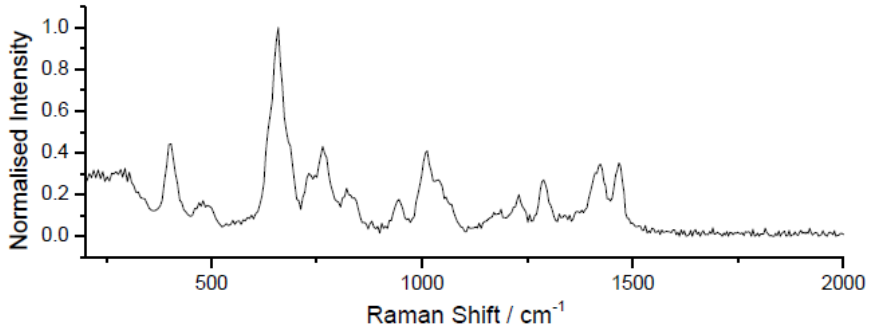
Introduction

- Raman Spectral Deconvolution

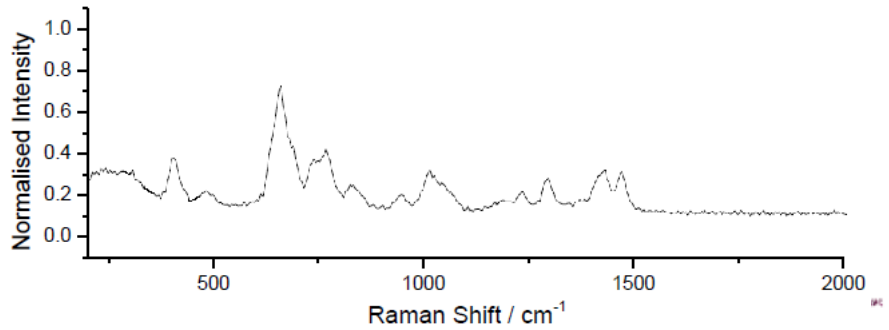
A



B



A&B Mixture



1. The mixture spectrum and a spectral library are given.

2. The goal is to qualitatively identify the mixture components and quantitatively predict the concentrations.

Introduction

- Raman Spectral Deconvolution

Existing quantitative techniques:

1. Direct Classical Least Squares (DCLS)

   Finds the linear combination of spectra that most closely matches the mixture spectrum.

2. Indirect Classical Least Squares

   First estimate the Raman spectra of the mixture components and then applies DCLS.

2. Principal Component Regression (training set required)

   First reduce the number of spectrum variables by using principal components analysis (PCA) and then estimate the analyte concentration as a function of the variables.

Objectives/Motivations

Only very few chemicals in the library are the true components in the mixture!

By exploiting this sparsity, we intended to realise

- Qualitative identification
- Quantitative analysis
- The extraction of possible unknown substances

* The sparse regularized model behaves as the complement and extension to other Raman regression models (e.g. least squares regression).
* No training required

Methodology

- Chemical composition model

By assuming the pure component spectra are not changed when the pure components are mixed together (linear superposition model), we have

$$y = \Theta(X, C) + e$$
$$= XC + e$$

where y is the Raman spectrum of a chemical mixture, Θ is the operator which represents the interaction between all components, C is the relative concentrations vector for the chemical components, $X=(x_1,x_2, \ldots x_m)$ with x_i represents the spectrum of the i-th mixture component, and e is the residual spectrum.

Methodology

- Direct Classical Least Squares (DCLS)

$$\min \left\| y - \Theta(D, \hat{C}) \right\|_2^2$$

The sample components are identified as the support of $\hat{C}$.

Methodology

- Sparse regularized model

By introducing the components sparsity and the non-negativity of the concentrations, we have

$$\min \quad \left\| y - \Theta(D, \hat{C}) \right\|_2^2 + \lambda \left\| \hat{C} \right\|_p$$

$$s.t. \quad \hat{c}_i \geq 0$$

where λ is the coefficient to control the trade-off between fitting the data and variable sparsity; p is often between 0 and 2 to give an appropriate solution. The smaller the p, the better L_p norm measures the sparsity.

Methodology

- Sparse regularized model

For p=0, the equivalent model:

$$\min \quad \left\| y - \Theta(D, \hat{C}) \right\|_2^2$$

$$s.t. \quad \left\| \hat{C} \right\|_0 \leq z$$

$$\hat{c}_i \geq 0$$

where the coefficient z bounds the number of components.

* It has been proven that solving the L_p ($0 \leq p < 1$) minimization problem is strongly NP-hard.
Solution1: exhaustively search the feasible space with a termination condition.

Methodology

- Sparse regularized model

Solution2 : commonly used compromise (L1-norm instead of L0) :

$$\min \quad \left\| y - \Theta(D, \hat{C}) \right\|_2^2 + \lambda \left\| \hat{C} \right\|_1$$

$$s.t. \qquad \hat{C}_i \geq 0$$

The choice of λ is justified by finding the sparsest possible Ĉ within acceptable level of residual spectrum intensity.

Methodology

- Unknown Substances

The combined unknown substances are likely to contribute a non-negative spectrum that is independent of the library.

$$y = \Theta(X, C) + e$$

$$\rightarrow \quad y = f + \Theta(X, C) + e$$

Where the spectrum contribution from unknown substances is denoted as f.

* The residual spectrum e and unknown spectrum f can be used for diagnostic purposes.

Methodology

- Unknown Substances

By assuming that the spectrum f mainly lies in the difference between y and Θ(X,C), we introduce:

$$\min \quad \|f\|_p + \lambda \left\| f + \Theta(D,\hat{C}) - y \right\|_2$$

$$s.t. \quad f \geq 0$$

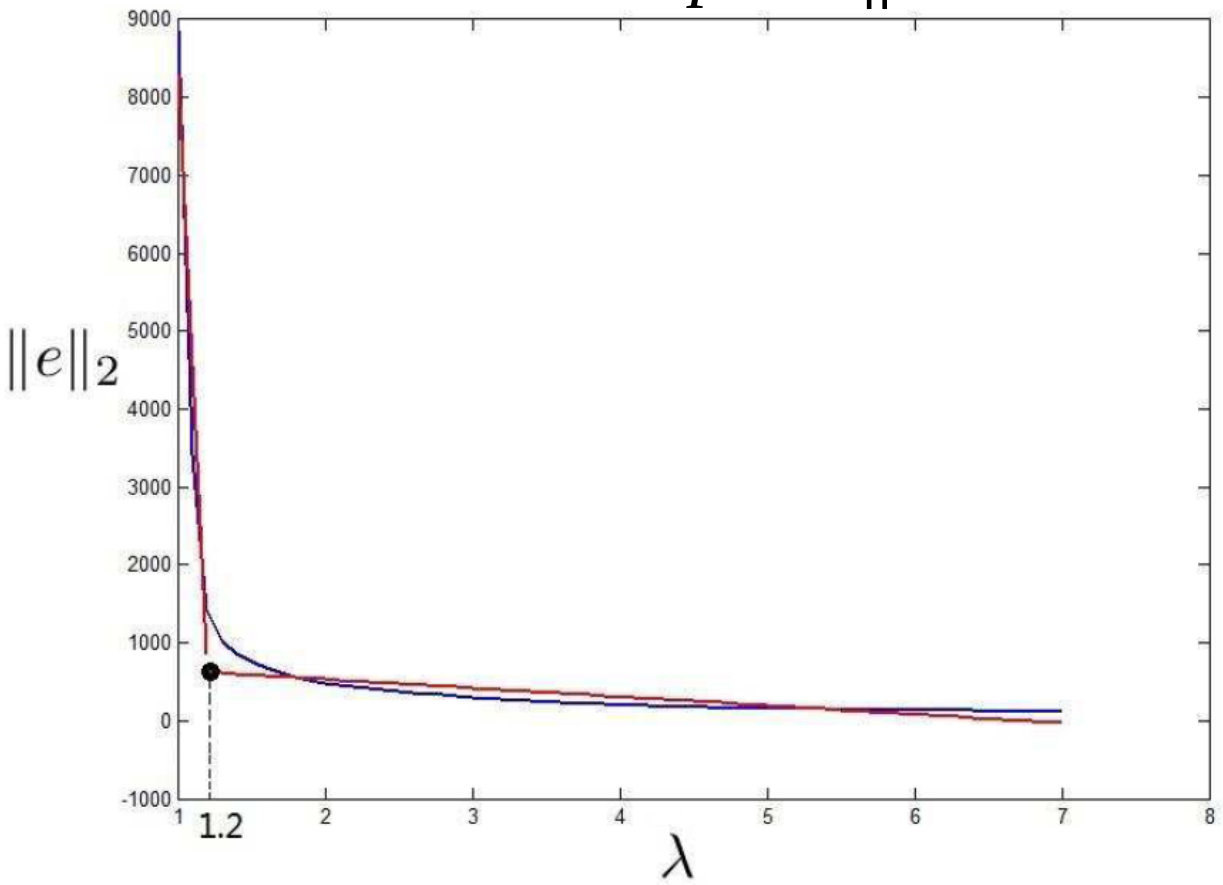
$$\hat{C}_i \geq 0$$

where 1≤ p ≤2 and λ is a positive tuning parameter.

Methodology

- Unknown Substances

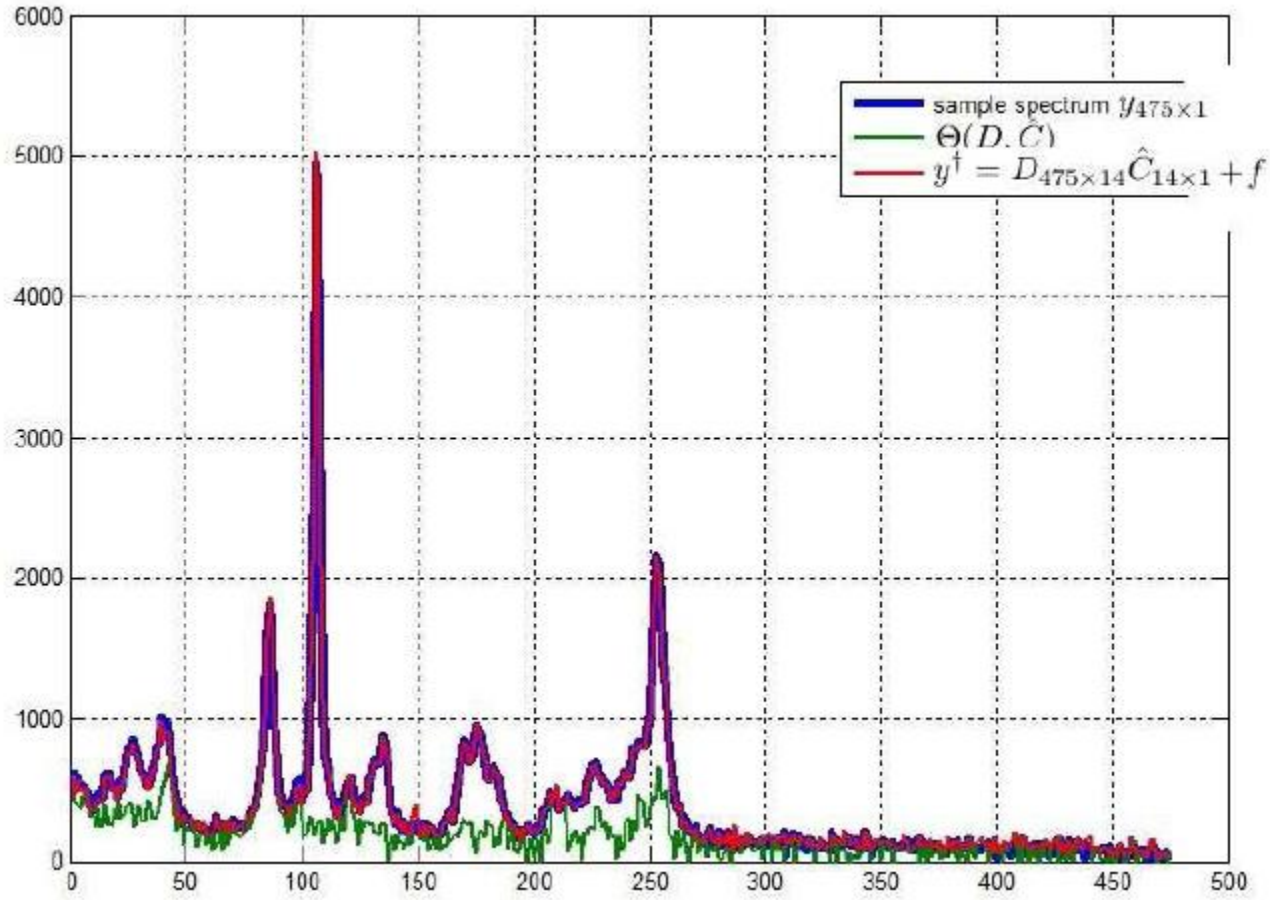
$$\min \quad \|f\|_p + \lambda \left\| f + \Theta(D, \hat{C}) - y \right\|_2$$



In practice we choose the "knee point".

Experiment and Results

- Real data set



A typical mixture with unknown substances

Blue line: original mixture
Green line: combined spectrum from the dictionary elements
Red line: the reconstructed spectrum

Qualitative: The estimated components and unknown signal f in this mixture match the ground truth data.

Experiment and Results

- Real data set

QUANTITATIVE RESULTS

	ground truth		estimation	
components	1st	2nd	1st	2nd
1st mixture	10%	90%	25%	75%
2nd mixture	50%	50%	58%	42%
3rd mixture	25%	75%	43%	57%

Experiment and Results

- Simulation

We compare the performance of the **sparse regularized regression models** and the **Direct Classical Least Squares method**.

Experimental Setup:

- All substances are known in this experiment
- Randomly mix 2 to 5 candidates from the 14 library spectra with a typical residual spectrum retrieved from real data experiments.
- Randomly generate the concentrations (remove the ones which contribute less than 10 percent to the overall spectrum intensity)

Experiment and Results

- Simulation

We borrow the definition of confusion matrix from the field of machine learning.

True positive (tp)	False positive (fp)	Precision =tp/(tp+fp)
False negative (fn)	True negative (tn)	Negative Predictive value (NPV)=tn/(fn+pn)
sensitivity =tp/(tp+fn)	specificity =fn/(tp+fn)	Accuracy (Acc)= (tp+tn)/(tp+fp+tn+fn)

Let tp be the number of mixture components which have been correctly identified, fp be the number of incorrectly identified components, tn be the number of correctly rejected library spectra, and fn be the number of incorrectly rejected library spectra.

Experiment and Results

- Simulation

	DCLS	sparse regularized non-negative least squares optimization Solution 2 : L1-norm instead of L0	sparse regularized non-negative least squares optimization Solution 1 : exhaustive search with a termination condition
sensitivity	1	1	0.9505
specificity	0.7669	0.9605	0.9992
precision	0.4918	0.8857	0.9958
NPV	1	1	0.9882
Acc	0.8107	0.9686	0.9893
F1 score	0.6514	0.9316	0.9673

The F1 score is harmonic mean of precision and sensitivity and it shows the overall identification performance.

Conclusions

- A sparse regularized model for Raman spectroscopy has been established to qualitatively identify mixture components and quantitatively predict the concentrations.

- The proposed model works as a complement to traditional regression methods.

- The model can extract the combined spectrum of unknown substances.

- Improved performance compared with Direct Classical Least Squares.

Thanks for your attention!

Questions?