A Sparse Regularized Model for Raman Spectral Analysis

Di Wu, Mehrdad Yaghoobi, Shaun Kelly and Mike Davies
School of Engineering
University of Edinburgh
UK, EH9 3JL
Email: {D.Wu, m.yaghoobi-vaighan, shaun.kelly, mike.davies}@ed.ac.uk

Rhea Clewes
Detection Department
DSTL
UK, SP4 0JQ
Email: RJCLEWES@mail.dstl.gov.uk

Abstract—**Raman spectroscopy has for a long time performed as a common analytical technique in spectroscopic applications. A Raman spectrum depends upon how efficiently a molecule scatters the incident light (electron rich molecules often produce strong signals) which results in difficulties for relating the spectrum to the absolute amounts of present substances. The spectrum is however a stable and accurate representation of the sample measured especially considering that each molecule is associated with a unique spectrum. State-of-the-art spectroscopic calibration methods include the principal component regression (PCR) and partial least squares regression (PLSR) methods which have been proved to be efficient regression methods to realise the quantitative analysis of Raman spectrum. In this paper we consider the problem of Raman spectra deconvolution to analyse the sample composition, as well as possible unknown substances. In particular, we propose a sparse regularized model as a complement to traditional regression methods by leveraging the components sparsity compared to the whole chemical library and the spectra sparsity, given that the chemical fingerprint of each spectrum is mainly determined by the peaks. Experimental results illustrate the effectiveness of this sparse regularized model.**

I. Introduction

Raman spectroscopy is a vibrational spectroscopy based technique which collects light radiation scattered from an illuminated sample. The scattered photons are shifted in frequency due to the interaction between the incident excitation and molecular vibrations which provides valuable identification information. Thus a single Raman spectrum often contains enough information about the sample composition to provide a unique fingerprint to distinguish the sample from others. Raman spectroscopy is commonly used in the analysis of materials in a variety of fields, such as chemistry, semiconductors, geosciences, medicine and biosciences. For example, the Raman spectrum has been used in the identification of artificial diamonds, the non-destructive forensic analysis of substances, and it is sufficiently characteristic to discriminate carcinoma from healthy tissue cells.

Due to the plethora of molecular information which can be obtained from Raman spectra, this technique is of value for the identification of materials found in civilian and military environments. As Raman is a non-contact technique for analysing materials, it provides a series of benefits when interrogating hazardous chemicals. Illicit, potentially dangerous, samples are unlikely to be found as well presented pure materials therefore it is important to be able to extract components from the composite Raman spectra.

Although Raman spectroscopy has received extensive use in qualitative analysis, quantitative analysis has not kept pace with those applications. The development of quantitative Raman spectroscopy is difficult as the intensity of a Raman spectrum is dependent upon the efficiency of the target chemical at scattering photons which cannot be directly related to the number of bonds. With the latest improvement in instrumentation and algorithms, more and more applications are moving towards quantitative analysis methods in Raman spectroscopy. Existing literatures [1][2][3] have reported that direct classical least squares (DCLS) method can give a good estimation under the assumption that pure component spectra are not changed when the different materials are mixed together. Other commonly used methods include partial least squares (PLS) and principal component regression (PCR) which are latent variable based methods that combine regression problems with dimension reduction techniques [3]. Some recent works in quantitative Raman analysis were focusing on rolling out new regression methods [4]. Also in 2010, Lyandres et al. proposed to adapt an optimization model to estimate a prediction range for the minimum and maximum concentrations for a given sample [5].

In this paper we present a model to introduce the exploitation of sparsities in Raman spectrum deconvolution. With this novel model, we could simultaneously realise both qualitative and quantitative composition analysis, and the identification of possible unknown substances. This sparse regularized model works as a complement to traditional Raman regression models and can be further developed to combine with other state-of-the-art techniques. Furthermore, we associate this model with the non-negative Direct Classical Least Squares (DCLS) regression method in experiments to showcase its effectiveness.

The remainder of this paper is organized as follows. In section two, based on traditional regression methods we introduce the sparse regularized models to better identify chemical mixture components and estimate their concentrations. In addition, a novel model is proposed to extract the spectra of unknown substances from the mixture spectrum. In section three, we conduct Raman spectra deconvolution experiments to illustrate how the model works. Conclusions and future work are presented in section four.

II. Sparse Regularized Raman Spectroscopy

Let $y \in \mathbb{R}^N$ be the Raman spectrum of a chemical mixture; $X = (x_1, x_2, ..., x_m)$, with $x_i \in \mathbb{R}^N$ represents the spectrum of the $i - th$ mixture component; $C = (c_1, c_2, ..., c_m)$ is the relative concentrations vector for the chemical components, with $c_i \geq 0$. The model of the mixture y can then be generalized as the combination of X with an extra term e:

$$y = \Theta(X, C) + e \tag{1}$$

where Θ is the operator which represents the interaction between all components, and $e \in \mathbb{R}^N$ is the residual spectrum. In practice, Θ could be general linear superposition, or non-linear combination due to chemical interactions.

Taking the linear superposition model as an example, the operator Θ performs as the weighted linear combination of all elements in X, i.e. $\Theta(X, C) = \sum_{i=1}^{m} x_i * c_i$, then we have the mixing process:

$$y_{N\times1} = X_{N\times m}C_{m\times1} + e_{N\times1} \tag{2}$$

The vector e represents the difference between the linear combination of the spectra and the observed spectrum y, due to any accuracies in the model Θ or the presence of any foreign substances. For all the experiments in the rest of the paper, we simply employ this linear superposition model.

A. Sparse Regularized Chemical Composition Analysis

We now consider the problem of identifying the chemical composition from the mixture spectrum $y \in \mathbb{R}^N$ and quantitatively estimating the concentrations of all components. Given that y follows the general mixing process (1), we assume that all the components x_i are chosen from a Raman Spectra library $D = (d_1, d_2, ..., d_{N_D}) \in \mathbb{R}^{N\times N_D}$. Also, we assume in the remainder of the paper that all spectra in the library are well preprocessed without much loss of quantitative information [6][7]. Then by instantiating the combination model Θ, the mixture components can be estimated by minimizing e^Te subject to appropriate constraints.

The general model to solve the simultaneous qualitative identification and quantitative analysis problem is:

$$\begin{aligned} &\underset{\hat{C}}{\operatorname{argmin}} \|y - \Theta(D, \hat{C})\|_2^2 \\ &= \{\hat{C} \mid \forall \hat{c}_i \in \hat{C}_{N_D\times1} : \ \hat{c}_i \geq 0\} \end{aligned} \tag{3}$$

where $\hat{c}_i$ is the absolute concentration of the i-th chemical in D. The sample components are identified as $\operatorname{supp}(\hat{C})$ which is the support of $\hat{C}$. Thus the resultant X in (1) is a subset of D which corresponds to the spectra with non-zero concentrations, and the relative concentrations can be computed by normalizing the corresponding elements in $\hat{C}$.

Given the fact that $\hat{C}$ is likely to be highly sparse, i.e. only very few chemicals in the library are the true components in the mixture, we propose to extend the general model (3) by imposing the sparse regularization:

$$\begin{aligned} &\min_{\hat{C}} \|y - \Theta(D, \hat{C})\|_2 + \lambda\|\hat{C}\|_p \\ &s.t. \quad \hat{c}_i \geq 0 \quad for\ i = 1, 2, ..., N_D \end{aligned} \tag{4}$$

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. With the L_p regularized objective function in (4), the sparsity of $\hat{C}$ is formulated as a penalty term, and the sparse regularized model is thus better able to distinguish chemical components. We thus have the model (4) which coincides with the non-negative Lasso (least absolute shrinkage and selection operator) model in compressed sensing techniques [8][9]. In particular, for $p = 0$ in (4), we have an equivalent model:

$$\begin{aligned} &\min_{\hat{C}} \|y - \Theta(D, \hat{C})\|_2 \\ &s.t. \quad \|\hat{C}\|_0 \leq z \\ &\hat{c}_i \geq 0 \quad for\ i = 1, 2, ..., N_D \end{aligned} \tag{5}$$

which finds the combination of elements in D that best approximates y with a limited number of components. The coefficient z bounds the number of components and it has a similar effect to the λ in (4).

It has been proven that solving the L_p $(0 \leq p < 1)$ minimization problem is strongly NP-hard [10]. However, if N_D is small enough, the model (5) can be implemented directly with an exhaustive search for all possible combinations in the feasible space. Otherwise, the commonly used compromise is to solve:

$$\begin{aligned} &\min_{\hat{C}} \|y - \Theta(D, \hat{C})\|_2 + \lambda\|\hat{C}\|_1 \\ &s.t. \quad \hat{c}_i \geq 0 \quad for\ i = 1, 2, ..., N_D \end{aligned} \tag{6}$$

With appropriate Θ, the model (6) can often be transformed to a convex programming problem which can be solved by conventional algorithms, such as interior-point methods, and a number of state-of-the-art approaches. The choice of λ is justified by finding the sparsest possible $\hat{C}$ within acceptable level of residual spectrum intensity. The model (6) is especially suitable when we are trying to resolve the Raman signals in a large database.

To summarise, the sparse regularized model stands as a natural complement to the Raman regression models. For a small Raman library, model (5) can accurately retrieve the composition information, while for a large Raman library, model (6) can be the substitute to rapidly obtain solutions with certain accuracy. Compared with simple regression models, both qualitative and quantitative Raman applications could benefit from extra sparse constraints given the appropriate interaction model Θ.

B. Sparse Regularized Extraction of Unknown Substances

Another problem worth considering is that there may exist unknown substances in the chemical mixture which could

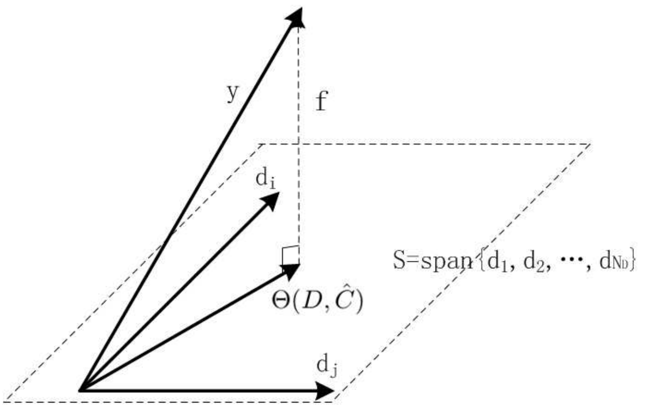

Fig. 1. Illustration of the approximation of f when $\Theta(D, \bullet)$ stands for the vector space $span\{d_1, d_2, ..., d_{N_D}\}$.

not be found in the Raman library. We cannot expect to fully identify the unknown substances. However, the combined unknown substances are likely to contribute a non-negative spectrum that is independent of the library D. The spectrum contribution from unknown substances is denoted as $f \in \mathbb{R}^N$ in the remainder of the paper. For a mixture spectrum $y \in \mathbb{R}^N$, we associate the chemicals combination model based on (1):

$$y = f + \Theta(X, C) + e \tag{7}$$

This model assumes that the spectrum f mainly lies in the difference between y and $\Theta(X, C)$. Due to the signal sparsity of Raman spectra, i.e. the Raman signals are mainly characterised by their peaks, we select an f which has the minimum overall energy and distinct peaks. We then introduce the following model to distinguish the f from (7):

$$\begin{aligned} &\min_{f,\hat{C}} \|f\|_{\widetilde{p}} + \widetilde{\lambda}\|f + \Theta(D, \hat{C}) - y\|_2 \\ &s.t. \quad f \geq 0 \\ &\hat{c}_i \geq 0 \quad for\ i = 1, 2, ..., N_D \end{aligned} \tag{8}$$

where $1 \leq \widetilde{p} \leq 2$, and $\widetilde{\lambda}$ is a positive tuning parameter.

Typically, for linearly combined components and $\widetilde{p} = 2$ in (8), f is approximately orthogonal to the vector space $span\{d_1, d_2, ..., d_{N_D}\}$ as shown in Fig. 1 and subject to the non-negativity constraint. To have a sparser estimation for f, we can further penalize the objective function by decreasing $\widetilde{p}$. In practice we find that $\widetilde{p} = 2$ is an effective configuration for resolving f.

The estimation for the parameter $\widetilde{\lambda}$ is also crucial for (8). Small $\widetilde{\lambda}$ yields very sparse f with energy focused in peaks rather than across the spectrum, while large $\widetilde{\lambda}$ gives small residual spectrum e but more sidelobes in f. In the Raman scenario, we hope to find the f which is sparse enough but also with bounded residual spectrum. Thus we can compute the $\|e\|_2$ values for a range of $\widetilde{\lambda}$ to generate a $\|e\|_2 - \widetilde{\lambda}$ curve, and the f corresponds to the "knee point" of the $\|e\|_2 - \widetilde{\lambda}$ curve which is a balance point, where the cost to alter the parameter $\widetilde{\lambda}$ is no longer worth the expected performance benefit. Then, as depicted in Fig. 3, the point can be approximated by fitting the curve to two line segments with minimum fitting errors and selecting the intersection point.

III. Experiment and Results

The model is developed in response to a DSTL (Defence Science and Technology Laboratory) challenge in which we aim to realise the Raman spectra deconvolution with a number of chemical mixtures and library spectra. To evaluate the effectiveness of the sparse regularized model in the qualitative and quantitative Raman analysis, in this section, we first use a typical mixture to illustrate the work flow for this challenge and demonstrate how to pull out unknown substances from the given mixture, and then verify the developed models by comparing them to the standard DCLS regression method using simulated data without unknown substances.

A. Experimental Setup

In this DSTL challenge, we are given the Raman spectra of several independent mixtures and 14 library spectra without any prior information. The mission is to find the components in each mixture and estimate their concentrations. The desired algorithm should work without training and be able to cope with unknown substances which are not in the library.

B. Sparse Regularized Spectra Deconvolution

In this subsection, we consider the deconvolution problem of a typical mixture spectrum $y_{475\times1}$ as shown in Fig. 2, and denote the given Raman library as $D_{475\times14}$. With the initial assumption that the mixing process Θ follows the linear combination model (2) and all mixture components are available in $D_{475\times14}$, we have $y_{475\times1} = D_{475\times14}\hat{C}_{14\times1} + e$ and the reconstructed spectrum $y^\dagger = D_{475\times14}\hat{C}_{14\times1}$.

Our first step is to make clear whether unknown substances exist. By implementing (3), we can tell if there are foreign chemicals in the mixture by comparing the spectra correlation (9), which is a convenient measure to describe how close the sample spectrum $y_{475\times1}$ matches the presumed mixing model, with a certain threshold. The threshold can be adjusted in accordance with the specific application. In this experiment, we believe unknown substances exist if $Corr(y, y^\dagger) < 0.85$.

$$Corr(y, y^\dagger) = \frac{\sum_{i=1}^{475}((y_i - \bar{y})(y_i^\dagger - \bar{y^\dagger}))}{\sqrt{(\sum_{i=1}^{475}(y_i - \bar{y})^2)(\sum_{i=1}^{475}(y_i^\dagger - \bar{y^\dagger})^2)}} \tag{9}$$

where $\bar{y}$ and $\bar{y^\dagger}$ are the means of y and $y^\dagger$.

Particularly, unknown substances are likely to exist in this mixture. Thus the reconstructed spectrum $y^\dagger = f + D_{475\times14}\hat{C}_{14\times1}$. We then utilize model (8) to resolve the f. Given the $\widetilde{p}$ value, different f can be obtained by initializing different $\widetilde{\lambda}$. Specifically, $\|e\|_2$ decreases rapidly with small $\widetilde{\lambda}$ and the $\|e\|_2 - \widetilde{\lambda}$ curve will flatten out when $\widetilde{\lambda}$ is large enough. Due to the shape of the curve, the two line segments approximately intersect at the "knee point" which is $\widetilde{\lambda} = 1.2$ given $\widetilde{p} = 2$ as shown in Fig. 3.

Subsequently, we include the derived f in the library D, then make use of the model (5) and exhaustive search to

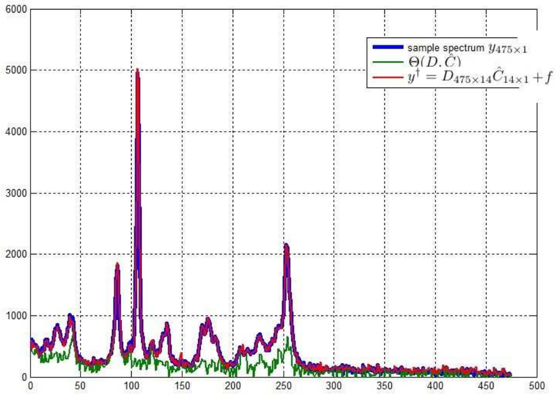

Fig. 2. Spectra plots for the sample spectrum $y_{475\times1}$ (blue), library contribution $\Theta(X,C)$ (green) and reconstructed spectrum $y^\dagger = f + \Theta(X,C)$ (red). The sample is a mixture of potentially dangerous chemicals that is of interest to DSTL.

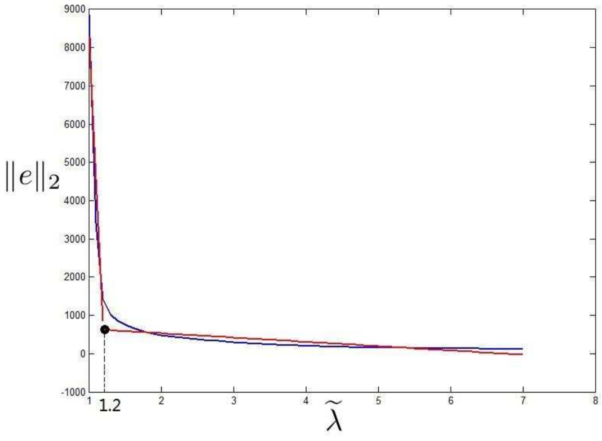

Fig. 3. Blue line shows the $||e||_2 - \widetilde{\lambda}$ curve given $\widetilde{p} = 2$. Red lines show the two line segments to fit the curve. The intersection point of red segments marks the approximated "knee point".

estimate the sample components and concentrations. A computationally cheap alternative is to employ (6) with appropriate λ to qualitatively identify the sample components.

Since the model (6) will always bias the magnitudes of the entries slightly, a standard debiasing process (10) needs to be done to eliminate the errors brought by irrelevant library entries after we have obtained the mixture components X (as presented in (1)) and the unknown substances contribution f. This process is however not necessary for (5).

$$\begin{aligned} &\min_{C} ||y - \Theta(X,C) - f||_2 \\ &s.t. \quad c_i \geq 0 \quad for\ i = 1,2,...,m \end{aligned} \tag{10}$$

where m is the number of the mixture components, C is the fixed support found from (6).

With (10), we exclude the influences from irrelevant library entries and give the quantitative estimation for the concentration vector C. The comparison between y, $\Theta(X,C)$ and the reconstructed spectrum $y^\dagger$ is shown in Fig. 2.

TABLE I
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%

In Table I, we show the quantitative results of three chemical samples using (3), (8) and (5) sequentially. The sample spectrum in Fig 2 corresponds to the 3rd mixture in this table. The three mixtures consist of same components with different concentrations, and the 2nd component is not in the library D. Based on the feedback from DSTL, our estimated components and f (in correspondence with $\widetilde{p} = 2$ in (8)) match the original 1st and 2nd component respectively, and the concentration estimation also provides certain accuracies.

C. Simulation Results

In this subsection, we compare the performance of the sparse regularized regression models and the DCLS regression model using simulated data. As DCLS is incapable of dealing with unknown components, we only consider known substances in this experiment. To generate the simulated instances of form (2), we randomly mix 2 to 5 candidates from the 14 library spectra with the average residual spectrum retrieved from our previous real data experiments. We randomly generate their concentrations and remove the components which contribute less than 10 percent to the overall spectrum intensity. Based on the linear superposition model, we then test the DCLS regression model (equivalent to (3) without the non-negativity constraint) and the sparse regularized models ((5) and (6)) with 200 randomly generated mixtures. All the spectra in the library D are normalized, the model (5) is combined with the exhaustive search and the λ in (6) is fixed as a typical constant.

We borrow the definition of confusion matrix, or error matrix [11] from the field of machine learning, and let true positive tp be the number of mixture components which have been correctly identified, false positive fp be the number of incorrectly identified components, true negative tn be the number of correctly rejected library spectra and false negative fn be the number of incorrectly rejected library spectra. In Table II, we show the statistics of our experimental results in which sensitivity $= \frac{tp}{tp+fn}$, specificity $= \frac{tn}{fp+tn}$, precision $= \frac{tp}{tp+fp}$, negative predictive value NPV $= \frac{tn}{tn+fn}$ and the proportion of true results $Acc = \frac{tp+tn}{tp+fp+tn+fn}$. Thus sensitivity represents how unlikely we are to miss the true components in our estimation, and specificity tells us how accurate our identifications are. The F1 score $f1 = \frac{2\times tp}{2\times tp+fp+fn}$ is the harmonic mean of precision and sensitivity, and can be used here as a single measure to evaluate the overall identification performance. The values in this table are calculated as the

TABLE II
STATISTICAL MEASURES OF THE SIMULATIONS

	DCLS	sparse regularized non-negative DCLS (6)	sparse regularized non-negative DCLS (5)
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

average numbers of all 200 mixtures. We can see from the table that both DCLS and sparse regularized non-negative DCLS models can achieve high sensitivities and negative predictive values in this experiment, but DCLS has relatively low precisions and specificities. It shows that the model hardly misses suspicious identifications, but may incorrectly identify redundant candidates. Our sparse regularized models work better than the standard DCLS in terms of the identification performance shown in Table II. Specifically, the combination of the model (5) and exhaustive search can achieve overall good performance (all measures are above 0.95 as shown in Table II).

IV. Conclusion

This paper presents a sparse regularized model for Raman spectroscopy to qualitatively identify mixture components and quantitatively predict the concentrations. We also show how the sparsity and non-negativity can be used to extract the spectra of possible unknown substances from the mixture spectrum. The experimental results illustrate its effectiveness to distinguish mixture compositions and retrieve possible unknown substances. Since the proposed model works as a complement to traditional regression methods, in the future work, we will combine the sparse regularizations with state-of-the-art regression techniques to further increase the capability of the model. The probability model to give detection confidences, appropriate noise cancelling and baseline shifts correction techniques for Raman spectra are also of interest to us.

Acknowledgment

The authors would like to extend their appreciation to the UK Defence Science & Technology Laboratory for helpful suggestions. This work was supported by the University Defence Research Collaboration (UDRC), EPSRC grants [EP/K014277/1].

References

[1] S. Keren, C. Zavaleta, Z. Cheng, A. de la Zerda, O. Gheysens, and S. S. Gambhir, “Noninvasive molecular imaging of small living subjects using raman spectroscopy,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 15, pp. 5844–5849, 2008.

[2] C. L. Zavaleta, B. R. Smith, I. Walton, W. Doering, G. Davis, B. Shojaei, M. J. Natan, and S. S. Gambhir, “Multiplexed imaging of surface enhanced raman scattering nanotags in living mice using noninvasive raman spectroscopy,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 32, pp. 13 511–13 516, 2009.

[3] M. J. Pelletier, “Quantitative analysis using raman spectrometry,” *Appl. Spectrosc.*, vol. 57, no. 1, pp. 20A–42A, Jan 2003.

[4] S. Li, J. Gao, J. O. Nyagilo, and D. P. Dave, “Probabilistic partial least square regression: A robust model for quantitative analysis of raman spectroscopy data,” in *Proceedings of the 2011 IEEE International Conference on Bioinformatics and Biomedicine*, ser. BIBM ’11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 526–531.

[5] O. Lyandres, R. P. Van Duyne, J. T. Walsh, M. R. Glucksberg, and S. Mehrotra, “Prediction range estimation from noisy raman spectra with robust optimization,” *Analyst*, vol. 135, pp. 2111–2118, 2010.

[6] T. Bocklitz, A. Walter, K. Hartmann, P. Rsch, and J. Popp, “How to pre-process raman spectra for reliable and stable models?” *Analytica Chimica Acta*, vol. 704, no. 12, pp. 47 – 56, 2011.

[7] N. K. Afseth, V. H. Segtnan, and J. P. Wold, “Raman spectra of biological samples: A study of preprocessing methods,” *Appl. Spectrosc.*, vol. 60, no. 12, pp. 1358–1367, Dec 2006.

[8] M. Slawski and M. Hein, “Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization,” *Electronic Journal of Statistics*, vol. 7, pp. 3004–3056, 2013.

[9] M. Fornasier and H. Rauhut, “Compressive sensing,” in *Handbook of Mathematical Methods in Imaging*. Springer, 2011, pp. 187–228.

[10] D. Ge, X. Jiang, and Y. Ye, “A note on the complexity of lp minimization.” *Math. Program.*, vol. 129, no. 2, pp. 285–299, 2011.

[11] S. V. Stehman, “Selecting and interpreting measures of thematic classification accuracy,” *Remote Sensing of Environment*, vol. 62, no. 1, pp. 77 – 89, 1997.