Supervised Raman Spectra Estimation based on Nonnegative Rank Deficient Least Squares

Barry Drake[a], Jingu Kim[b], Mahendra Mallick[a] and Haesun Park[bc]

[a]Sensors and Electromagnetic Applications Laboratory
Georgia Tech Research Institute
Georgia Institite of Technology
Atlanta, GA, U.S.A.
{barry.drake,mahendra.mallick}@gtri.gatech.edu

[b]School of Computational Science and Engineering
Georgia Institute of Technology
Atlanta, GA, U.S.A.
{jingu,hpark}@cc.gatech.edu

[c]School of Computational Sciences
Korea Institute for Advanced Study (KIAS)
Seoul, Korea

Abstract – *Raman spectroscopy is a powerful and effective technique for analyzing and identifying the chemical composition of a substance. In this paper, we focus on supervised methods for estimating Raman spectra and present a supervised method that can handle rank deficiency for estimating the Raman spectra. Earlier work has mostly assumed that the reference spectra matrix whose columns consist of the library of reference spectra are of full rank. However in practice, methods that can handle rank deficient cases, and the special case of an over complete library, are needed. We present our theoretical discovery that the active set method with a proper starting vector can actually solve the rank deficient nonnegativity-constrained least squares problems without ever running into rank deficient least squares problems during iterations. Experimental results illustrate the effectiveness of the proposed approaches.*

Keywords: Chem/Bio Detection, Raman Spectroscopy, Machine Learning, Classification, Constrained Parameter Estimation, Weighted Least Squares, Nonnegative Weighted Least Square, Rank Deficient Least Squares with Nonnegativity Constraint, Active-set Methods, Generalized Likelihood Ratio Test, Measures of Performance.

1 Introduction

The Raman effect or Raman scattering represents the inelastic quantum scattering of a photon by molecules in liquids, gases, or solids [2]. When light is incident on a molecule, most photons are scattered elastically so that the energy or frequency of the scattered photon is the same as that of the incident photon. This is known as Rayleigh scattering. A small fraction, about one in a million, is scattered inelastically, causing the frequency of the scattered photon to be different from (usually lower than) the frequency of the incident photon. This is known as Raman scattering. The frequency change is due to the change in energy levels of the vibrational or rotational energy of the molecule. Therefore, Raman spectroscopy is a powerful tool for analyzing the chemical composition of liquids, gases, or solids using a laser [2, 13, 14, 15, 16]. A Raman spectrum is a plot of the intensity of the scattered photon as a function of frequency shift. The measured Raman spectrum can be used as a fingerprint to uniquely identify the chemical composition of a substance. Application of Raman spectroscopy to analyze chemical compositions of various substances has seen rapid growth in recent years [2, 13, 14, 15, 16]. This is primarily due to the development of inexpensive and effective lasers and charge-coupled device (CCD) detectors [2]. Raman spectroscopy is also popular because measurement collection is fast and does not require contact with the chemical substance.

Suppose we have the measured Raman spectrum of a substance and we are interested in determining the chemical composition of the substance. The measured spectrum contains various error sources. Therefore, it is necessary to use a statistical measurement model that expresses the measurement as a function of the true spectrum and dominant error sources.

Raman spectrum estimation algorithms can be grouped into two types: supervised and unsupervised algorithms [12]. In the supervised approach, a library of reference Raman spectra are used and the true target spectrum is expressed as a linear combination of the reference spectra. Each reference spectrum is assumed to be error-free. In practice, this is not feasible. If the errors in a measured reference spectrum are very small compared with the signal values, then it is a good approximation to treat the measured reference spectrum as error-free. Otherwise, one must model the errors in the reference spectra. Supervised algorithms assume that the library contains all reference spectra that may be encountered in data collection. A supervised algorithm estimates the nonnegative expansion coefficients or mixing coefficients using the reference spectra and a statistical measurement model. The unsupervised approach estimates the spectra and mixing coefficients di-

rectly from measurements.

This paper examines estimation of Raman spectra using the supervised approach. In particular, we address the cases where the reference spectra matrix is rank deficient due to some columns being linearly dependent on other columns. A special case of rank deficient problem occurs when the reference spectra matrix is over complete, i.e., when there are more reference spectra (columns) than the number of the CCD array bins. This current work studies rank deficiency of the underdetermined problem, which is relevant to the application area considered. Our theoretical discovery explains that the active set method for nonnegativity constrained least squares with a proper starting vector can actually handle these rank deficient cases.

We discuss algorithms, present test results and a comparative analysis of the following supervised Raman spectra estimation methods

1. Least squares (LS)
2. Nonnegative least squares with active set method (NLS)
3. Generalized likelihood ratio test (GLRT)
4. Nonnegative generalized likelihood ratio test (NGLRT)

We use simulated data and perform Monte Carlo simulations to compare the performance of the algorithms we discuss. The measure of performance used in this study is the root mean square error (RMSE) for the mixing coefficients, i.e., the estimated solution vector of the minimization problem. We also present visual representations of the computed mixing coefficients that illustrate clearly the theoretical results developed in this paper.

The outline of the paper is as follows. In Sections 2 and 3, we describe the measurement model and measurement function for Raman spectra, respectively. We then present various algorithms that were used for Raman spectra estimation, and present the properties of the active set method as a method for rank deficient nonnegativity constrained least squares problems in Section 4. Finally, Sections 5 and 6 present numerical test results and discussions.

2 Measurement Model for Raman Spectrum

The Raman spectroscopy sensor system transmits a laser pulse and produces a measured Raman spectrum from the energy scattered by the chemical substance. The spectrum is spread across the bins of a CCD detector. The response on each bin corresponds to the amount of energy scattered at a particular frequency or wave number. The measurement model for the measured Raman spectrum is based on [17, 18] and is described in detail in [10, 12]. For the sake of completeness, we summarize the model here.

Let $\mathbf{y} \in \mathbb{R}^M$ denote a measured spectrum with values at M bins

$$\mathbf{y} = [\ y_1 \quad y_2 \quad \cdots \quad y_M\ ]^T. \tag{1}$$

The measurement model [10, 12, 17, 18] for the i^{th} element of $\mathbf{y}$ is described by

$$y_i = n_i^s + n_i^b + g_i,\ i = 1, 2, \cdots, M, \tag{2}$$

where n_i^s and n_i^b represent the number of photoelectrons generated by the signal and background noise, respectively, and g_i is the Gaussian readout noise from the amplifier. The random variables (RVs) n_i^s and n_i^b have Poisson distribution with parameters λ_i^s and λ_i^b, respectively. The Gaussian RV g_i has mean m and variance σ^2, which are assumed to be known. We assume that the RVs n_i^s, n_i^b, and g_i are independent and also independnet of similar RVs at other bins.

Based on our assumptions, we have

$$n_i^s \sim p_{\text{Poisson}}(n_i^s; \lambda_i^s), \tag{3}$$

$$n_i^b \sim p_{\text{Poisson}}(n_i^b; \lambda_i^b), \tag{4}$$

$$p_{\text{Poisson}}(x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!},\ x = 0, 1, 2, \cdots \tag{5}$$

$$g_i \sim N(g_i; m, \sigma^2), \tag{6}$$

$$E\{(g_i - m)(g_j - m)\} = \delta_{ij}\sigma^2, \tag{7}$$

where λ_i^s and λ_i^b represent the expected numbers of counts for the signal and background noise, respectively. We note that the mean and variance of a Poisson distributed RV x with parameter λ are equal to λ.

If $\lambda_i^s + \lambda_i^b$ is large (e.g. greater than 50), then $n_i^s + n_i^b$ is well approximated by a Gaussian distribution

$$n_i^s + n_i^b \sim N(n_i^s + n_i^b; \lambda_i^s + \lambda_i^b, \lambda_i^s + \lambda_i^b). \tag{8}$$

This approximation is called the large signal approximation. We can show that under the large signal approximation [10]

$$\mathbf{y} = \boldsymbol{\lambda}^s + \boldsymbol{\lambda}^b + \mathbf{m} + \mathbf{v}, \tag{9}$$

where

$$\begin{aligned}
\mathbf{m} &= m[\ 1 \quad 1 \quad \cdots \quad 1\ ]^T, \\
\boldsymbol{\lambda}^s &= [\ \lambda_1^s \quad \lambda_2^s \quad \cdots \quad \lambda_M^s\ ]^T, \\
\boldsymbol{\lambda}^b &= [\ \lambda_1^b \quad \lambda_2^b \quad \cdots \quad \lambda_M^b\ ]^T, \\
\mathbf{v} &\sim N(\mathbf{v}; \mathbf{0}_{M\times 1}, \mathbf{R}), \\
\mathbf{R} &= \text{diag}(\lambda_1^s + \lambda_1^b + \sigma^2, \cdots, \lambda_M^s + \lambda_M^b + \sigma^2).
\end{aligned} \tag{10}$$

3 Measurement Function for Raman Spectra

Suppose we have N reference spectra $\{\mathbf{s}_j \in \mathbb{R}^M\}_{j=1}^N$ in our library corresponding to N chemical substances. Then the true target spectrum $\mathbf{s}$ can be expressed as a linear combination of the reference spectra by

$$\mathbf{s} = \Sigma_{j=1}^N x_j \mathbf{s}_j. \tag{11}$$

We can write (11) in the matrix form

$$\mathbf{s} = \mathbf{A}\mathbf{x}, \tag{12}$$

where

$$\begin{aligned} \mathbf{A} &= [\ \mathbf{s}_1 \quad \mathbf{s}_2 \quad \cdots \quad \mathbf{s}_N\ ], \\ \mathbf{x} &= [\ x_1 \quad x_2 \quad \cdots \quad x_N\ ]^T \geq \mathbf{0}. \end{aligned} \tag{13}$$

The parameter vector $\boldsymbol{\lambda}^s$ and the true specrum $\mathbf{s}$ are ralated by

$$\boldsymbol{\lambda}^s = \mathbf{P}\mathbf{s}, \tag{14}$$

where $\mathbf{P}$ is the $M \times M$ *point spread function matrix* of the diffraction grating used to spread the spectral energy across the CCD bins. Substitution of (12) in (14) gives

$$\boldsymbol{\lambda}^s = \boldsymbol{\Phi}\mathbf{x}, \quad \text{where} \quad \boldsymbol{\Phi} = \mathbf{P}\mathbf{A}.$$

Not all photons that hit the CCD array are converted to photoelectrons. The quantum efficiency or flat-field response varies along the CCD array. This non-uniform detector efficiency is modeled by

$$\lambda_i^s = \beta_i(\boldsymbol{\Phi}\mathbf{x})_i, \tag{15}$$

where β_i is known from calibration measurements. We can write (15) in the matrix form

$$\boldsymbol{\lambda}^s = \mathbf{C}\mathbf{x}, \tag{16}$$

where

$$\mathbf{C} = \mathbf{B}\boldsymbol{\Phi} = \mathbf{B}\mathbf{P}\mathbf{A}, \quad \mathbf{B} = \text{diag}(\beta_1, \beta_2, \cdots, \beta_M). \tag{17}$$

Under the large signal approximation, substitution of (16) in (9) gives

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \boldsymbol{\lambda}^b + \mathbf{m} + \mathbf{v}.$$

Define the new measurement vector $\mathbf{z}$

$$\mathbf{z} = \mathbf{y} - \boldsymbol{\lambda}^b - \mathbf{m}.$$

Then

$$\mathbf{z} = \mathbf{C}\mathbf{x} + \mathbf{v}. \tag{18}$$

Thus, under the large signal approximation, the measurement model is linear with additive Gaussian measurement noise. An estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$ can be obtained using the maximum likelihood estimator (MLE) [11, 4] or weighted least squares (WLS) [11, 4] with nonnegativity constraints (13). Thus, the estimation problem is a constrained estimation problem due to (13), and the use of classical MLE or WLS would yield approximate results. In this paper, we address estimation of Raman spectra under the large signal assumption. Future work will address the more general case where the large signal assumption is not valid.

4 Raman Spectra Estimation Algorithms

Since the measurement model in (18) is linear with additive Gaussian noise, the estimates from the MLE and WLS are the same provided that the weight matrix $\mathbf{W}$ in WLS has a certain form, see [4, 11]. Then using (11),

$$\mathbf{W} = \mathbf{U}^T\mathbf{U} = \text{diag}(w_1, w_2, \cdots, w_M), \tag{19}$$

where

$$\begin{aligned} w_i &= 1/(\lambda_i^s + \lambda_i^b + \sigma^2),\ i = 1, 2, \cdots, M, \\ \mathbf{U} &= \text{diag}(\sqrt{w_1}, \sqrt{w_2}, \cdots, \sqrt{w_M}). \end{aligned}$$

The cost function for parameter estimation is

$$J(\mathbf{x}) = (\mathbf{z} - \mathbf{C}\mathbf{x})^T\mathbf{W}(\mathbf{z} - \mathbf{C}\mathbf{x}) \tag{20}$$

$$= \|\mathbf{U}(\mathbf{C}\mathbf{x} - \mathbf{z})\|_2^2 = \|\mathbf{H}\mathbf{x} - \mathbf{g}\|_2^2, \tag{21}$$

where the weighted measurement vector $\mathbf{g} \in \mathbb{R}^M$ and the weighted measurement matrix $\mathbf{H} \in \mathbb{R}^{M \times N}$ are defined by

$$g_i = \sqrt{w_i}z_i,\ i = 1, 2, \cdots, M,$$

$$d_{ij} = \sqrt{w_i}c_{ij},\ i = 1, 2, \cdots, M,\ j = 1, 2, \cdots, N.$$

As shown in the above equation, the WLS cost function in (20) is equivalent to the LS cost function in (21). Thus, in the following algorithm development, for simplicity of discussion, we will not differentiate between the unweighted and the weighted least squares problems. In addition, although these algorithms assume the same measurement variances for all measurements, i.e., $\mathbf{W} = \boldsymbol{\sigma}_v^2\mathbf{I}$, they can be easily modified to handle non-uniform weights, which is application dependent, e.g., the application discussed in this paper.

4.1 Least Squares (LS) and the Generalized Likelihood Ratio Test (GLRT)

Unconstrained least squares (LS) [4, 11] solves the following problem:

$$\min_{\mathbf{x}} J(\mathbf{x}). \tag{22}$$

Since LS does not enforce the nonnegativity constraints, the estimate obtained by LS might contain negative values. The algorithms for solving (22) can be divided into two groups depending on the rank of matrix $\mathbf{H}$: When $\mathbf{H}$ is of full column-rank, then the QR decomposition method is recommended although the method of the normal equations is commonly used as well. When $\mathbf{H}$ is rank deficient, then algorithms that reveal the rank of the matrix are needed. The best algorithm in terms of numerical stability would be based on the singular value decomposition (SVD) of the matrix $\mathbf{H}$. As a faster approximation, several methods that approximates the SVD can be used which include those based on the complete orthogonal decompositions and rank-revealing URV and ULV decompositions [19].

We describe briefly the subspace version of the GLRT [5]. The likelihood function is $p(\mathbf{z}; \hat{\mathbf{x}}, \mathbf{R}, H_1)$ for the

measurement model (18), with hypothesis H_1, where all of the reference spectra are the columns of $\mathbf{C}$ and $\hat{\mathbf{x}}$ is the estimate of the mixing coefficients. A leave-one-out strategy is used to leave the i^{th} column out of $\mathbf{C}$ and the likelihood function $p(\mathbf{z}; \hat{\mathbf{x}}_\mathbf{0}, \mathbf{R}_\mathbf{0}, H_0)$ is formed for the alternative or null hypothesis. $\hat{\mathbf{x}}_\mathbf{0}$ is estimated and the log of the ratio is tested against a threshold α

$$\ln p(\mathbf{z}; \hat{\mathbf{x}}, \mathbf{R}, H_1) - \ln p(\mathbf{z}; \hat{\mathbf{x}}_\mathbf{0}, \mathbf{R}_\mathbf{0}, H_0) > \alpha \qquad (23)$$

If this is satisfied, then the i^{th} chemical is assumed to be present and that spectra is used to form one of the columns of the measurement matrix $\mathbf{C}$. If the test is satisfied p times then $\mathbf{C}_\mathbf{p}$ is formed and the estimate $\hat{\mathbf{x}}_\mathbf{p}$ is determined. The estimation algorithm can be any of the usual techniques. Here we choose least squares (GLRT) and nonnegative least squares (NGLRT) for comparison.

4.2 Nonnegative Least Squares (NLS) for Full Rank Problems

For the measurement model (18), the nonnegativity-constrained least squares (NLS) or nonnegative MLE (NMLE) solves the problem

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\geq 0} J(\mathbf{x}). \qquad (24)$$

Due to the nonnegativity constraints, algorithms for (24) are more complicated than those for unconstrained problems. In this section, we describe algorithms for solving (24) for both cases where $\mathbf{H}$ is of full column rank and $\mathbf{H}$ is rank deficient.

A standard algorithm for (24) is the active-set method described in Lawson and Hanson [9], and its implementation is included in MATLAB® as function *lsqnonneg*. We summarize the algorithm in Algorithm 1 and will refer to it when we discuss the rank deficient case.

A key objective of Algorithm 1 is to identify what variables are zero or non-zero in the solution. For the solution $\mathbf{x}^*$ in (24), the index set

$$\mathcal{E}^* = \{j | x_j^* = 0, j = 1, 2, \cdots, N\} \qquad (25)$$

is called the *active set* since the nonnegativity constraints are actively satisfied in those indices. Similarly, the index set

$$\mathcal{S}^* = \{j | x_j^* \neq 0, j = 1, 2, \cdots, N\} \qquad (26)$$

is called the *passive set*. In the algorithm, we maintain *working sets* $(\mathcal{E}, \mathcal{S})$ as candidates for $(\mathcal{E}^*, \mathcal{S}^*)$ and iteratively exchange variables between $\mathcal{E}$ and $\mathcal{S}$ until $(\mathcal{E}^*, \mathcal{S}^*)$ is found. We typically start from the all zero solution, i.e., $\mathcal{S} = \emptyset$.

The algorithm is composed of two nested loops: the inner loop (Steps 7 to 11) and the outer loop (Steps 3 to 14). In the inner loop, the unconstrained least squares solution with respect to the current passive set $\mathcal{S}$ is computed (Step 11). If the unconstrained solution is feasible, the inner loop is terminated (Step 13); otherwise, a step length is chosen so that at least one passive variable becomes active (Step 8), and the loop is repeated. In the outer loop, a check is made to determine if the current solution obtained from the inner loop is the desired solution (Step 2). If it is not the desired solution, then one index is chosen from the active set and moved to the passive set (Steps 3-4).

For the case that the matrix $\mathbf{H}$ is full column rank, the correctness of Algorithm 1 is proved in [9]. We briefly discuss the key ideas of the proof. These results will be essential in establishing the fact that Algorithm 1 is also valid for rank deficient cases with no modification. As the first step, we verify that all the steps in Algorithm 1 are well defined when the matrix $\mathbf{H}$ has full rank. In particular, the following lemma plays a key role. For the full proof, see [9] Chapter 23.

Lemma 1. *In Algorithm 1, the solution* $\mathbf{z}$ *obtained in Step 5 satisfies* $z_t > 0$ *where* t *is the index chosen in Step 3.*

To understand the important implication of Lemma 1, let us assme that the statement is not true, i.e., $z_t \leq 0$. In this case, the step length α in Step 8 is zero, and therefore the current solution candidate $\mathbf{x}$ is not updated in Step 9 making further updates impossible. Therefore, Lemma 1 shows that α can be positive and the steps are well defined.

In addition, the following statements show that Algorithm 1 terminates in a finite number of iterations. For the inner loop, observe that at least one index is removed from $\mathcal{S}$ at each iteration. Hence, the inner loop terminates in at most $|\mathcal{S}|$ steps. The finiteness of the outer loop can be shown by considering the value of the cost function $J(\mathbf{x})$. Because the value of $J(\mathbf{x})$ is strictly reduced after each iteration, set $\mathcal{S}$ at Step 4 is different from all the previous instances of itself. Since only a finite number of cases are possible for set $\mathcal{S}$, the outer loop terminates in a finite number of iterations. In practice, the number of iterations of the outer loop is usually the same or slightly bigger than the size of the passive set, $|\mathcal{S}^*|$.

4.3 Rank Deficient NLS

When $\mathbf{H}$ in (21) is rank deficient, Algorithm 1 is applicable without modification. We now prove this by asserting and proving a lemma regarding the state of the inner-loop subproblem in the presence of the overall rank deficiency of $\mathbf{H}$.

A key issus is whether $\mathbf{H}_\mathcal{S}$ ever becomes rank deficient during the execution of Algorithm 1. If this happens, a problem arises because the solutions in Steps 5 and 11 are not uniquely determined, and then Lemma 1 might not hold. If Lemma 1 does not hold, it is difficult to show the finite termination property. In the following, however, we show that the columns in $\mathbf{H}_\mathcal{S}$ indeed remain linearly independent throughout all iterations.

Lemma 2. *In Algorithm 1, the column corresponding to the index* t *chosen in Step 3 is linearly independent of the columns indexed by the current* $\mathcal{S}$*.*

Proof. Assume that $k \notin \mathcal{S}$ and denote the corresponding column of $\mathbf{H}$ by $\mathbf{h}_k$. We will show that if $\mathbf{h}_k$ is lineary dependent on the columns in $\mathbf{H}_\mathcal{S}$, then k is not selected in Step 3.

Note that, at the end of the previous iteration of the inner loop, $\mathbf{x}$ is feasible ($\mathbf{x} \geq 0$) and is the optimal solution with respect to the current passive set $\mathcal{S}$. We therefore have that

$$\frac{\partial \|\mathbf{H}_\mathcal{S}\mathbf{x}_\mathcal{S} - \mathbf{g}\|_2^2}{\partial x} = \mathbf{H}_\mathcal{S}^T\mathbf{H}_\mathcal{S}\mathbf{x}_\mathcal{S} - \mathbf{H}_\mathcal{S}^T\mathbf{g} = 0. \qquad (27)$$

Algorithm 1 NLS : This algorithm [9] solves $\min_{\mathbf{x}\geq 0} \|\mathbf{Hx}-\mathbf{g}\|_2$ where $\mathbf{H}\in\mathbb{R}^{M\times N}$, $\mathbf{g}\in\mathbb{R}^M$

1: $\mathbf{x}:=0$, $\mathcal{E}:=\{1,2,\cdots,N\}$, $\mathcal{S}:=\emptyset$, $\mathbf{w}:=\mathbf{H}^T(\mathbf{g}-\mathbf{Hx})$.
2: **while** $\mathcal{E}\neq\emptyset$ and $\exists j\in\mathcal{E}$ such that $w_j>0$ **do**
3: Find $t\in\mathcal{E}$ such that $w_t=\max\{w_j : j\in\mathcal{E}\}$.
4: Move the index t from $\mathcal{E}$ to $\mathcal{S}$.
5: Let $\mathbf{H}_\mathcal{S}$ denote the $M\times|\mathcal{S}|$ submatrix of $\mathbf{H}$ containing only the columns indexed by $\mathcal{S}$. For $\mathbf{z}\in\mathbb{R}^N$, let $\mathbf{z}_\mathcal{S}$ be the subvector of $\mathbf{z}$ indexed by $\mathcal{S}$. Define $\mathbf{z}_\mathcal{E}$ similarly. Then, solve $\min_{\mathbf{z}_\mathcal{S}}\|\mathbf{H}_\mathcal{S}\mathbf{z}_\mathcal{S}-\mathbf{g}\|_2$ and set $\mathbf{z}_\mathcal{E}:=0$.
6: **while** $z_j\leq 0$ for any $j\in\mathcal{E}$ **do**
7: Find $q\in\mathcal{S}$ such that $x_q/(x_q-z_q)=\min\{x_j/(x_j-z_j) : z_j\leq 0, j\in\mathcal{S}\}$.
8: $\alpha:=x_q/(x_q-z_q)$.
9: $\mathbf{x}:=\mathbf{x}+\alpha(\mathbf{z}-\mathbf{x})$.
10: Move from $\mathcal{S}$ to $\mathcal{E}$ all indices $j\in\mathcal{S}$ for which $x_j=0$.
11: Solve $\min_{\mathbf{z}_\mathcal{S}}\|\mathbf{H}_\mathcal{S}\mathbf{z}_\mathcal{S}-\mathbf{g}\|_2$ and set $\mathbf{z}_\mathcal{E}:=0$.
12: **end while**
13: $\mathbf{x}:=\mathbf{z}$.
14: $\mathbf{w}:=\mathbf{H}^T(\mathbf{g}-\mathbf{Hx})$.
15: **end while**

Now, if $\mathbf{h}_k$ is lineary dependent on the columns in $\mathbf{H}_\mathcal{S}$, then $\mathbf{h}_k=\mathbf{H}_\mathcal{S}\mathbf{c}$ with some vector $\mathbf{c}$. Then, the k^{th} element of $\mathbf{w}$ in Step 14 is

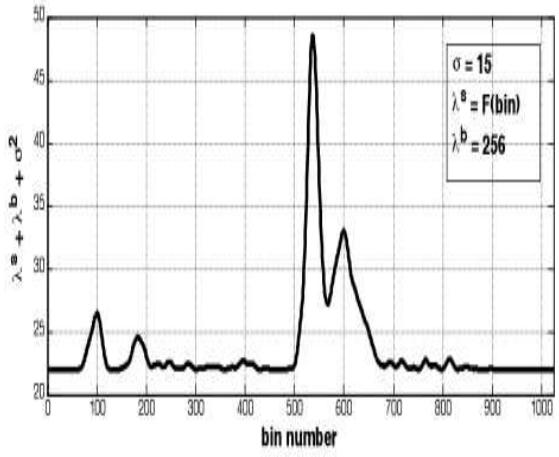
$$\begin{aligned} w_k &= (\mathbf{h}_k)^T(\mathbf{H}^T\mathbf{x}-\mathbf{g}) \\ &= (\mathbf{H}_\mathcal{S}\mathbf{c})^T(\mathbf{H}_\mathcal{S}^T\mathbf{x}_\mathcal{S}-\mathbf{g}) \\ &= \mathbf{c}^T\mathbf{H}_\mathcal{S}^T(\mathbf{H}_\mathcal{S}^T\mathbf{x}_\mathcal{S}-\mathbf{g})=0, \end{aligned}$$

using (27). Therefore, k is not selected in Step 3. □

Lemma 2 shows that $\mathbf{H}_\mathcal{S}$ does not become rank deficient after Step 3. Because the inner loop only reduces the passive set $\mathcal{S}$, $\mathbf{H}_\mathcal{S}$ does not become rank deficient during the inner loop. Hence, Lemma 2 is enough to show that $\mathbf{H}_\mathcal{S}$ remains full-column rank throughout the iterations. The remaining argument of the finite termination property of Algorithm 1 for the rank deficient case is the same as that of the full rank case described above.

Our proof provides further understanding regarding the initialization of the active-set method. Although $\mathbf{x}$ is initialized with a zero vector in Algorithm 1, it is possible to use prior information and initialize $\mathbf{x}$ by a non-zero vector. When $\mathbf{H}$ is rank deficient, however, care must be taken if $\mathbf{x}$ is initialized with a non-zero vector. If $\mathbf{x}$ is set to be a zero vector initially, then $\mathcal{S}$ is initially empty and the column rank of $\mathbf{H}_\mathcal{S}$ remains full as we have shown above. If $\mathbf{x}$ is initialized with a non-zero vector for which the corresponding $\mathbf{H}_\mathcal{S}$ is rank deficient, then the steps of Algorithm 1 might not be well defined. Therefore, unless we have other information that an initial value of $\mathbf{x}$ can be set to non-zero and the corresponding $\mathbf{H}_\mathcal{S}$ has full column rank initially, Algorithm 1 needs to be started from $\mathbf{x}=0$, i.e., with $\mathcal{S}=\emptyset$. For this case, the algorithm will correctly find a solution even when the matrix is rank deficient without ever running into rank deficient subproblems.

Figure 1: *Variation of measurement error variance with bin index.*



5 Numerical Simulation and Results

We implemented several algorithms discussed in this paper in MATLAB® 7.9 (R2009b) and compared them for several datasets for supervised Raman spectra estimation. All experiments were conducted on an Intel Core 2 Quad processor with the Windows XP operating system and 4GB of RAM. One thousand Monte Carlo trials were used to calculate measures of performance for each spectral estimation algorithm.

5.1 Data Sets and Experimental Settings

We used 67 reference Raman spectra, $\{\mathbf{s}_j\in\mathbb{R}^M\}_{j=1}^N$, $N=67$. Each spectrum has values at $M=1024$ bins. In the Monte Carlo simulations, the mean and variance of the Gaussian measurement noise are 10 and 225, respectively. We used a constant value of 256 for the Poisson parameter λ_i^b for all bin values. We then calculated λ^s by selecting and substituting a true $\mathbf{x}$ vector into (16)-(17) and (13).

In previous work we studied various cases with multiple chemicals present. After discussions with field experts, in this present work we focus on the case with 3 chemicals present, which turns out to be the scenario often encountered in the field. These will be referred to as chemicals spec3, spec5, and spec30. The concentrations were set at 0.3 g/m^2 with a uniform distribution for each chemical species.

Two sets of reference spectra have been generated: one from laboratory samples and another derived from the first with a perturbation to force the resulting reference spectra matrix to be rank deficient. The former will be referred to as refFR (Full Rank); the latter as refRD (Rank Deficient). The perturbation was performed by modifying the singular value decomposition of refFR so that reconstruction of the new reference spectra matrix resulted in a reduced rank reference spectra matrix.

All of this can be summarized in the following equations: as in the above sections, let $\mathbf{H}$ be full column rank (refFR) and $\mathbf{H_r}$ be the rank reduced matrix of reference spectra (refRD). Consider the singular value

Algorithm	Estimation RMSE	
	refFR	refRD
LS	0.0908	0.1894
GLRT	0.2197	0.4064
NGLRT	0.0162	0.0154
NNLS	0.0024	0.0051

Table 1: Overall errors for unweighted versions of the algorithms.

decomposition (SVD) of $\mathbf{H}$. Then $\mathbf{H}$ and $\mathbf{H_r}$ are related by

$$\begin{aligned} \mathbf{H} &= \mathbf{U}\mathbf{S}\mathbf{V}^T \\ \mathbf{H_r} &= \mathbf{U}(\mathbf{S}+\mathbf{E})\mathbf{V}^T \end{aligned} \tag{28}$$

where $\mathbf{S} = diag(\sigma_1, ..., \sigma_{n-2}, \sigma_{n-1}, \sigma_n)$, $\mathbf{E} = diag(0_1, ..., 0_{n-2}, \alpha + \epsilon - \sigma_{n-1}, -\sigma_n)$, ϵ is machine precision, and α is a small constant determined to make the matrix just on the "edge" of rank deficiency in the second to last singular value.

Finally, we characterize the variation in quantum efficiency of the CCD array in Figure 1, which shows the variation of the measurement error variance with bin index for the current scenario. We observe that the measurement error variance changes significantly with the bin index - this shows that the weighted algorithm methods model the physics more realistically than unweighted methods [10].

5.2 Performance Evaluation Measures

Let M_s be the total number of Monte Carlo simulations and $\hat{\mathbf{x}}_m$ the estimate of $\mathbf{x}$ in the mth Monte Carlo simulation. The estimation error in the jth component of $\mathbf{x}$ in the mth Monte Carlo simulation is defined by

$$\tilde{x}_{m,j} := x_j - \hat{x}_{m,j}, \ j = 1, 2, ..., N. \tag{29}$$

The root mean square error (RMSE) for the jth coefficient and the overall RMSE for the coefficients are defined, respectively, by

$$\text{RMSE}_{x,j} := \left[\frac{1}{M_s} \sum_{m=1}^{M_s} \tilde{x}_{m,j}^2 \right]^{1/2}, \ j = 1, 2, ..., N, \tag{30}$$

$$\text{RMSE}_x := \left[\frac{1}{NM_s} \sum_{j=1}^{N} \sum_{m=1}^{M_s} \tilde{x}_{m,j}^2 \right]^{1/2}. \tag{31}$$

5.3 Experimental Results

Table 1 summarizes overall RMSE results for the parameter estimation, i.e. the estimation of the mixing coefficients for each algorithm averaged over all Monte Carlo runs. The algorithms are listed on the left and estimation RMSE results are listed for each algorithm in the two columns to the right. Note that there is a slight increase for the NGLRT algorithm going from the full rank to rank deficient cases. More Monte Carlo simulations would probably show that this is not significant.

In Figure 2, we observe that, for the full rank case (refFR), the least squares algorithm performs fairly well at determining the correct mixing coefficients for spec3, spect5, and spec30 and their concentrations. This can be seen by the blue spikes at chemical numbers 3, 5, and 30 on the abscissa. However, for refRD, the least squares algorithm completely misses spec30 and finds additional chemicals at spec7 and spec55.

Figure 2: *Least squares solution vectors for the full rank (blue) and rank deficient (red) cases.*

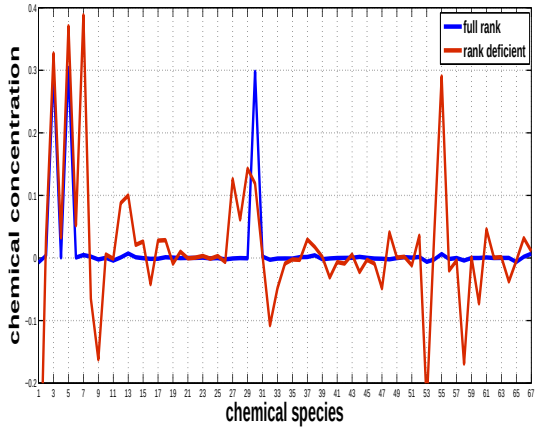


Figure 3: *Generalized likelihood ratio test solution vectors for the full rank (blue) and rank deficient (red) cases.*

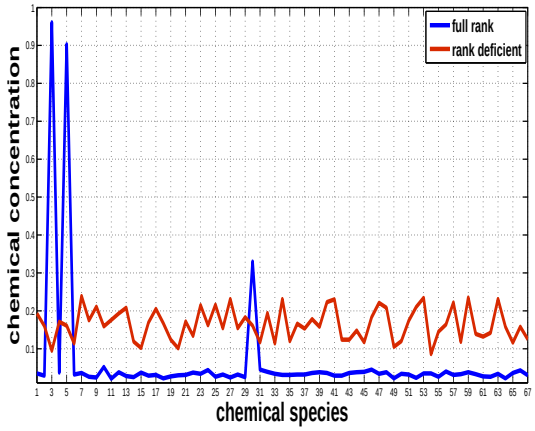


Figure 3 shows the results for the generalized likelihood ratio test. The results for the full rank matrix of spectra are similar to the least squares results except that the concentrations are not determined correctly (blue curve). When the matrix is rank deficient, GLRT completely fails to estimate the mixing coefficients (red curve).

In Figures 4 and 5 we see that both the NGLRT and the NLS algorithms were unaffected by rank deficiency. However, the NLS was able to both determine the chemical species present and their concentrations. In these two figures the plots were so close that a color scheme was used to show the results with the full rank case in thick green and the rank deficient results in red and "inside" the green curve.

Figures 6 and 7 are surface plots of the solution vectors for all 1000 Monte Carlo runs as viewed from

Figure 4: *Nonnegative generalized likelihood ratio test solution vectors for the full rank (blue) and rank deficient (red) cases.*

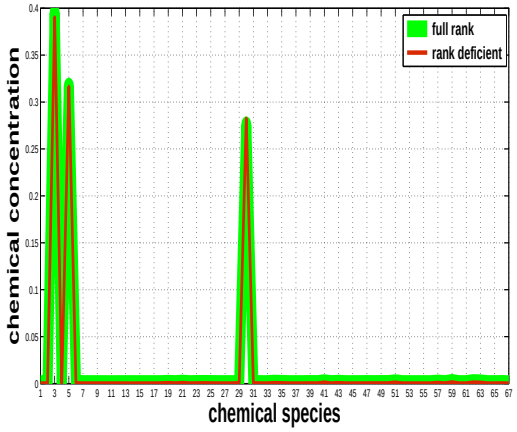


Figure 5: *Nonnegative least squares solution vectors for the full rank (blue) and rank deficient (red) cases.*

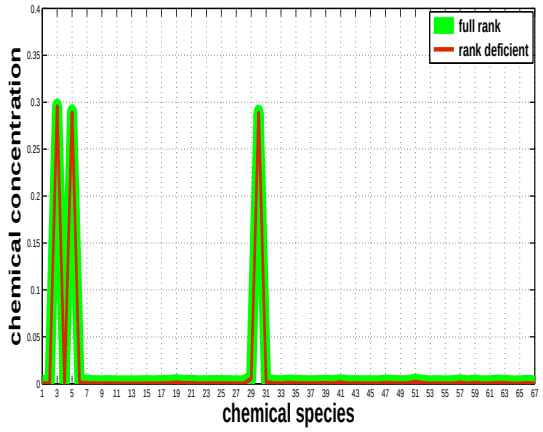


Figure 6: *Least squares solutions for the full rank case for all 1000 Monte Carlo runs (green) and the same for NLS (black).*

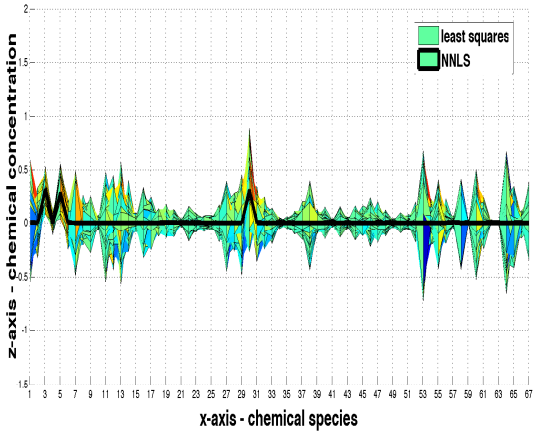
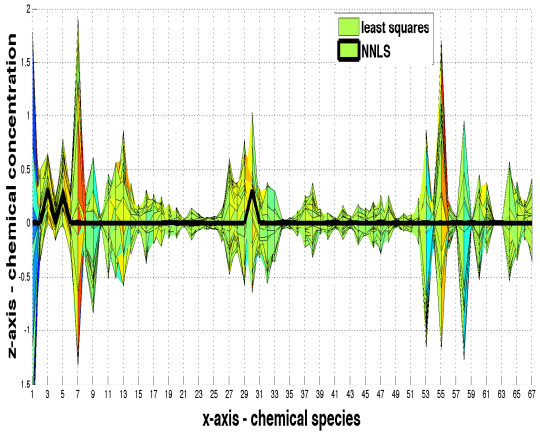


Figure 7: *Least squares solutions for the rank deficient case for all 1000 Monte Carlo runs (green) and the same for NLS (black).*



the x,z-plane (chemical number, concentration). These plots are for the least squares algorithm operating on the full rank (Fig. 6) and rank deficient (Fig. 7) cases. Superimposed in black are the same plots of the NLS Monte Carlo runs for the corresponding cases. This demonstrates the large variation over Monte Carlo runs when rank deficiency is introduced when using the least squares algorithm. In contrast the NLS algorithm exhibits very little variation and is unaffected by the rank deficiency, as anticipated by the theoretical results in the previous sections.

6 Conclusions and Discussions

In this paper we presented theoretical and experimental results for estimating the mixing coefficients of chemical species sampled using a Raman spectroscopy instrument. We reviewed the measurement model as implemented in our simulation. The mixing coefficients were estimated using four algorithms: least squares, generalized likelihood ratio test, nonnegative least squares, and nonegative generalized likelihood ratio test.

The theoretical results show that the NLS-based algorithms should be robust in the presense of a rank deficient coefficient matrix. For the applicatoin under study, this is the reference Raman spectra matrix used to find the mixing coefficients that allow matching a sample to chemical species in the reference Raman spectra "library".

Experimental results verify the theoretical results by demonstrating a reduction of the overall RMSE for the NLS-based methods over the LS and GLRT algorithms. The experiments were conducted using 1000 Monte Carlo simulations for each algorithm with a scenario that is similar to those found in practice. Also presented were visual confirmation of the results using graphs of the computed mixing coefficients averaged over the 1000 Monte Carlo simulations and surface plots of all 1000 MC runs.

The results demonstrate that for supervised learning of the mixing coefficients NLS using the active set method should be preferred due to its robustness in the presence of rank degeneracy.

Future research will focus on imposing sparsity constraints on the nonnegativity constrained least squares problems. This is a better model for the problem, which is expected to improve the solution, especially in the presence of multiple chemicals.

Acknowledgment

The work of B. Drake was supported in part by the ONR Grant N00014-07-1-0378. The work of J. Kim and H. Park was supported in part by the National Science Foundation grants CCF-0732318 and CCF-0808863 and in part by the Korea Institute for Advanced Study, Seoul, Korea. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The work of J. Kim was also supported in part by the Samsung Foundation of Culture scholarship awarded to him.

References

[1] R. Bro and S. De Jong. A fast non-negativity-constrained least squares algorithm. *Journal of Chemometrics*, 11(5):393–401, 1997.

[2] J.R. Ferraro, K. Nakamoto, and C.W. Brown. *Introductory Raman Spectroscopy*. Academic Pr, 2003.

[3] J. J. Judice and F. M. Pires. A block principal pivoting algorithm for large-scale strictly monotone linear complementarity problems, Computers and Operations Research, 21(5):587596, 1994.

[4] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume 1: Estimation Theory*, Prentice Hall, 1993.

[5] S. M. Kay, *Fundamentals of Statistical Signal Processing, Vol 2: Detection Theory*, Prentice Hall, 1998.

[6] H. Kim and H. Park, Sparse Non-negative Matrix Factorizations via Alternating Non-negativity-constrained Least Squares for Microarray Data Analysis, *Bioinformatics*, Vol. 23, No. 12, pp. 1495-1502, 2007.

[7] H. Kim and H. Park, "Non-negative Matrix Factorization based on Alternating Non-negativity Constrained Least Squares and Active Set Method," *SIAM J. on Matrix Analysis and Applications*, Vol 30, No. 2, pp. 713-730, 2008.

[8] J. Kim and H. Park. Toward faster nonnegative matrix factorization: a new algorithm and comparisons. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pp. 353–362. IEEE Computer Society, 2008.

[9] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, 1974.

[10] M. Mallick, B. Drake, H. Park, A. Register, D. Blair, P. West, R. Palkki, A. Lanterman, and Darren Emge, Comparison of Raman Spectra Estimation Algorithms, *Proc. 2009 International Conference on Information Fusion*, July 6-9, Seattle, USA.

[11] J.M. Mendel. Lessons in estimation theory for signal processing, communications, and control(Book). *Englewood Cliffs, NJ: Prentice Hall PTR*, 1995.

[12] R. D. Palkki, A. D. Lanterman, "Algorithms and Performance Bounds for Chemical Identification under a Poisson Model for Raman Spectroscopy," *Proc. of the Twelfth International Conference on Information Fusion*, July 6-9, Seattle, USA.

[13] P.L. Ponsardin, NS Higdon, T.H. Chyba, W.T. Armstrong, A.J. Sedlacek III, S.D. Christesen, and A. Wong. Expanding applications for surface-contaminant sensing using the laser interrogation of surface agents (LISA) technique. In *Proceedings of SPIE*, volume 5268, pp. 321–327, 2004.

[14] A.J. Sedlacek III, S.D. Christesen, T. Chyba, and P. Ponsardin. Application of UV-Raman spectroscopy to the detection of chemical and biological threats. In *Proceedings of SPIE*, volume 5269, p.23, 2004.

[15] M.A. Slamani, T.H. Chyba, H. LaValley, and D. Emge. Spectral unmixing of agents on surfaces for the Joint Contaminated Surface Detector (JCSD). In *Proceedings of the SPIE*, volume 6699, 2007.

[16] M-A Slamani, B. Fisk, and T. Chyba, D. Emge, and S. Waugh, "An algorithm benchmark data suite for chemical and biological (chem/bio) defense applications," *Proc. Signal and Data Processing of Small Targets*, Vol. 6969, March 18-20, 2008, Orlando, FL, USA.

[17] D.L. Snyder, A.M. Hammoud, and R.L. White. Image recovery from data acquired with a charge-coupled-device camera. *Journal of the Optical Society of America A*, 10(5):1014–1023, May 1993.

[18] DL Snyder, TJ Schulz, and JA O'Sullivan. Deblurring subject to nonnegativity constraints. *IEEE Transactions on signal processing*, 40(5):1143–1150, 1992.

[19] G.W. Stewart. An updating algorithm for subspace tracking. *Transactions on signal processing*, 40(6):15351541, 1992.

[20] M.H. Van Benthem and M.R. Keenan. Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems. *Journal of chemometrics*, 18(10):441–450, 2004.