Automated Analysis of Reflection Mode Terahertz Hyperspectral Images

Mark Berman, Krunal Radhanpura, and David Farrant, *Senior Member, IEEE*

***Abstract*—A suite of algorithms and associated procedures, originally developed for mineral exploration applications, are adapted for application to terahertz hyperspectral images measured in reflection mode. Such data are often quite noisy due to the low reflectivity of many materials at terahertz frequencies. The algorithms and procedures are based on an extended linear mixture model consisting of two parts. The first part, called the "foreground", models the distinguishing parts of the spectra of materials (including mixtures) of interest (especially their diagnostic absorption features). The second part, called the "background", models parts of the spectra that are typically of lesser interest, such as variation in low frequencies and water vapor. The model and procedures are exemplified with a spectral library of six materials and are applied to three hyperspectral images, one consisting only of pure pellets, some of which are not in the library, and two of which contain both pure and mixed pellets of three of the materials in the library. The associated procedures include the following: estimating the number of materials in the mixture at each pixel; identifying pixels with materials that are well modeled by the background terms only; identifying pixels with materials not in the library; and identifying pixels containing metal. Finally, this article concludes with a discussion of some outstanding issues.**

***Index Terms*—Hyperspectral image, mixture analysis, reflection mode, terahertz time-domain spectroscopy (THz-TDS).**

I. Introduction

In RECENT decades, there has been increasing interest in the use of terahertz (THz) spectroscopy for a range of areas including medical, security, communications, astronomy, and industrial applications such as quality, sensing, monitoring, and process control [1]–[3]. A topic of growing interest is the analysis of the THz spectra of chemical mixtures [4]–[21]. Usually, the aim of such an analysis is to identify what materials are present in the mixture and the relative abundances ("proportions") or concentrations of the mixture components. In fields such as remote sensing and microscopy, this process is often called (spectral) "unmixing" [22], [23]. We will use the same term in this article.

Manuscript received November 18, 2021; revised April 10, 2022; accepted April 20, 2022. Date of publication May 10, 2022; date of current version July 5, 2022. (*Corresponding author: Mark Berman.*)

Mark Berman is with the CSIRO Data61, Marsfield, NSW 2122, Australia (e-mail: mark.berman@data61.csiro.au).

Krunal Radhanpura and David Farrant are with the CSIRO Manufacturing, CSIRO, Lindfield, NSW 2070, Australia (e-mail: krunalkradhanpura@gmail.com; david.farrant@csiro.au).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TTHZ.2022.3172006.

Digital Object Identifier 10.1109/TTHZ.2022.3172006

The 18 THz articles analyzing mixtures referenced previously can be subdivided in several ways. First, only one of the articles analyzes spectra measured in reflection mode [19], while the remaining 17 articles analyze transmission mode spectra. Measuring samples in reflection mode is essential when the material is too thick and/or too absorbing to allow transmission of THz radiation, or when only one side of a material is physically accessible. Reflection mode spectra are often quite noisy due to the low reflectivity of many materials at THz frequencies.

A second subdivision is that 14 of the articles use time-domain spectroscopy (THz-TDS) to measure their spectra, while two use Fourier transform infrared spectrometers measurements [14], [15] and two use THz parametric generation [4], [5]. The last two of these articles combine THz spectroscopy with imaging, as does [6], to produce what are variously called spectroscopic images [4], [5], hyperspectral images [24], chemical images [25], and chemical maps [6]. None of the other 15 articles use spectra with any spatial structure. The fourth way to subdivide these articles is based on whether or not they use a database of spectra of *known* materials/chemicals to aid in their identification of *unknown* pure and mixed materials. In some disciplines, such a database is called a "spectral library," e.g., mineral exploration and remote sensing [26] and proteomics [27]. We will use this term in this article. Fourteen of the articles use a spectral library, while four do not; they carry out what is sometimes called "blind (hyperspectral) unmixing" [28]. The basic aim of blind unmixing is to find the "purest" materials in a hyperspectral image without reference to a spectral library. This is usually done with the aid of a linear mixture model.

In many of our THz spectroscopy applications, the aim is to identify and often map *automatically* materials of interest in a hyperspectral image, and to do so in reasonable time. As interesting as blind unmixing is as a research topic, its fundamental drawback in many *real-world* applications is that, once the purest materials have been identified, an expert in the spectroscopic aspects of the relevant application needs to *manually* identify what these materials are by examining their spectra.

There are two types of spectral library used in the 14 articles that employ them. In three of the articles, the library contains spectra of both pure materials and mixtures of them [18], [20], [21]. The spectral libraries used in the remaining 11 articles consist of the spectra of pure materials *only* and model the mixing of these. Seven of these articles use (either implicitly or explicitly) a linear mixture model to analyze their data [4]–[9], [14]. We will use an extension of the linear mixture model as the basis for our analysis; this is described in detail in Section V.

Of the other four articles using spectral libraries of pure materials, [10] and [11] use a microgenetic algorithm [29] to maximize a fairly complex "fitness" function. From our perspective, the algorithm has three drawbacks. First, it appears to take many iterations to converge. The authors show plots with good convergence after 1000 iterations, but there are no statistics on how long the process takes. Second, "random starting seeds might result in very different outputs," which is not desirable. Finally, the algorithm requires the allowed proportions in the mixture model to be *discrete*, e.g., multiples of 0.05; presumably, the finer the discrete resolution, the longer the algorithm will take to converge. In [12], absorption spectra are modeled as a combination of Lorentzian peaks and a low-order polynomial; we could not find the order of the polynomial used. Their method is applied to mixtures of lactose and theophylline. In [19], a complex procedure is used to distinguish the amplitude spectra (in reflection mode) of the two ternary mixtures (of explosives): RDX+TNT+HMX and RDX+TNT+PETN. Note that the two mixtures have two explosives in common. The method described is very focussed on this particular application. By contrast, the method that we propose in this article is a general purpose one.

Among the articles using libraries of *pure* spectra, all except [10] and [11] contain at most four materials. They fit *all* the materials in their libraries. In most applications, the number of materials in a mixture is small (2, 3, or possibly 4). However, libraries can be quite large. In [10] and [11], the number of materials in their libraries (N_l) are 22 and 28, respectively. In such a case, it is important to omit those materials in the sample which are *not* in the library (NIL). We will refer to such materials as "absent" materials. To some extent, the microgenetic algorithm used by [10] and [11] achieves this by discretizing the proportions (see aforementioned) so that the estimated proportions of absent materials equal 0 (hopefully), rather than 0.05 say.

In this article, we use an extension of a linear mixture model in combination with specific testing for the presence/absence of materials. In the statistical literature, this process is often called "subset selection" [30]. Our model and associated testing procedures also incorporate terms that address variability in relative humidity, materials in the dataset that are absent from the library, and spectra that show no distinctive absorption features (i.e., which appear to be low-frequency curves only). The unmixing methodology is applied to hyperspectral images, measured in reflection mode, consisting of pure and mixed pellets of different materials. The methodology presented in this article is an adaptation of methodology developed for mineral exploration applications, based on visible and infrared spectra [31]. This methodology, in particular, the extension of the linear mixture model and the use of subset selection, has not been published in the THz literature previously. In addition, what little automated analysis that has been published in the THz literature has been mostly confined to transmission spectra.

The article is structured as follows. Section II describes sample preparation and the experimental setup, while Section III introduces our spectral library and three test images that we will use. Preprocessing, parts of which are important in motivating our model, is described in Section IV. Our mixture model and associated training and testing procedures are outlined in Section V. These are applied to the test images in Section VI. Conclusions are drawn and outstanding issues discussed in Section VII.

II. Sample Preparation and Experimental Setup

A. Sample Preparation

Samples of D(-)fructose, D-(+)glucose, α-lactose monohydrate, D-(+)maltose monohydrate, sucrose, and L-tartaric acid with $> 99\%$ purity were acquired from Sigma-Aldrich. A spectrally neutral filler material, Polyblend (PB), was acquired from Micro Powders. These samples were prepared as pellets using a press (ICL's Port-A-Press) with approximately 42-MPa pressure to make pellets of 10-mm diameter. The thicknesses of pellets were 0.5–1.0 mm for transmission measurements and 4.0–4.5 mm for reflection measurements. For fructose, sucrose, and tartaric acid, the powders were ground finer using a mortar and pestle before preparing the pellets. This reduces scattering of the THz radiation, which has wavelengths in the range of 100–500 μm. A brass pellet was used as a reference for reflection configuration measurements.

B. Experimental Setup

THz spectroscopy measurements were performed using a pulsed time-domain system (Menlo Systems) based on fiber-coupled photoconductive antennas as emitter and detector [32], [33]. THz-TDS generates coherent broadband THz pulses by means of short pulses of excitation radiation. The time-resolved THz field amplitude with femtosecond (fs) resolution was obtained by using part of the excitation beam as a gating pulse and mapping the change in field amplitude by scanning the probe beam along the THz beam. The system can thus measure both phase and amplitude simultaneously. This allows calculation of the frequency-dependent complex refractive index, absorption coefficient, and other material properties in the THz frequency range.

In transmission configuration, the THz beam was collimated and focused onto the samples using two TPX lenses. The THz beam transmitted through the samples was also collimated and focussed using another two TPX lenses before impinging on the detector antenna.

In reflection configuration, the THz beam was collimated using a TPX lens and focused onto the samples using a parabolic mirror. The THz beam reflected from the samples was collimated using another parabolic mirror and focused on the detector antenna using a second TPX lens.

The measurements were performed with continuous purging using dry nitrogen gas to suppress water vapor absorption lines in the spectra. For reflection configuration measurements, the reflectance spectra were calculated from the ratio of sample to reference, where the reference measurement was a measurement using a brass pellet.

The time-domain traces (TDTs) were collected every 0.03333 ps over 50, 100, or 200-ps windows (depending on the sample and measurement configuration), and then Fourier transformed

to produce frequency spectra, resulting in a spectral resolution of 20, 10, or 5 GHz, respectively. Each spectrum was then averaged over 10, 20, or 50 s measurements over the area of each pixel.

In order to obtain hyperspectral images, the samples were scanned laterally using a *X–Y* stage. The total measurement time depends on the averaging time (see aforementioned) and on the total number of pixels, which itself depends on the physical size of the sample and the required spatial resolution. Discussion about these issues in relation to particular images is given in Section III-B. All the measurements presented in this article were taken with a 1-mm spatial resolution.

III. Spectral Library and Test Images

A. Spectral Library

We have examined the THz reflectance spectra of a large number of pure materials. Unfortunately, most of them have a low reflectance and unless their diagnostic absorption features are strong, they are hard to distinguish from noise. For illustrative purposes, in this article, we will demonstrate a variety of issues with a small library, consisting of five saccharides [fructose (F), glucose (G), lactose (L), maltose (M), and sucrose (S)], plus tartaric acid (TA). Three of these (L, M, and TA) have strong diagnostic absorption features, while the other three have weak absorption features. PB was originally in the library. However, due to its nature as a spectrally neutral filler material, its reflectance spectrum is particularly featureless, so we decided to exclude it from the library. Details of the sample preparation of all these materials are given in Section II-A.

The library has been created from six separate "training" images, each of size 27×27 pixels, and each containing a single pellet of the relevant material. We will call these "single-pellet" images. A seventh single-pellet brass reference (R) image (also of size 27×27) was also measured. All seven images were measured with 10-s averaging. For each of the seven single-pellet images, the "length" (i.e., no of measurement points) of each TDT was 3000, making a total measurement window of 100 ps. The relative humidity (RH) was approximately a very low 4.5% for each image.

The pellet boundaries are somewhat fuzzy due to the width of the illuminating beam. Effectively, pixels near the *true* boundary are actually mixtures of the holder and material in the pellet. However, the radius of each pellet in the seven images appears to be about 10 pixels. To minimize the effect of mixing caused by the beam width, we have built the library using spectra around the center (manually estimated) of each pellet. Specifically, we have taken the 25 reflectance spectra in the 5×5 window about the center of each pellet to build the library. The calculation of the reflectance spectra from the TDTs is described in Section IV. Having more than one spectrum per material gives us some idea of the variability of spectra of the same material, hopefully with minimal influence from the holder. We explain how we use the variability information in the training process in Section V.C.

B. Three Test Images

Three test images have been used to test our unmixing procedure against the library. All three images contain either six or eight pellets. We will refer to these as "multipellet" images, as opposed to the single-pellet images used to create our pure spectral library. Summary information about the three test images, as well as the training images, are given in Table I. Note that there is significant variability in the spatial dimensions, lengths, amount of averaging ("mode"), and RH in the four sets of images.

TABLE I
Summary of Analyzed Datasets

Data set	Single or Multipellet	Spatial Dimensions	Length	Mode	RH (%)
Pure	7 Single	27×27	3000	10s	4.5%
Pure	Multi	62×121	1504	10s	15%
L/TA	Multi	34×44	6000	20s	56%
L/M/TA	Multi	48×37	6000	50s	46%

The first test image consists of eight *pure* pellets, six containing the six materials in the library, a seventh brass reference pellet, and the eighth containing PB. This image contains many more pixels than the other images, and so it has the shortest averaging time and length. By comparison, the seven single-pellet images from which the spectral library has been obtained, discussed in Section III-A, are each much smaller, so we were able to almost double the length of each TDT.

Note that our library was originally extracted from this multipellet image, but we found that some of the pellets had low signals due to slight variations in alignment from pellet to pellet as a result of their being mounted on a polymer tape substrate. We then made the single-pellet measurements, where each pellet could be independently aligned for the maximum reflected signal.

The impact of the alignment issue is shown in Fig. 1(a)–(f). Each plot shows 25 reflection coefficient (RC) spectra (the square root of the reflectance spectra) taken from the centers of each of the single pellets (in red) and from the centers of the multipellets (in green). (Because most of the spectra have a low reflectance, it is often easier to see differences between RC spectra than between reflectance spectra.) See (1) and the steps leading to it for an explanation of how the RC is calculated. All plots are shown in the frequency range (0.25–2.00) THz, because outside this range, the reflection mode spectra are very noisy, due to a lower THz dynamic range. The means of the single-pellet and multipellet spectra are also shown (in black). In addition, in each plot, a transmission mode RC spectrum of the relevant material is shown in black. This is calculated from the refractive index, which is derived from the transmission mode measurement of the material [34], [35]. This spectrum is included because as well as being less noisy, it is used as a guide to where one hopes to observe significant absorption features in the noisier reflection mode spectra.

Before discussing each of the six plots, we make some general observations about all the plots. The first thing to note is that, in each plot, both sets of reflection mode spectra become noisier as the frequency increases, which is why we have excluded higher frequencies from each plot. The data are also noisy below 0.25 THz, so they have also been excluded. A second thing to note is that, for each material, the three datasets broadly differ in a number of ways. In particular, their low-frequency components

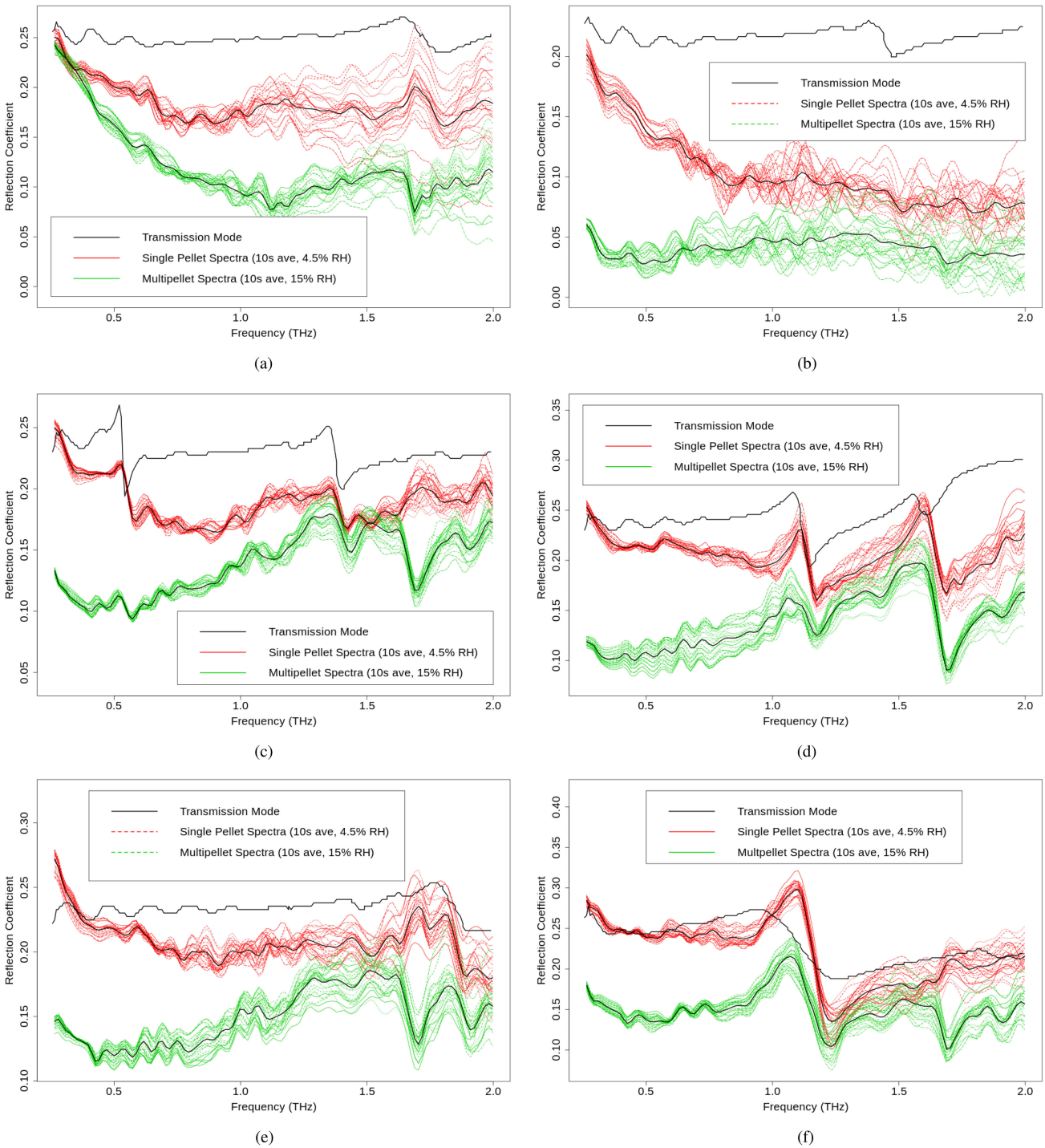


Fig. 1. RC spectra for six materials in (0.25–2.00) THz from three datasets (Transmission mode-black. Reflection mode: 25 single pellet (library) spectra—red; 25 multipellet (test) spectra—green, with corresponding mean values in black). See (1) and the steps leading to it for relevant calculations. (a) Fructose. (b) Glucose. (c) Lactose. (d) Maltose. (e) Sucrose. (f) Tartaric acid.

differ quite significantly. We have included a cubic term in our extended linear model (described in Section V-A) to account for this variation. Note also that the multipellet spectra (in green) of many of the materials show a strong water vapor absorption feature at approximately 1.7 THz, because their RH is 15%. However, this feature is usually not present in the single-pellet spectra (in red), because their RH is a much lower, 4.5%. This difference is also accounted for by our model. Another difference is that the multipellet spectra tend to have lower amplitude (and hence, noisier) than the single-pellet spectra. This is due to the alignment problem mentioned previously. The glucose spectra have particularly low amplitudes, with their single-pellet and multipellet means being 0.104 and 0.041, respectively. The single-pellet means of the other five materials vary between

0.184 and 0.220, while the multipellet means vary between 0.120 and 0.150.

We now turn to a more detailed discussion of the six plots in Fig. 1, and in particular where we can see strong absorption features consistently across the three datasets. Such features can be seen in lactose (at approximately 0.6 and 1.4 THz). Also, these features in the reflection mode spectra are weaker than they are in the transmission mode spectrum, especially for the multipellet spectra. There is a strong feature at approximately 1.2 THz in all three sets of maltose spectra. Similarly, there is a strong feature at 1.25 THz in all three sets of TA spectra.

The features are not as strong or consistent for the other three materials. Fructose has a weak feature at about 1.8 THz in the transmission mode and single-pellet spectra. However, it is not obvious in the multipellet spectra. Sucrose appears to have a weak feature just below 2.0 THz in all three datasets, but its shape is not consistent. Glucose has a weak feature at about 1.5 THz in transmission mode. However, this is not obvious in the two reflection mode datasets.

A summary image of the multipellet image of pure pellets is shown in Fig. 2(a). This image shows what we will call the "absolute maxima." These are the maximum at each pixel of the absolute value of the original TDT after it has been truncated to include the relevant parts of the data. (Some details about this automated truncation procedure are given in Section IV-A.) We mostly use a "heat" color map to display the maxima. The smallest values are shown in blue, while the largest values are shown in red, except for the (global) maximum value, which is shown in white; we will explain why in Section IV-C. Spectra that do not satisfy certain criteria that must be met by the truncation procedure are excluded from further analysis. These are shown in black. For this dataset, they largely correspond to the polymer holder. Abbreviated labels for each of the eight pellets are also included in the image. The following three things are worth noting:

1) the absolute maxima, together with the truncation procedure, provide a reasonable segmentation of the image into the eight pellets and holder;
2) there is a dark "halo" around the edges of many of the pellets due to the width of the beam of the THz source;
3) the (brass) reference is clearly brighter than the other pellets, as expected, due to its high reflectivity.

The second test image contains six pellets: three pure pellets containing lactose (L), tartaric acid (TA), and a brass reference (R), and three pellets containing mixtures of L and TA. The overall *weight* proportions of L of the mixed pellets are 0.25, 0.5, and 0.75, respectively. We will refer to this as the L/TA (mixtures) image. Its summary image is shown in Fig. 2(b), where the color map is the same as that used in Fig. 2(a). Its bottom row contains zeroes only and so it is shown in black. As in the pure multipellet image [see Fig. 2(a)], all the pellets show a dark blue "halo" (due to the width of the beam) and the reference pellet is much brighter than the other pellets. However, unlike in Fig. 2(a), the holder pixels have "passed" the truncation process. The reason for this will be discussed shortly.

The third test image also contains six pellets: four pure pellets containing lactose (L), maltose (M), tartaric acid (TA), and a brass reference (R), and two pellets containing mixtures of L, M, and TA in *equal weight* proportions. We will refer to this as the L/M/TA (mixtures) image. Its summary image is shown in Fig. 2(c). Its characteristics are similar to those of the L/TA mixtures image.

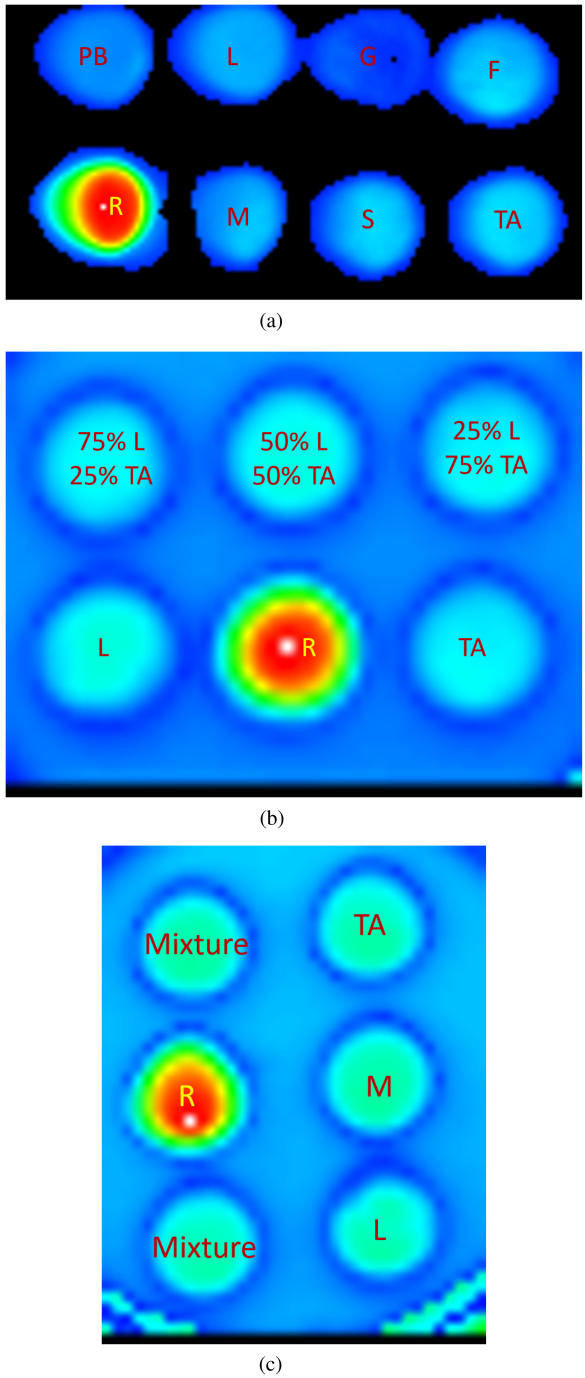


Fig. 2. Maxima of absolute time-domain traces (after contiguous subset selection) for the three multipellet images. (a) Pure pellets. (b) L/TA mixtures. (c) L/M/TA mixtures.

The second and third test images were measured using a horizontal mounting configuration for the pellets. The pellets were placed on a rigid polymer sheet, which was supported by a metal plate. To capture the full signal from this configuration

required a longer measurement window. It was then decided that all future TDTs would have a length of 6000. This greater length is part of the reason why the holder pixels in both these images pass the truncation process. Also, because the number of pixels in both images is relatively small, the averaging time was increased from 10 to 20 s.

IV. Preprocessing

Each spectrum in the four datasets has been subjected to the following preprocessing.

A. Truncating TDTs

The lengths of the four datasets vary between 1504 and 6000; see Table I. It is useful to have all the TDTs that are further analyzed of the same length because it is easier to then apply the same algorithms to all TDTs in all datasets. In particular, once the TDTs have been transformed into the Fourier domain, the frequency spacing within each spectrum will be the same. In addition, many of the longer TDTs within pellets contain the effects of the holder, which is undesirable. These effects need to be omitted by truncating each TDT to at most 1504 *contiguous* observations. In fact, we will reduce them to 1501 observations, where possible.

We have developed an automated truncation algorithm, to deal with a variety of situations. Because it is moderately complicated and is not the main focus of this article, we omit the details. However, we give the flavor of the algorithm with two examples drawn from the L/TA mixture dataset. Fig. 3(a) and (b) shows a TDT near the middle of a pellet and one in the holder, respectively. The red peak in Fig. 3(a) represents reflection from the front surface of the pellet, while in Fig. 3(b), the red peaks represent reflection from the front of the pellet and the front of the polymer holder (both showing low amplitudes due to the width of the beam covering only the edge of the pellet). The black peaks following the red peaks are due to internal reflection from the pellets, as well as reflection from the polymer holder and the metal plate. In each figure, the selected contiguous section of the TDT (of length 1501) is shown in red. In some cases, it is not possible to find a contiguous section of length (at least) 1501 meeting the requirements of the algorithm. These TDTs, such as the holder spectra in Fig. 2(a), are omitted from further processing.

B. Converting TDTs to Amplitudes

Before transforming each TDT (of length 1501) to the Fourier domain, we first apply a Hanning window to them to make the data "periodic" [36], and then, "pad" them with additional zeroes to a length of 2048 ($= 2^{11}$). This is a common technique for interpolating Fourier data. The amplitude spectra obtained after padding and taking the Fourier transform of each TDT produce data at 1025 frequencies over the range [0.00–15.03] THz. As mentioned previously, the observations outside (0.25–2.00) THz are too noisy for practical use. So, we only use the $N_f = 119$ frequencies inside this interval in our analysis.

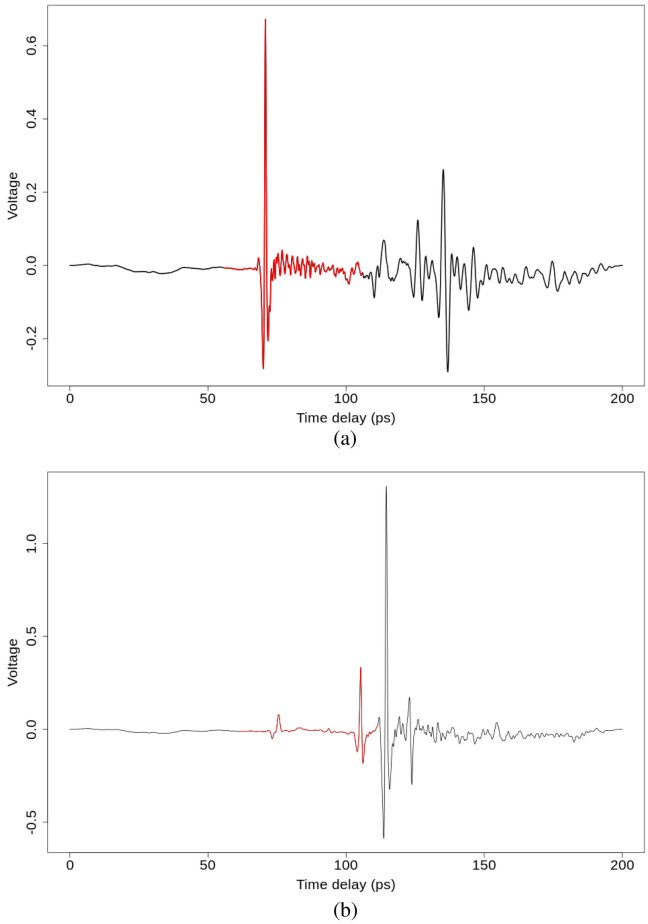


Fig. 3. Two TDTs from the L/TA mixture dataset. Contiguous truncated sections (used in subsequent processing) shown in red. (a) Pixel (22, 8): Near the middle of a pellet. (b) Pixel (28, 18): In the holder.

C. Converting Amplitudes to Reflection Coefficients

Let $\mathbf{E}_{Si}, i = \ldots, N$ denote the amplitude vector of the sample material (of length 119) at pixel i out of N pixels in an image (after masking out those that are unsuitable for further processing; see above), and let $\mathbf{E}_R$ denote the amplitude vector of the reference (assumed to be the same for all pixels in the image). Then, the RC at pixel i can be estimated by [35]

$$\mathbf{RC}_i = \mathbf{E}_{Si}/\mathbf{E}_R. \tag{1}$$

The reference is designed to remove features that are common in all the spectra prior to further processing and analysis, most notably low-frequency components and water vapor, which varies with RH.

For the training data, the reference spectrum used is the *geometric* mean of the nine driest brass spectra (those with $RH < 10\%$) from an experiment we call the water vapor experiment, where the measurement volume was purged with dry nitrogen gas. This reference is used because it is of high quality and very dry. The water vapor experiment is described in Section V-B. In the three test images, the reference spectrum that we have used is the geometric mean of the nine amplitude spectra centered on the pixel with the largest absolute maximum TDT

after the contiguous section has been selected; see Section IV-A. For each of the three datasets, this pixel is shown in white in Fig. 2(a)–(c), respectively. We will explain why we have used the geometric mean in Section V-B.

V. Extended Linear Mixture Model, and Associated Training and Testing Procedures

The extended linear mixture model and associated procedures are based on a set of algorithms and associated software called The Spectral Assistant (TSA) [31]. They have been developed for mineral exploration applications, originally in the shortwave (SWIR) infrared region (1300–2500) nm (with a library of 60 materials, mostly minerals), and more recently in the visible and near infrared region (450–1100) nm (with a library of 17 materials) and the thermal infrared region (6000–14000) nm (with a library of 81 materials). TSA is a significant part of a commercial package called The Spectral Geologist (TSG),[1] which is mainly applied to (typically) tens of thousands of spectra measured in individual drill cores with an instrument called the HyLogger [37].

This section is structured as follows. In Section V-A, we outline the extended linear model. The model includes a water vapor term. Section V-B describes an experiment we carried out that enabled us to model the water vapor term. In Section V-C, we explain how the library is trained within the constraints of the model. We explain the algorithm for unmixing spectra in Section V-D. Two related issues arising from this are overfitting and how to estimate the number of materials in the mixture. These issues are addressed in Section V-E. The following three miscellaneous issues are briefly discussed in the final three subsections: identifying spectra with no significant absorption features; spectra that are NIL; and metal spectra.

A. Extended Linear Model

Let

$$\mathbf{Y}_i \equiv \log(\mathbf{R}\mathbf{C}_i). \tag{2}$$

(We will explain why we have used the logarithm ("log") of the RC spectrum in Section V-B.) Then, the extended linear model that we will use is

$$\mathbf{Y}_i = \sum_{j=1}^{N_m} \alpha_{ij}\boldsymbol{\mu}_j + \sum_{k=1}^{N_b} \beta_{ik}\mathbf{B}_k + \epsilon_i. \tag{3}$$

The first term in (3) ($\sum_{j=1}^{N_m} \alpha_{ij}\boldsymbol{\mu}_j$), which is the usual linear mixture model and which we will sometimes call the "foreground", represents the mixture of the N_m chosen materials (out of $N_l = 6$). $N_m = 1$ corresponds to *pure* spectra. $\boldsymbol{\mu}_j$ is the "typical" spectrum of the material j in the mixture. We will describe how it is estimated in Section V-C. α_{ij} represents, in a semiquantitative way, the amount of the material j in the mixture. We do not constrain it to be a proportion so that we can account for brightness variations in different spectra of the same material. However, we do constrain it to be *positive*.

[1][Online]. Available: https://research.csiro.au/thespectralgeologist/

The second term ($\sum_{k=1}^{N_b} \beta_{ik}\mathbf{B}_k$) represents "background" terms, which are of no or lesser interest. Unlike the coefficients in the foreground term, there are *no* constraints on β_{ik}. The most obvious parts of the spectra that can be represented as part of the background are the low-frequency components of each spectrum. Here, we represent those components as a *cubic* function, which itself can be represented as the sum of four terms. However, we have found it necessary to include an additional term to better model the effect of water vapor. This will be discussed in Section V-B. The addition of water vapor means that for the spectra analyzed in this article $N_b = 5$.

The last part of (3) is the error term (ϵ_i). This consists of both instrumental noise and "natural" variation of spectra with the same composition. Note that the spectra in Fig. 1(a)–(f) generally become noisier with increasing frequency. So, it is natural to downweight the observations at higher frequencies. Also, because spectra are curves, they are highly correlated locally. We can account for both noise variability and local correlation in the spectra by an error covariance matrix. In this article, we will assume normally distributed errors with zero mean and a *common* error covariance matrix. We show how we estimate this matrix in Section V-C.

B. Water Vapor

Water vapor is a significant issue in the analysis of THz spectra (see [38]), which discusses the physics of water vapor at THz frequencies in detail. We have carried out an experiment that provides us with the rationale for incorporating a water vapor term in our model and a means of estimating it. It is summarized here.

Multiple TDTs were measured for the following materials: brass, fructose, lactose, maltose, and TA. For each material, the RH was adjusted (and measured) before a *single* TDT was measured, by varying the amount of purging of the system using dry nitrogen gas. The TDTs were then converted into amplitude spectra, as described in Section IV-B. For brass, 46 TDTs were measured with RH in the range [8.3–42.6] %. There were between 39 and 44 TDTs of each of the other four materials measured with RH in the range [6.9–30.9] %.

For each material, we then removed the low-frequency variation in the amplitude spectra by dividing those spectra by the spectrum with the smallest measured RH for that material. This enables us to better observe the effects of water vapor. We call these the *relative* amplitude spectra. These are shown for the brass (reference) spectra in Fig. 4.

This plot raises a number of issues. First, we see water vapor features in a number of frequency regions. These occur at well-known frequencies [39]. Second, are the depths of the features a monotonically increasing function of RH, and if so, can we parametrize this relationship? To investigate monotonicity, in Fig. 5, we plot the relative amplitudes at five "key" frequencies: 0.57, 1.11, 1.44, 1.69, and 2.23 THz (the last of these is not shown in Fig. 4).

Each of the brass relative amplitude sequences do appear to be approximately monotonically decreasing functions of RH. There are two major exceptions, the first being very low RH values

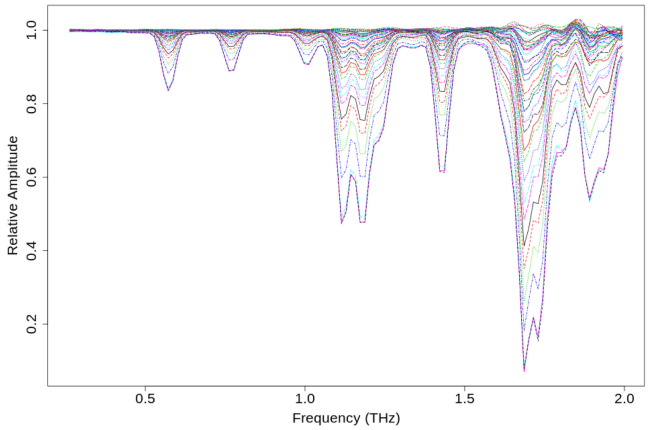


Fig. 4. 45 brass relative amplitude spectra measured in the RH range [8.3–42.6]%, showing prominent water vapor absorption lines.

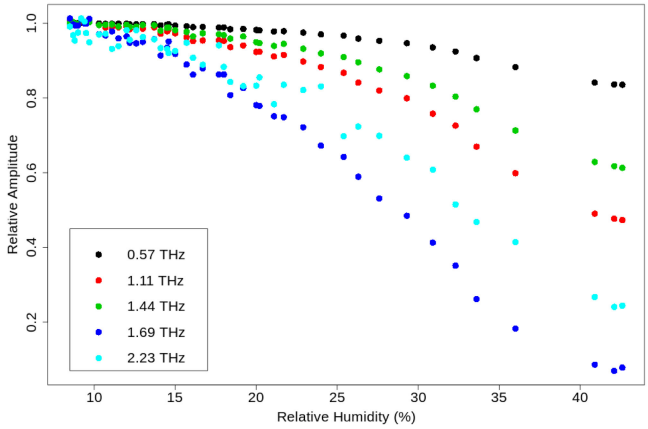


Fig. 5. 45 brass relative amplitudes versus RH at five key frequencies.

where *strict* monotonicity does not apply, probably either due to small inaccuracies in the RH measurements or due to noise in the spectra . (These are also consistent with the relative amplitudes being occasionally above 1 *between* the water vapor features in Fig. 4.) The second major exception is when RH $> 40\%$. Here, the values appear to flatten out and possibly even start to increase at the higher frequencies. We will discuss this issue further later.

The approximate monotonic relationship between the brass relative amplitudes and RH can be parametrized; it turns out that the log of the relative amplitudes is approximately a linear function of RH. Fig. 6 shows the log relative amplitudes ("Y") at the key frequencies for four RH values approximately equally spaced between the minimum recorded RH (8.3%) and the maximum recorded RH (42.6%) versus the log relative amplitudes at the key frequencies for RH $= 42.6\%$ ("X"). We also plot the least-squares (LS) line when fitting each of the four Y's against X (without an intercept). The approximate log-linear relationship is clearly demonstrated. It is because of this relationship that we have taken logs in (2) and why we have used *geometric* means of neighboring brass RC spectra to obtain reference spectra in various datasets.

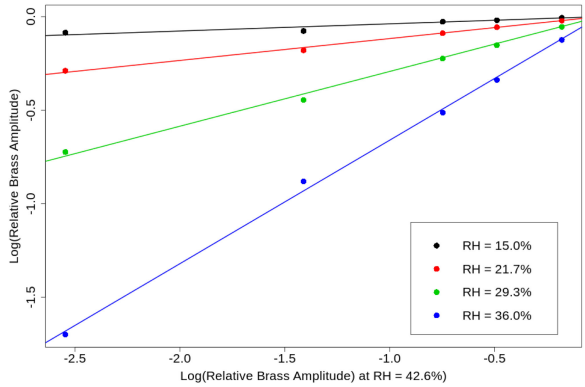


Fig. 6. Log relative amplitudes for four RH values versus log relative amplitudes for RH $= 42.6\%$.

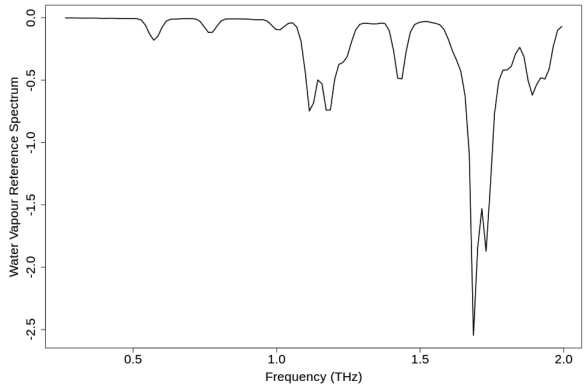


Fig. 7. Water vapor reference spectrum. This is the log of the deepest spectrum in Fig. 4.

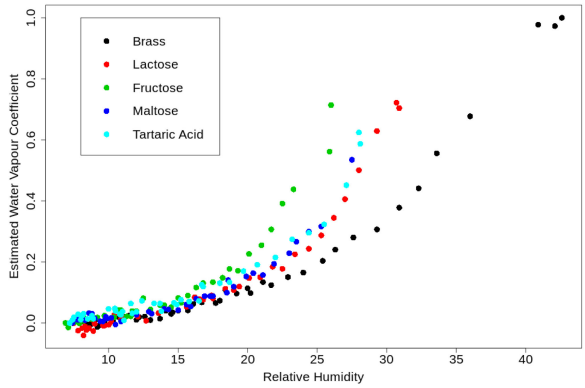


Fig. 8. Water vapor coefficients in LS fits of log amplitude spectra of five materials versus RH.

Based on the aforementioned results, we have chosen the brass log relative amplitude spectrum at RH $= 42.6\%$ as the (background) water vapor term in (3). It is shown in Fig. 7.

A fundamental question is how well does the (brass) water vapor term model the water vapor components of the RC spectra of other materials? To answer this question, we carried out an LS fit of the log amplitude spectra of brass, fructose, lactose, maltose, and TA at all RH values against their *driest* versions and the water vapor term, i.e., the black curve in Fig. 7. Fig. 8

shows the coefficient of the water vapor term for each material versus RH. We observe that all five sequences are approximately monotonic functions of RH. However, they do not increase at the same rate. Brass is slowest and fructose is fastest, with the other three materials in between, rising at about the same rate. This reflects the fact that different materials absorb moisture at different rates. It also highlights the fact that dividing a sample amplitude spectrum by a brass (or other) reference [see (1)] is insufficient to model the effect of water vapor on its own. This is why we have included a water vapor term in the model. Its coefficient [see (3)] allows the water vapor depths of a material to be a *multiple* of those of the reference [rather than being constrained to be equal to it, which is implied by (1)].

Note that the water vapor term was measured at RH $= 42.6\%$, while two of our test datasets were measured at higher RH values (i.e., 56% and 46%); see Table I. The water vapor term *by itself* does not adequately model water vapor behavior at higher RH values. However, as we shall see in Section VI, the water vapor term together with division of the amplitude spectra of the samples by the reference spectrum (1) does appear to do an adequate job of modeling that behavior. So effectively, we need two terms in our model to account adequately for the behavior of water vapor at higher RH values.

We suspect that this is related to the fact that water vapor is a mixture of two components. If the nuclear spins of both hydrogen atoms in a water molecule are in the same direction, it is called ortho-H_2O. If they are in different directions, it is called para-H_2O. In related experiments to ours, the authors in [39] and [40] have demonstrated how these two components impact water vapor frequency depths as RH varies.

C. Training the Library

This section summarizes the process of training the library. More mathematical details can be found in [31, Sec. V].

Because the spectra in our library are (notionally) pure, they satisfy (3) with $N_m = 1$. However, for the training process, it will be useful to include additional notation. Let $\mathbf{Y}_{ij}, i = 1, \ldots, n_j, j = 1, \ldots, N_l$ denote the ith log RC spectrum in class (material) j, where $n_j (= 25)$ is the number of samples in class j. Then, for library spectra in class j, (3) becomes

$$\mathbf{Y}_{ij} = \alpha_{ij}\boldsymbol{\mu}_j + \sum_{k=1}^{N_b} \beta_{ijk}\mathbf{B}_k + \epsilon_{ij} \tag{4}$$

where $\boldsymbol{\mu}_j$ is the "typical" spectrum in the class j (as yet unknown), α_{ij} is a positive scale parameter, $\mathbf{B}_k, k = 1, \ldots, N_b$ $(= 5)$ are the background functions (all known), β_{ijk} are their (unknown) coefficients, and ϵ_{ij} is an error term. The LS solution to (4) minimizes (in matrix notation)

$$\sum_{i=1}^{n_j} \left(\mathbf{Y}_{ij} - \alpha_{ij}\boldsymbol{\mu}_j - \sum_{k=1}^{N_b} \beta_{ijk}\mathbf{B}_k \right)^T \times \left(\mathbf{Y}_{ij} - \alpha_{ij}\boldsymbol{\mu}_j - \sum_{k=1}^{N_b} \beta_{ijk}\mathbf{B}_k \right). \tag{5}$$

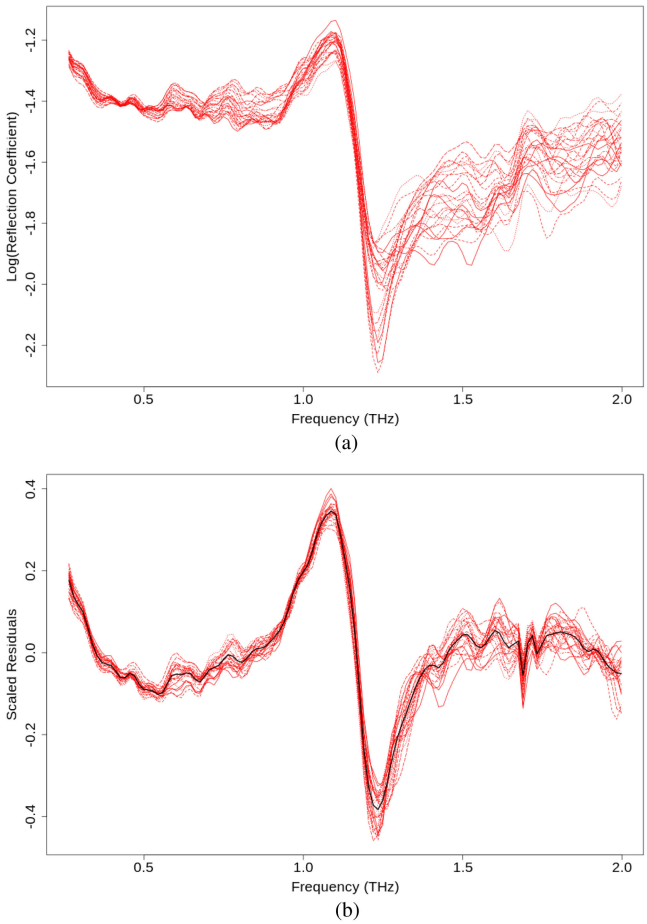


Fig. 9. 25 TA log RC spectra from the library and their scaled residuals, given by (7). (a) 25 TA log RC spectra. (b) 25 TA scaled residuals.

The minimizer of (5) is *not* unique! However, this makes no difference to the unmixing procedures discussed in the following subsections. Further discussion about this issue can be found in [31, Sec. V-A]. Probably the simplest solution is as follows. First, carry out an LS regression of *each* $\mathbf{Y}_{ij}$ on the N_b background functions only. Let

$$\mathbf{R}_{ij} \equiv \mathbf{Y}_{ij} - \sum_{k=1}^{N_b} \hat{\beta}_{ijk}\mathbf{B}_k \tag{6}$$

denote the residual of this fit, where $\hat{\beta}_{ijk}$ is the LS estimate of β_{ijk}. Then, substitute the $\mathbf{R}_{ij}$'s into (5) and minimize it with respect to α_{ij} and $\boldsymbol{\mu}_j$. The solution $(\hat{\boldsymbol{\mu}}_j)$ is just the first principal component (PC) of the $\mathbf{R}_{ij}$'s [41, Sec. 2.1]. Note that the term $\alpha_{ij}\boldsymbol{\mu}_j$ in (5) has a scale indeterminacy between the two terms. Unfortunately, the standard PC scaling used to resolve this makes the amplitude of $\hat{\boldsymbol{\mu}}_j$ much larger than the typical amplitudes of the spectra. We use a scaling that overcomes this problem; see [31, Sec. V.A] for further details. As an example, Fig. 9(a) shows the 25 TA log RC spectra in the library. These are the logs of the red spectra in Fig. 1(f). The "typical" TA spectrum, $\hat{\boldsymbol{\mu}}_j$, is the black curve in Fig. 9(b).

We see from (5) and (6) that each $\mathbf{R}_{ij}$ is an estimate of $\alpha_{ij}\boldsymbol{\mu}_j$. Then, if $\hat{\alpha}_{ij}, j = 1, \ldots, n_j$ are the values minimizing (5), the *scaled* residuals

$$\hat{\gamma}_{ij} = \mathbf{R}_{ij}/\hat{\alpha}_{ij} = \left(\mathbf{Y}_{ij} - \sum_{k=1}^{N_b} \hat{\beta}_{ijk}\mathbf{B}_k\right)/\hat{\alpha}_{ij} \tag{7}$$

are all estimates of $\boldsymbol{\mu}_j$. The red spectra in Fig. 9(b) are the scaled residuals derived from the log RC spectra in Fig. 9(a). Note how the background removal and scaling has considerably reduced the variation among the training spectra. Note also how, broadly speaking, the variation in the scaled residuals tends to increase with frequency. In addition, because the spectra are curves, they are highly correlated locally. The way to deal with these two issues is by first calculating an average error or noise covariance. Let

$$\hat{\delta}_{ij} = \hat{\gamma}_{ij} - \hat{\boldsymbol{\mu}}_j, i = 1, \ldots, n_j; j = 1, \ldots N_l (= 6) \tag{8}$$

denote the "final" residuals about the estimated "typical" spectrum in each class, $\hat{\boldsymbol{\mu}}_j$, which minimizes (5). The (estimated) error covariance matrix, denoted by $\hat{\boldsymbol{\Sigma}}_e$, is just the covariance matrix of the final residuals derived from *all* $150 (= 6 \times 25)$ samples in the library; see [31, eq. (9)].

We will see in Section V-D that we need to use the inverse of $\hat{\boldsymbol{\Sigma}}_e$ in the unmixing process. Unfortunately, background removal causes it to be "singular" (i.e., noninvertible). In fact its rank is $N_f - N_b (= 114)$. There are at least two ways of overcoming this. We will use both. The first way is to omit $N_b (= 5)$ frequencies (after background removal). We typically omit about half each of the lowest and highest frequencies. This has minimal impact on plots. For the data analyzed in this article, we omit the first two and last three frequencies.

The second way is to "ridge" Σ_e, i.e., add a small multiple of the identity matrix [42]. The effect is to add this quantity to each of its eigenvalues, making them all positive.

We also add a second matrix to Σ_e, which is designed to take advantage of the fact that the data are curves. This is a modification of an approach called penalized discriminant analysis (PDA) [43].

The ridge and PDA matrices are scaled so that they have the same total variance (i.e., trace) as Σ_e. The final "regularized" estimated error covariance matrix (denoted by Σ) is a linear combination of the three matrices with weights π_R, π_P, and $1 - \pi_R - \pi_P$, respectively. In all the examples shown in this article, $\pi_R = \pi_P = 0.01$ so that the contribution of Σ_e to Σ is 98% of the total.

Further details of the calculation of Σ are given in [31, Sec. V-B].

D. Unmixing Spectra

Assuming that there are $N_m (\leq N_l)$ materials in pixel i, we propose to find the best fitting N_m materials by minimizing the Mahalanobis distance (MD)

$$\text{MD}_i = \epsilon_i^\top \Sigma^{-1} \epsilon_i \tag{9}$$

where Σ is the estimated error covariance matrix described previously and

$$\epsilon_i = \mathbf{Y}_i - \sum_{j=1}^{N_m} \alpha_{ij}\hat{\boldsymbol{\mu}}_j - \sum_{k=1}^{N_b} \beta_{ik}\mathbf{B}_k \tag{10}$$

where $\hat{\boldsymbol{\mu}}_j$ is the typical spectrum of the material j in the subset. For a given subset of size N_m, (9) is mimimized with respect to $\alpha_{ij}, j = 1, \ldots, N_m$ (which are constrained to be nonnegative) and $\beta_{ik}, k = 1, \ldots, N_b$ (whose values are unconstrained). We need to do this for all subsets of size N_m and choose the one minimizing (9). We will discuss implementation issues shortly. Finally, we need to choose the value of N_m, which is most consistent with the data. This is discussed in Section V-E.

Of the 18 THz articles analyzing mixtures reviewed in Section I, seven implicitly or explicitly use (3) but without any background terms [4]–[9], [14]. Six of the articles place no constraints on the coefficients and use ordinary LS regression, while the seventh [5] does constrain the coefficients to be nonnegative and uses a method called nonnegative LS (NNLS) [44] to do this. None of these articles use any subset selection. We shall investigate this issue shortly.

LS or NNLS minimize Euclidean distance (ED), i.e., with Σ replaced by the identity matrix in (9). We have used MD because it has *minimum error rate* in the following special case:

1) $N_m = 1$;
2) the errors are Gaussian with a *common* and *known* within-class covariance matrix;
3) each of the N_l classes is equally likely.

Another interpretation is that MD transforms the data into a space that *maximally separates* the classes/materials. This should also improve the chances of unmixing the data into their "true" classes. This has been demonstrated empirically to be true in the minerals context (see [31, Tables II and III]), where MD consistently outperforms ED.

We now turn to implementation issues. The first thing to note is that MD can be converted to ED by a suitable linear transform so that (9) can be represented as a residual sum of squares (RSS), in our case as the sum of 114 squared residuals. The second thing to note is that the RSS can be decomposed into two parts. This is done using an *extension* of a technique called canonical variate (CV) analysis [45, Sec. 3.9.2], which is just a compressed version of the widely used classification technique called linear discriminant analysis [46, Sec. 4.3]. In our *extended* CV transformation, the dimension of the data is reduced from 114 to $N_l + N_b = 11$. We can then write

$$\text{MD}_i = \text{RSS}_{i,\text{CV}} + \text{RSS}_{i,O} \tag{11}$$

where $\text{RSS}_{i,\text{CV}}$ is RSS in 11-D CV space and $\text{RSS}_{i,O}$ is RSS in the 103-D space orthogonal to it. The important property of this decomposition is that, while $\text{RSS}_{i,\text{CV}}$ depends on the particular subset whose MD is being calculated, $\text{RSS}_{i,O}$ does *not*. Hence, $\text{RSS}_{i,O}$ only needs to be calculated once for each pixel, while the calculation of $\text{RSS}_{i,\text{CV}}$ is much faster for each subset than it is for MD_i, because it is the sum of 11 terms rather than of 114 terms. Mathematical details about the decomposition can be found in [31, Sec. VI].

Having converted minimization of (9) into a (lower dimensional) RSS/LS minimization problem, we can utilize fast subset selection procedures described in [30]. Fortran software implementing these procedures are currently freely available at http://wp.csiro.au/alanmiller/ and http://jblevins.org/mirror/amiller/, and in the R package "leaps" [47]. In the implementation of TSA found in TSG, for each of $N_m = 1, 2, 3$ and sometimes 4, we find the 20, 30, or 40 best fits, and among these find the best fit (if it exists), for which the estimated values $\hat{\alpha}_{ij}, j = 1, \ldots, N_m$ are all *nonnegative*. However, in the version applied to the data analyzed in this article, we examine *all* possible combinations satisfying these nonnegativity constraints. It is not much slower; see [48]. It has been implemented in the R package "groupsubsetselection" [49].

For $N_m > 1$, it is more meaningful to transform the $\hat{\alpha}_{ij}$'s into proportions for reporting purposes

$$\hat{p}_{ij} = \hat{\alpha}_{ij} / \sum_{k=1}^{N_m} \hat{\alpha}_{ik}, j = 1, \ldots, N_m. \tag{12}$$

This scaling makes the procedure invariant to gain effects, while the inclusion of a constant term among the background functions (as part of the cubic function) makes the procedure invariant to offset effects.

E. Estimating the Number of Materials in Each Pixel

Before proceeding, it will be useful to divide MD/RSS by the number of frequencies used in each pixel. This is usually called mean squared error (MSE)

$$\text{MSE}_i = \text{MD}_i / (N_f - N_b). \tag{13}$$

The value of this statistic is smaller and more interpretable. Its value is also less sensitive to small changes in the number of frequencies used.

Before discussing ways of estimating the number of materials in each pixel, we show why it is necessary, with the aid of the pure multipellet image; see Fig. 2(a). We fit all six materials in the library *without any subset selection*, i.e., we minimize (9) with $N_m = 6$ and the coefficients $\alpha_{ij}, j = 1, \ldots, 6$ constrained to be nonnegative. Although this can be converted into an NNLS problem, it is easier to use quadratic programming methods [50, ch. 16].

When all pixels are pure, it is possible to show their classifications in a *single* image. However, it is difficult to show the results of mixtures (and in particular, the proportions of each material in each pixel) in a single image. Therefore, in Fig. 10, we show separate proportion maps [using (12)] for each of the six library materials. We also show the MSE map; we will discuss it shortly. We also show the "heat" color scale that we use for both the proportion and MSE maps, together with the scale used for the proportion maps: dark blue, light blue, green, yellow, and red correspond approximately to the proportions 0, 0.25, 0.5, 0.75, and 1, respectively. Excluded pixels are shown in black. These pixels either did not satisfy the requirements of the truncation procedure (see Section IV-A) or $\hat{\alpha}_{ij} = 0, j = 1, \ldots, 6$, in which case (12) is undefined.

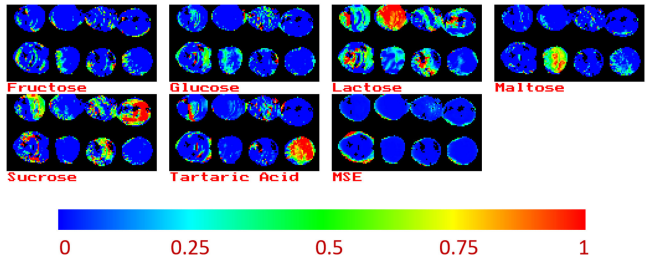


Fig. 10. Proportion and MSE maps for the pure multipellet image [Fig. 2(a)] with no subset selection, together with an annotated color scale.

Of the materials with strong features, the L and TA pellets are mostly red (i.e., pure or almost pure), with significant green patches (mainly around the mixed edges). Most of the M pixels are green, except for some red and yellow pixels near the pellet center. On the other hand, many of the pellets with weaker features show significant amounts of these materials (especially L). The pellets of materials whose RC spectra have weaker features are poorly identified. The worst example is the F pellet, which is mostly identified as S. This figure demonstrates several needs to reclassify those parts of the L, M, and TA pellets currently identified as mixed as pure, and to reclassify the F, G, PB, and S pellets as either "NIL" or what will call "aspectral." It will also be useful to classify the brass reference pellet as "metal." These issues are addressed over the next few subsections.

Before doing so, we discuss the MSE image in Fig. 10. Apart from the excluded pixels, a heat color map has also been applied to a linearly stretched version. Specifically, the stretch is between 0 and the upper 1% point of the distribution of (included) MSE values. Any values above the upper value are shown in red. We do not stretch it between 0 and the maximum MSE value, because there are usually some very large RSE values and stretching to this upper limit will make a few pixels red, with the rest shown in dark blue. Note that most of the pixels in this image that are not black or dark blue occur around the (mixed pixel) edge.

There are several ways of estimating the number of materials at each pixel. We have found the following method reasonably reliable. It is based on comparing "consecutive" models, i.e., containing N_m and $N_m + 1$ materials, respectively. In what follows, we will omit the subscript i for notational simplicity. Call the best fitting mixture of N_m materials "model N_m," and let MSE_{N_m} denote its MSE. Our estimate is based on

$$\rho_{N_m, N_m+1} = \text{MSE}_{N_m} / \text{MSE}_{N_m+1}. \tag{14}$$

The model $N_m + 1$ is preferred to model N_m if

$$\rho_{N_m, N_m+1} > \rho_0 \tag{15}$$

where ρ_0 is chosen according to some criterion. For each pixel, ρ_{N_m, N_m+1} is calculated for $N_m = 1, \ldots, N_{\max}$. In this article, we will use $N_{\max} = 2$. The estimated number of materials in the mixture is the last value of $N_m + 1$ satisfying (15). If this inequality is never satisfied, the estimated number of materials is 1.

There is statistical theory suggesting a suitable value of ρ_0. However, it is based on the assumption that the errors in the

training and test data have the same distribution. Unfortunately, this does not apply here because of slight variations in pellet alignments causing differences in the pure training and test spectra; see Fig. 1(a)–(f). Empirically, we have found that $\rho_0 = 1.125$ works reasonably well. This is the value in all the examples that we will analyze in Section VI.

F. Aspectral Pixels

"Aspectral" pixels are those that do not have detectable absorption features, i.e., they are well fitted by the background terms in (3) alone, or at least including the best fitting pure material does not significantly improve the fit. The aspectral test is just (15) with $N_m = 0$, i.e., only the background terms are fitted. We use the same value of ρ_0 (1.125) as aforementioned. This enables the easy integration of the aspectral test into the estimation of the number of aforementioned endmembers. All that needs to be done is to calculate ρ_{N_m,N_m+1} for $N_m = 0, \ldots, N_{\max}$. The estimated number of materials in the mixture is still the last value of $N_m + 1$ satisfying (15). However, if this inequality is never satisfied, the estimated number of materials is now 0, i.e., the pixel is classified as aspectral.

Although we have not seen the term "aspectral" in the literature, it is in fact used in the version of TSA contained in TSG.

G. Spectra That are NIL

NIL spectra are conceptually different from aspectral spectra. The latter have no distinctive absorption features (relative to the noise), while the former have features but their frequencies are significantly different from those in the library or those of water vapor. Typically, they do not fit as well and have a higher MSE than those spectra that are in the library or are well fitted by the background terms only. The NIL concept is needed in this article to mask out holder pixels in the two test images containing mixed pellets. (Most of the holder pixels in the pure pellet test image are masked out because they fail the truncation test; see Section IV-A.) In the examples analyzed in this article, a pixel is masked out if $\text{MSE}_3 > 1.3$.

H. Metal Spectra

As we can see in the absolute maxima images in Fig. 2(a)–(c), the metal pellets are significantly brighter than the other objects in those images. So, for all three images, we classify a pixel as "metal" if its absolute maximum is greater than 0.8. This gives a reasonably good separation between metal pellets and the other objects. The test is carried out *before* the mixture model (3) is fitted.

VI. Application of the Methodology to the Test Images

Fig. 11(a)–(c) shows the proportion and MSE maps (based on the theory and associated procedures in Section V) for the three test images, respectively. The same color scales that have been used in Fig. 10 have also been used here. Pixels (shown in black) have been masked out for one of the following reasons:

1) the TDT consists entirely of zeroes;
2) criteria to pass the truncation procedure have not been met (see Section IV-A);
3) if no *pure* material (with positive weights) can be fitted;
4) if $\text{MSE}_3 > 1.3$ (see Section V-G).

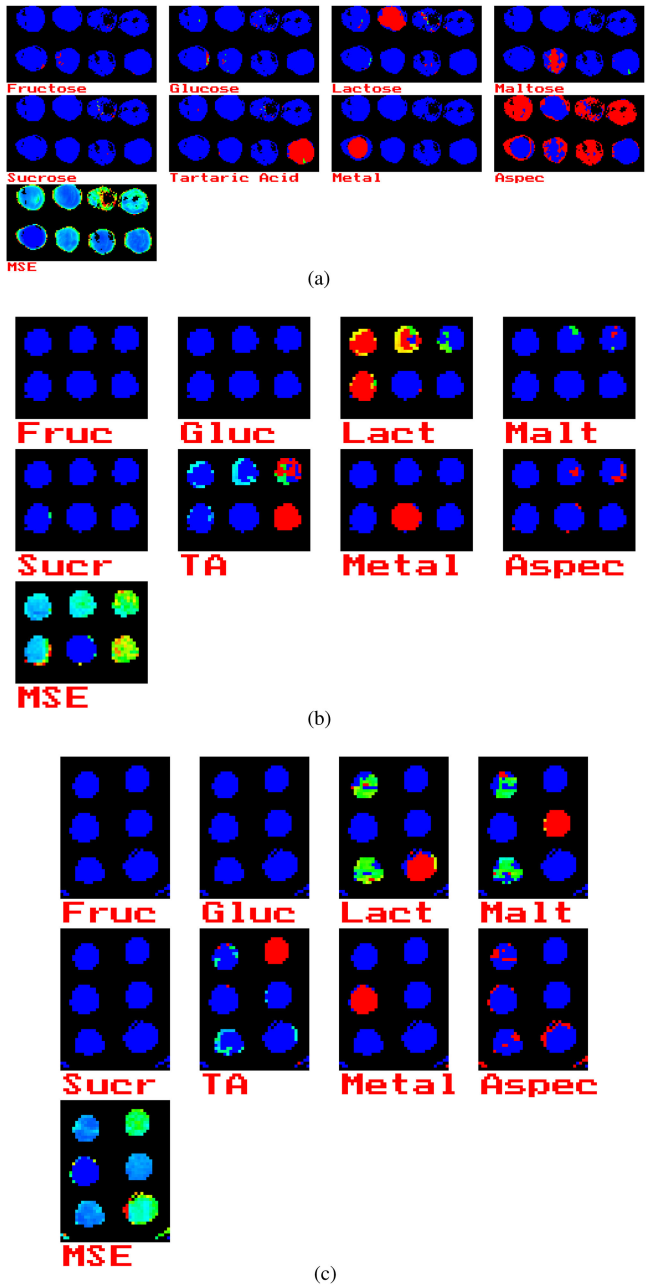


Fig. 11. Proportion and MSE maps for three multipellet images after subset selection. The color scale used in Fig. 10 has also been used here. (a) Pure pellets. (b) L/TA mixtures. (c) L/M/TA mixtures.

Note also that, for all three MSE images, the values are usually higher in pixels near the pellet edges. This is due to mixing with light from the holder, which is not in the library.

The proportion maps for the pure pellet image in Fig. 11(a) are a considerable improvement on those in Fig. 10 in terms of accuracy and interpretability. Almost all pixels that have not been masked out are either dark blue (0) or red (1), which is

what one would expect for pure pixels. In particular, the L, TA, and brass (R) pellets are almost entirely red in the appropriate proportion map. The central section of the M pellet is correctly classified, while the remaining parts are classified as aspectral. The F, G, PB, and S pellets are also almost entirely classified as aspectral, as are many pixels near the edges of the L and R pellets.

In the two images containing some mixed pellets [see Fig. 11(b) and (c)], the pure L, M, TA, and R pellets are almost entirely correctly classified; some pixels near edges are classified as being mixed. The mixing of the materials in mixed pellets is seldom uniform, so we do not expect the colors to be the same throughout such pellets. In Fig. 11(b), the pellets that are 75% L and 75% TA, respectively, are indeed dominated by these two materials, with smaller amounts of the minor (25%) material present. However, in the pellet with equal (50%) amounts of L and TA, L appears to dominate. Note that these percentages are by *weight*. Originally, we thought that this discrepancy might be due to the fact that TA has a higher density (1.79 g/cm^3) than L (1.52 g/cm^3). The theoretical *volume* percentages of L and TA then become 54% and 46%, respectively. However, it is straightforward to average the percentages in any of the (estimated) pellets, because each of them is surrounded by black pixels. For the pellet in question, the estimated percentages are L: 71%, TA: 22%, and Other: 7%, which is very different from the expected percentages. One possible explanation is that, when the two materials were mixed together, many of the denser TA grains may have "sunk" beneath L grains and so were less detectable by the measurement system. We intend to investigate this issue further in the near future.

A similar phenomenon can be observed in the L/M/TA mixture image. In theory, the two mixed pellets have equal amounts of L, M, and TA (by *weight*). We see that most of the pixels in these two pellets are approximately 50/50 mixtures of L and M. Only small amounts of TA are found and most of these are near the pellet edges. The density of M is 1.54 g/cm^3, which is similar to that of L. So what we observe is consistent with the theory that many of the TA grains many have "sunk" and so were less detectable by the system.

Finally, we note that, unlike in the pure pellet proportion maps [see Fig. 11(a)], very few pixels in Fig. 11(b) and (c) have been classified as aspectral, and most of these are near the pellet edges.

VII. Conclusion

In this article, we outlined an automated system for analyzing reflection mode THz hyperspectral images of pure and mixed materials. The system is based on an extended linear model (3), which incorporated terms representing low-frequency variation and water vapor. The methodology also estimates the number of materials present in each pixel and tests for the presence of aspectral and NIL pixels, as well as identifying metals.

The proportion maps in Fig. 11(a)–(c) are encouraging in that they appear to produce reasonable approximations to the truth. However, we note again the potential limitation that the estimated proportion of the denser material (TA) was considerably lower than the actual proportion when it was mixed with less dense materials. We speculate that this may be due to the heavier TA grains tending to "sink" below the lighter grains.

Most of the theory applied in this article was developed for mineral exploration applications. The TSA software that implemented it is widely used as part of TSG, and especially in conjunction with a number of HyLoggers, mainly in Australia,[2] but also in several other countries. It is encouraging that it appears to also work well with THz *reflection* mode spectra with little modification. This robustness suggests that it should also work well with THz *transmission* mode spectra, which are often less noisy than their reflection mode counterparts.

An issue needing further research is an improved model for water vapor. The water vapor term in our model is based on a brass measurement when RH $= 42.6\%$. This is insufficient for higher RH levels, probably due to their being two types of water vapor (ortho-H_2O and para-H_2O). We needed to add a brass reference term [see (1)] to ensure that water vapor effects are adequately modeled at higher RH levels. Ideally, we would like to be able to incorporate a suitable additional water vapor term in our model so that we do not need to rely on a reference in every dataset. We intend to extend the water vapor experiment to higher RH levels and to build on the work of [39] and [40] to enable this.

Finally, we discuss the issue of speed. This is mainly determined by the following three quantities: the number of samples/pixels in the dataset (N); the size of the library (N_l); and the maximum number of materials assumed to be in the mixture in any pixel ($N_{\max}$). The datasets analyzed in this article were fairly small with $N = 7502, 1496$, and 1776 for the three datasets, respectively; see Table I. The library size was also small ($N_l = 6$) and $N_{\max} = 3$. Consequently, the average elapsed times for the three datasets (run on a Dell Latitude 7400 laptop with a 1.6 GHz Intel Core i5-83650 CPU) were 7.93, 3.63, and 4.07 s, respectively.

As N increases, the time taken will grow linearly, unless of course parallel computing is used, since the algorithm was applied to each spectrum separately. However, the time taken grows more quickly if either N_l or N_m is increased. The key parameter is $\binom{N_l}{N_m}$. For the datasets analyzed here, $\binom{6}{3} = 20$. However, for the SWIR mineral application mentioned previously, $N_l = 60$, and then, $\binom{60}{3} = 34\,220$, while $\binom{60}{4} = 487\,635$. A few years ago, we timed a dataset with $N = 109\,344$ spectra, analyzing pairs of spectra in parallel. It took 130 s, when $N_m = 3$, and 950 s, when $N_m = 4$. Hence, in recent versions of TSG, $N_m = 3$ is the default with $N_m = 4$ offered as an option.

There are faster approximations to the *full* subset selection algorithms used in this article. Probably the most promising is a method called L1 regularization, where a penalty term discouraging large values of N_m is added to the term in (9); see [31, eq. (28)]. Although it is faster, it is not as accurate. A comparison was carried out with some test data in [31, Sec. VIII.C]. However, more comparisons with other datasets are needed to understand the loss of accuracy better.

[2][Online]. Available: https://www.csiro.au/en/work-with-us/industries/mining-resources/Resourceful-magazine/Issue-18/Virtual-core

ACKNOWLEDGMENT

The authors would like to thank P. Fairman and T. Gretzinger for helpful discussions, and P. Rusconi for technical support.

REFERENCES

[1] A. Rostami, H. Rasooli, and H. Baghban, *Terahertz Technology: Fundamentals and Applications*, vol. 77. Berlin, Germany: Springer, 2010.
[2] H.-J. Song and T. Nagatsuma, *Handbook of Terahertz Technologies: Devices and Applications*. Boca Raton, FL, USA: CRC Press, 2015.
[3] S. L. Dexheimer, *Terahertz Spectroscopy: Principles and Applications*. Boca Raton, FL, USA: CRC Press, 2017.
[4] K. Kawase, Y. Ogawa, and Y. Watanabe, "Non-destructive terahertz imaging of illicit drugs using spectral fingerprints," *Opt. Exp.*, vol. 11, no. 20, pp. 2549–2554, 2003.
[5] Y. Watanabe *et al.*, "Component analysis of chemical mixtures using terahertz spectroscopic imaging," *Opt. Commun.*, vol. 234, no. 1–6, pp. 125–129, 2004.
[6] Y. Shen, P. Taday, D. Newnham, and M. Pepper, "Chemical mapping using reflection terahertz pulsed imaging," *Semicond. Sci. Technol.*, vol. 20, no. 7, 2005, Art. no. S254.
[7] Z. Zeng-Yan, J. Te, Y. Xiao-Han, X. Ti-Qiao, and X. Hong-Jie, "A method for quantitative analysis of chemical mixtures with time domain spectroscopy," *Chin. Phys. Lett.*, vol. 23, no. 8, 2006, Art. no. 2239.
[8] D. Jiang, S. Zhao, and J. Shen, "Quantitative analysis of the mixtures of illicit drugs using terahertz time-domain spectroscopy," *Proc. SPIE*, vol. 6840, 2008, Art. no. 68400U.
[9] G. Liu *et al.*, "Quantitative measurement of mixtures by terahertz time-domain spectroscopy," *J. Chem. Sci.*, vol. 121, no. 4, pp. 515–520, 2009.
[10] Y. Chen, Y. Ma, Z. Lu, Z.-N. Xia, and H. Cheng, "Chemical components determination via terahertz spectroscopic statistical analysis using micro-genetic algorithm," *Opt. Eng.*, vol. 50, no. 3, 2011, Art. no. 034401.
[11] Y. Chen, Y. Ma, Z. Lu, B. Peng, and Q. Chen, "Quantitative analysis of terahertz spectra for illicit drugs using adaptive-range micro-genetic algorithm," *J. Appl. Phys.*, vol. 110, no. 4, 2011, Art. no. 044902.
[12] L. Qiao, Y. Wang, Z. Zhao, and Z. Chen, "Identification and quantitative analysis of chemical compounds based on multiscale linear fitting of terahertz spectra," *Opt. Eng.*, vol. 53, no. 7, 2014, Art. no. 074102.
[13] X. Li, D. Hou, P. Huang, J. Cai, and G. Zhang, "Component spectra extraction from terahertz measurements of unknown mixtures," *Appl. Opt.*, vol. 54, no. 30, pp. 8925–8934, 2015.
[14] Y. Peng *et al.*, "Terahertz identification and quantification of neurotransmitter and neurotrophy mixture," *Biomed. Opt. Exp.*, vol. 7, no. 11, pp. 4472–4479, 2016.
[15] A. Pohl, N. Deßmann, K. Dutzi, and H.-W. Hübers, "Identification of unknown substances by terahertz spectroscopy and multivariate data analysis," *J. Infrared, Millimeter, Terahertz Waves*, vol. 37, no. 2, pp. 175–188, 2016.
[16] L. A. Sterczewski, M. P. Grzelczak, K. Nowak, B. Szlachetko, and E. Plinski, "Bayesian separation algorithm of spectral sources applied to -glucose monohydrate dehydration kinetics," *Chem. Phys. Lett.*, vol. 644, pp. 45–50, 2016.
[17] Y. Ma *et al.*, "Spectral data analysis and components unmixing based on non-negative matrix factorization methods," *Spectrochimica Acta A, Mol. Biomol. Spectrosc.*, vol. 177, pp. 49–57, 2017.
[18] Y. Peng *et al.*, "Qualitative and quantitative identification of components in mixture by terahertz spectroscopy," *IEEE Trans. THz Sci. Technol.*, vol. 8, no. 6, pp. 696–701, Nov. 2018.
[19] V. A. Trofimov and S. A. Varentsova, "A possible way for the detection and identification of dangerous substances in ternary mixtures using pulsed spectroscopy," *Sensors*, vol. 19, no. 10, 2019, Art. no. 2365.
[20] Z. Tang, H. Deng, Q. Liu, J. Guo, and L. Shang, "Quantitative analysis of low-concentration α- based on terahertz spectroscopy," *Anal. Methods*, vol. 12, no. 47, pp. 5684–5690, 2020.
[21] Y. Sun *et al.*, "Quantitative analysis of bisphenol analogue mixtures by terahertz spectroscopy using machine learning method," *Food Chem.*, vol. 352, 2021, Art. no. 129313.
[22] N. Keshava and J. F. Mustard, "Spectral unmixing," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 44–57, Jan. 2002.
[23] T. Zimmermann, "Spectral imaging and linear unmixing in light microscopy," *Microsc. Techn.*, vol. 95, pp. 245–265, 2005.
[24] C.-I. Chang, *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*, vol. 1. Berlin, Germany: Springer, 2003.
[25] A. Gowen, C. O'Donnell, P. J. Cullen, and S. Bell, "Recent applications of chemical imaging to pharmaceutical process monitoring and quality control," *Eur. J. Pharmaceutics Biopharmaceutics*, vol. 69, no. 1, pp. 10–22, 2008.
[26] R. F. Kokaly *et al.*, "USGS spectral library version 7," U.S. Geological Surv., Denver, CO, USA, Data Series 1035, 2017.
[27] K. H. Sim *et al.*, "A comprehensive spectral library for robust quantitative profiling of 10,000 proteins," *Sci. Data*, vol. 7, no. 1, pp. 1–13, 2020.
[28] L. Drumetz *et al.*, "Blind hyperspectral unmixing using an extended linear mixing model to address spectral variability," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3890–3905, Aug. 2016.
[29] K. Krishnakumar, "Micro-genetic algorithms for stationary and non-stationary function optimization," *Proc. SPIE*, vol. 1196, pp. 289–296, 1990.
[30] A. Miller, *Subset Selection in Regression (Monographs on Statistics and Applied Probability)*, 2nd ed. Boca Raton, FL, USA: Chapman & Hall, 2002.
[31] M. Berman *et al.*, "A comparison between three sparse unmixing algorithms using a large library of shortwave infrared mineral spectra," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3588–3610, Jun. 2017.
[32] N. Chimot *et al.*, "Terahertz radiation from heavy-ion-irradiated in 0.53 GA 0.47 as photoconductive antenna excited at 1.55 μm," *Appl. Phys. Lett.*, vol. 87, no. 19, 2005, Art. no. 193510.
[33] N. M. Burford and M. O. El-Shenawee, "Review of terahertz photoconductive antenna technology," *Opt. Eng.*, vol. 56, no. 1, 2017, Art. no. 010901.
[34] C. Baker *et al.*, "Detection of concealed explosives at a distance using terahertz technology," *Proc. IEEE*, vol. 95, no. 8, pp. 1559–1565, Aug. 2007.
[35] N. Palka, "THz reflection spectroscopy of explosives measured by time domain spectroscopy," *Acta Physica Polonica A*, vol. 120, no. 4, pp. 713–715, 2011.
[36] A. V. Oppenheim, J. R. Buck, and R. W. Schafer, *Discrete-Time Signal Processing*, vol. 2. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.
[37] M. Schodlok *et al.*, "Hylogger-3, a visible to shortwave and thermal infrared reflectance spectrometer system for drill core logging: Functional description," *Australian J. Earth Sci.*, vol. 63, no. 8, pp. 929–940, 2016.
[38] W. Withayachumnankul, B. M. Fischer, and D. Abbott, "Numerical removal of water vapour effects from terahertz time-domain spectroscopy measurements," *Proc. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 464, no. 2097, pp. 2435–2456, 2008.
[39] X. Xin, H. Altan, A. Saint, D. Matten, and R. Alfano, "Terahertz absorption spectrum of para and Ortho water vapors at different humidities at room temperature," *J. Appl. Phys.*, vol. 100, no. 9, 2006, Art. no. 094905.
[40] X. Miao, J. Zhu, K. Zhao, H. Zhan, and W. Yue, "Determining the humidity-dependent ortho-to-para ratio of water vapor at room temperature using terahertz spectroscopy," *Appl. Spectrosc.*, vol. 72, no. 7, pp. 1040–1046, 2018.
[41] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer, 2002.
[42] A. Hoerl and R. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, pp. 55–67, 1970.
[43] T. Hastie, A. Buja, and R. Tibshirani, "Penalized discriminant analysis," *Ann. Statist.*, vol. 23, pp. 73–102, 1995.
[44] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems (Classics in Applied Mathematics)*. Philadelphia, PA, USA: SIAM, 1995.
[45] G. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. New York, NY, USA: Wiley, 1992.
[46] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York, NY, USA: Springer-Verlag, 2001.
[47] T. Lumley, "Package 'leaps'," 2009. [Online]. Available: https://cran.r-project.org/web/packages/leaps/leaps.pdf
[48] Y. Guo, M. Berman, and J. Gao, "Group subset selection for linear regression," *Comput. Statist. Data Anal.*, vol. 75, pp. 39–52, 2014.
[49] Y. Guo and M. Berman, "Package 'groupsubsetselection'," 2017. [Online]. Available: https://cran.r-project.org/web/packages/groupsubsetselection/groupsubsetselection.pdf
[50] J. Nocedal and S. Wright, *Numerical Optimization*. New York, NY, USA: Springer, 1999.

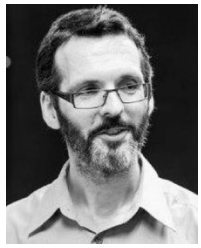
Mark Berman received the B.Sc. (Hons.) degree with University Medal in mathematical statistics and the Master of Statistics degree from the University of New South Wales, Sydney, NSW, Australia, in 1974 and 1976, respectively, and the Ph.D. and D.I.C. degrees in mathematical statistics from the Imperial College of Science and Technology, London, U.K., in 1978.

He was a Visiting Lecturer with the Department of Statistics with the University of California, Berkeley, CA, USA, during 1978–1979. Between 1979 and 2014, he was with the CSIRO Division of Mathematics, Informatics and Statistics (CMIS), Sydney, where he reached the position of Chief Research Scientist. He also led CMIS' Image Analysis Group from 1989 to 2000. In 1988, he took leave from CSIRO to establish the Image Processing and Data Analysis Group, Melbourne Research Laboratories of Broken Hill Proprietary Ltd. He has given Ph.D. courses in spectroscopy and hyperspectral imaging with the Technical University of Denmark, in 2007, and with Stanford University, in 2008 and 2014. He is currently a Visiting Scientist with CSIRO Data61, Marsfield, NSW, and an Adjunct Fellow with the School of Computing, Engineering and Mathematics, University of Western Sydney, Penrith, NSW. He is the author of several patents. His research interests include image analysis (especially hyperspectral), spectroscopy, and spatial data analysis.

Dr. Berman was an Associate Editor for *Computational Statistics and Data Analysis* in 2001–2006 and an Associate Editor for *Environmetrics* in 2010–2015. He was the recipient of Best Paper awards from IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING in 1990 and the *Journal of Chemometrics* in 2011.

Krunal Radhanpura received the B.Sc. degree with major in physics and the M.Sc. degree with major in physics—electronics from Gujarat University, Ahmedabad, India, in 2005 and 2007, respectively, and the Ph.D. degree in physics—terahertz optoelectronics from the University of Wollongong, Wollongong, NSW, Australia, in 2013.

The primary focus of his Ph.D. work was to identify and characterize new materials as high-power terahertz sources and sensitive detectors. He worked as a Postdoctoral Researcher with the Department of Engineering, University of Wollongong between 2013 and 2015. In 2015, he worked as a Research Associate with the Macquarie University Photonics Research Centre. Between 2016 and 2021, he was with the CSIRO Optical Systems Group, first as a Postdoctoral Research Fellow and more recently as a Research Scientist. His research interests include the field of terahertz science and technology. He has worked on identifying new terahertz sources and detectors, establishing novel terahertz systems, and performing terahertz spectroscopy, imaging, and hyperspectral imaging for a range of applications.



David Farrant (Senior Member, IEEE) received the B.Eng. degree in electronics from the Royal Melbourne Institute of Technology, Melbourne, VIC, Australia, in 1986, the M.Eng. degree in electrical from the University of Technology Sydney, Ultimo, NSW, Australia, in 1992, and the Ph.D. degree in mechanical engineering - optics from Loughborough University, Loughborough, U.K., in 2004.

He joined CSIRO Sydney in 1987, researching Fourier optical image processing, followed by development of metrology for LIGO optics. From 2004 onwards, he led projects developing tunable electro-optic filters for the IMaX Sunrise mission and the PHI instrument for the ESA Solar Orbiter mission. From 2011 to 2014, he led the Optics Research Group, covering a range of space and astronomy projects, including Advanced LIGO. He is currently a Senior Principal Research Scientist and leads the Optical Systems Team, CSIRO, Lindfield, NSW. His research interests include optical design and modeling, interferometry, and spectroscopy, covering optical through to terahertz wavelengths.

Dr. Farrant is a Member of the Australian Optical Society. He and his colleagues were the recipient of the SPIE Technology Achievement Award in 2012.