Autoregressive Modeling of Raman Spectra for Detection and Classification of Surface Chemicals

Quan Ding, *Student Member, IEEE,* Steven Kay, *Fellow, IEEE,* Cuichun Xu, *Member, IEEE,* and Darren Emge

Abstract

This paper considers the problem of detecting and classifying surface chemicals by analyzing the received Raman spectrum of scattered laser pulses received from a moving vehicle. An autoregressive (AR) model is proposed to model the spectrum and a two-stage (detection followed by classification) scheme is used to control the false alarm rate. The detector decides whether the received spectrum is from pure background only or background plus some chemicals. The classification is made among a library of possible chemicals. The problem of mixtures of chemicals is also addressed. Simulation results using field background data have shown excellent performance of the proposed approach when the signal-to-noise ratio (SNR) is at least -10 dB.

Index Terms

autoregressive model, detection and classification, likelihood function, Raman spectra

This work was sponsored by the US Army Research, Development and Engineering Command, under contracts W91ZLK-08-P-1214 and W91ZLK-06-P-0916.

Quan Ding and Steven Kay are with the Department of Electrical Computuer and Biomedical Engineering, University of Rhode Island, Kingston, RI 02881, USA. E-mail: dingqqq@ele.uri.edu, kay@ele.uri.edu.

Cuichun Xu is with the Air Traffic Control Group, ARCON Corporation, Waltham, MA 02541, USA. E-mail: ccxu@arcon.com.

Darren Emge is with the Edgewood Chemical Biological Center, Aberdeen Proving Ground, MD 21010, USA. E-mail: darren.emge@us.army.mil.

This paper extends the results of [1].

I. Introduction

Raman spectroscopy has been widely used in detection and classification of chemical agents in the presence of a material, termed the background [2], [3], [4], [5]. Many spectral data analysis techniques have been developed for this application. Supervised approaches such as regression analysis [6] and the generalized likelihood ratio test (GLRT) [7], [8] can be used when the background spectrum is known since this is a standard subspace detection problem. Unsupervised approaches such as independent component analysis (ICA) [9], canonical correlation [10], [11] and a correlation scheme based on a Gaussian filter [12] can be used when the background spectrum is unknown or varies due to noise.

In this paper, we study the unsupervised problem of detecting and classifying surface chemicals based on Raman spectral returns received from a moving vehicle. The Raman spectral data are collected by the laser interrogation of surface agents (LISA) system developed by ITT Industries. LISA provides standoff detection and identification of surface-deposited chemical agents based on short-range Raman sensing (see [13] for more information about the system). This detection and classification problem is complicated by many factors. Some of these are:

1. A background surface whose spectrum is unknown a priori and is changing with time.
2. Target chemicals that, even if present, are presented to the detector only a fraction of the time. This is due to an uneven and incomplete distribution of deposited surface chemicals.
3. The energy in the target return varying with the amount of chemical, the type of chemical, and the range to the chemical.
4. The possible presence of more than one chemical, i.e., a chemical mixture.
5. Impurities in the background that present themselves as unknown chemical deposits.

In order to design algorithms that are able to handle this multitude of unknown situations we rely heavily on *adaptive processing*. The approaches to be described take advantage of any information that is known and that can reasonably be assured to be valid in an operational environment. For the remaining uncertainties the algorithms will estimate *on-line* the information necessary for their successful implementation. We will first discuss detection and classification (identification) of a *single chemical from a library of possible chemicals*. Next, we will extend the results to the mixture problem, i.e, when one or two or possibly three chemical targets may be present in a single scattered spectrum.

The paper is organized as follows. Section II describes the problem and the two-step detection followed by classification scheme that is proposed. An AR model that models the Raman spectrum is described in Section III. In Section IV we derive the detection test statistic and the overall algorithm in order to maintain a low false alarm rate which [1] did not consider. The experimental detection performance for field background data is shown in Section V. Simulation results in Section VI show that a very low false alarm rate can be obtained. The classification algorithm is derived in Section VII. Here we extend the case of a single chemical present to mixtures of chemicals, which was not treated in [1]. In Section VIII, we present the classification performance for field background data. Finally, Sections IX draws the conclusion.

II. Problem Statement and Rationale of Approach

Consider the case when a moving vehicle is mounted with a Raman spectroscopy unit that probes the ground surface every short time interval (40 milliseconds in our case). A Raman spectrum or a pulse $I_i(F)$ is received at the i^{th} probe and consecutive Raman spectra of the road surface are received as the vehicle moves. Each Raman spectrum is an $N_f \times 1$ vector given at equally space wavenumbers F. We assume that the background is relatively stationary in composition, that is, it is a road of the same type for a certain time interval. There are also possibly some of M target chemicals present on the background. As a result, the received spectrum at the i^{th} probe could be from background plus noise or background plus noise and one or several chemicals. We wish to design a testing procedure that decides if no chemicals are present or if chemicals are present, which chemicals are deposited on the background.

Current approaches to the detection problem have been plagued with high false alarm rates. Indeed for any operational system the false alarm rate must be controlled or else the system is deemed unreliable and cannot be used. Nearly identical considerations arise in sonar [14] and radar [15]. It has been generally accepted, and this philosophy is reflected in the design of these systems, that one first performs a decision of either a detection or no detection and then follows this with a classification. In this way the false alarm rate can be controlled since the initial step does not consider which target may be present but only that *some target is present*. This initial binary hypothesis test then allows one to control the false alarm rate and to reduce it to a reasonable level. This is in contrast to attempting to decide whether no target is present versus a subset of M possible targets. The latter approach requires one to formulate a decision strategy that can decide among multiple hypotheses, for which an error rate or false alarm rate will be much higher.

III. Spectral Modeling

As mentioned in Section 1 the background spectrum is unknown and can change in time. For the algorithms to accommodate this uncertainty, it is necessary to estimate the spectrum on-line. To do so a spectral estimator that can estimate the spectrum from a single pulse accurately and with reasonable computation to allow a real-time implementation is the *autoregressive (AR)* spectral estimator [16]. Similar approaches have been used in radar [17] and sonar [18]. To implement this estimator it is assumed that spectral data from the output of the Raman spectroscopy unit is available over a spatial frequency band, i.e., wavenumber band, which by letting F denote spatial frequency, extends from $F = 0$ to $F = F_c$, the cutoff frequency. This spectral data $I(F)$ is also called the periodogram in analogy with Fourier based methods of spectral estimation. Given $I(F)$ for $0 \leq F \leq F_c$, the AR spectral estimate is found as follows, with details given in [16]:

1. Assume a model order, denoted by p, for the AR spectral estimate. This order is an integer, with smaller values preferred since it relates to the number of parameters in the model and hence the number of unknowns to be estimated.

2. Based on $I(F)$ find the real-valued autocorrelation sequence, denoted as $\{r[0], r[1], \ldots, r[p]\}$, which is a sampled version (at a rate of $1/\Delta$ samples per sec) of the inverse *continuous-time* Fourier transform of $I(F)$ as
$$r[k] = \int_0^{2F_c} I(F) \exp(j2\pi F k\Delta) dF \qquad k = 0, 1, \ldots, p \tag{1}$$
where Δ is the interval in time between successive samples of the autocorrelation function. The sample interval should be chosen to be less than $1/(2F_c)$. Note that since the spectral data $I(F)$ has a spectrum that is one-sided, we let $I(F) = I(2F_c - F)$ for $F_c \leq F \leq 2F_c$. In this way $I(F)$ can be viewed as one period of a periodic spectrum and therefore $r[k]$ becomes real-valued. The implied sampling rate is then $2F_c$.

3. Solve the Yule-Walker equations to estimate the AR filter parameters $\{a[1], a[2], \ldots, a[p]\}$ from
$$\begin{bmatrix} r[0] & r[-1] & \ldots & r[-(p-1)] \\ r[1] & r[0] & \ldots & r[-(p-2)] \\ \vdots & \vdots & \ddots & \vdots \\ r[p-1] & r[p-2] & \ldots & r[0] \end{bmatrix} \begin{bmatrix} a[1] \\ a[2] \\ \vdots \\ a[p] \end{bmatrix} = - \begin{bmatrix} r[1] \\ r[2] \\ \vdots \\ r[p] \end{bmatrix} \tag{2}$$
and then use these estimated filter parameters to find the excitation noise variance σ_u^2 as
$$\sigma_u^2 = r[0] + \sum_{k=1}^{p} a[k] r[-k]. \tag{3}$$
Note that the matrix is symmetric and Toeplitz since $r[-k] = r[k]$.

4. Once the parameters $\{a[1], a[2], \ldots, a[p], \sigma_u^2\}$ have been found the estimated AR spectrum is
$$P(F) = \frac{\sigma_u^2 \Delta}{|1 + a[1]\exp(-j2\pi F\Delta) + \cdots + a[p]\exp(-j2\pi pF\Delta)|^2} \tag{4}$$

for $0 \leq F \leq F_c$.

Note that this procedure estimates the AR spectrum for a given AR model order p. However in practice, we also need to estimate the appropriate order p since a large p will cause overfitting and a small p will cause underfitting. We next assume that the frequencies have been normalized to discrete frequencies as $f = F\Delta$ so that $0 \leq f \leq 1$ and hence digital techniques can be used. Clearly, the upper cutoff frequency F_c corresponds to $f = 1/2$. The AR model order can be estimated as follows (see Appendix A for the derivation):

1. For the spectral data $I(f)$, and for each model order p, estimate the AR filter parameters or $\{a[1], a[2], \ldots, a[p]\}$ using (1) and (2), and then estimate the AR filter frequency response as
$$\hat{A}_p(f) = 1 + a[1]\exp(-j2\pi f) + \cdots + a[p]\exp(-j2\pi fp) \tag{5}$$
2. Calculate the generalized likelihood ratio $l_{G_p}(\mathbf{x})$ for each model order p by
$$l_{G_p}(\mathbf{x}) = -N \ln \frac{\sum_{k=1}^{N_f} |\hat{A}_p(f_k)|^2 I(f_k)\Delta f}{\sum_{k=1}^{N_f} I(f_k)\Delta f} \tag{6}$$
where N is the unknown number of samples in the time domain since $\mathbf{x}$ is fictitious. We will use $N = 2N_f$, which produces good results.
3. Choose the model order with the largest of the following
$$EEF(p) = \begin{cases} l_{G_p}(\mathbf{x}) - p\left[\ln\left(\frac{l_{G_p}(\mathbf{x})}{p}\right) + 1\right] & \text{if } \frac{l_G(\mathbf{x})}{p} > 1 \\ 0 & \text{if } \frac{l_{G_p}(\mathbf{x})}{p} \leq 1 \end{cases} \tag{7}$$
This is the exponentially embedded families (EEF) as a model order selection criterion that has been recently proposed [19].

As an example, using an estimated model order of $p = 40$ and a *single pulse* from a background of asphalt, the AR spectral estimate and original periodogram data are shown in Figure 1. Note that the AR spectral estimate is able to model the general shape of the data spectrum as well as the prominent peaks and valleys. Additionally, if an artificial signal is included in the spectral data, then the AR spectral estimate (with a different estimated AR model order of $p = 44$) appears as in Figure 2. Similar results have been obtained for other surfaces such as gravel and grass. What this says is that *the AR spectral model with appropriate order p is adequate for representing the main details of a spectrum using Raman spectroscopy*. This includes the cases of background only being present as well as a target chemical deposited on a background. Consequently, for the development of signal processing algorithms it allows us *to consider the spectral data as having been obtained from a hypothetical AR time series that has been Fourier transformed and magnitude-squared.* If we further assume that this hypothetical time series is

Gaussian, then many of the powerful techniques of statistical signal processing [20], [7] can be brought to bear upon this problem. As we will see later, the Gaussian assumption is not entirely accurate but algorithms based on it still perform exceptionally well.

IV. Detection Algorithm

The detection algorithm consists of two parts. This is necessary to avoid a high false alarm rate as described previously. It is assumed that when a chemical is present it must be present in a certain percentage of the returned pulses. A detector based upon a single pulse with a reasonably low false alarm rate would require a high threshold and hence a poorer probability of detection. Hence, the chemical present condition is defined to be in effect when a certain percentage of *successive pulse returns indicate a chemical*. The pulse returns which do not indicate a chemical, when indeed a chemical present condition is in effect, results from the lack of presence of a chemical in the illuminated area of the laser imaging system. Thus, we have designed a detection system that

1. Examines each successive pulse for a threshold crossing of a test statistic.
2. Registers a chemical present condition when a suitable number of threshold crossings are present over a fixed interval of time

We next examine each of these procedures in detail.

A. Test Statistic

The test statistic is computed for each pulse return or sequentially in time. To estimate the background, we will need spectral data from the M_B previous pulse returns that *do not have a threshold crossing*. The choice of M_B is made to ensure that the background has not changed over this time period. For example, if $M_B = 25$, then for a laser firing rate of 25 pulses/sec, we have effectively assumed that the background spectral shape is stationary over the time interval of $M_B/25 = 1$ second. Analysis of field data supports this assumption. However, it has also been found that although the background spectral *shape* is stationary over a short period of time, its *overall level may change significantly from pulse to pulse.* This necessitates us to base any test statistic *on the shape of the spectrum but not its total power*. This can be done by assuming for the background a fixed set of AR *filter* parameters from pulse to pulse but with a time varying excitation noise variance. Also, if some of the previous pulse returns have threshold crossings of the test statistic, then we exclude them from the M_B pulses used in estimating the background. The test statistic is computed as follows (see Appendix B for the derivation and explicit statistical assumptions):

1. Using the previous M_B pulses that *do not have threshold crossings*, compute the average Raman spectrum. Because the overall background power level can change from pulse to pulse we must first normalize the power before we average. To do so we set the total power of each pulse to one by scaling appropriately. Let $I_{B_i}(f)$ represent the Raman spectrum for the ith pulse *after power normalization*. Then, we compute the sample average of the background spectral data as

$$\bar{I}_B(f_k) = \frac{1}{M_B} \sum_{i=1}^{M_B} I_{B_i}(f_k) \tag{8}$$

for $k = 1, 2, \ldots, N_f$, where N_f is the number of spectral data points of the Raman spectrum. Also, $I_{B_i}(f_k)$ is the Raman spectral data for the ith pulse at frequency f_k, assuming that it previously did not produce a threshold crossing.

2. Estimate the AR model order p using the procedure described in (5), (6) and (7). For this estimated order p, use the procedure described in (1) and (2) to find the AR filter parameters of $\bar{I}_B(f_k)$. These are denoted by $\{a_B[1], a_B[2] \ldots, a_B[p]\}$, where the subscript refers to the background spectral model. Note that these may change in time and therefore will have to be updated periodically. The estimated background AR filter frequency response then becomes

$$A_B(f) = 1 + a_B[1]\exp(-j2\pi f) + \cdots + a_B[p]\exp(-j2\pi fp) \tag{9}$$

3. Using the Raman spectrum for the return pulse under consideration, which we denote as $I_T(f)$ and where T refers to a potential target, estimate the AR model order q using the procedure described in (5), (6) and (7). Compute the AR parameters, again using (1) and (2). Note that power normalization is not needed since only the AR *filter* parameters are estimated. This produces the AR filter parameters or $\{a_T[1], a_T[2], \ldots, a_T[q]\}$ and the estimated AR filter frequency response for the current pulse under consideration as

$$\hat{A}_T(f) = 1 + a_T[1]\exp(-j2\pi f) + \cdots + a_T[p]\exp(-j2\pi fq) \tag{10}$$

4. The generalized likelihood ratio test (GLRT) statistic is finally computed as

$$T_D = \ln \frac{\sum_{k=1}^{N_f} |A_B(f_k)|^2 I_T(f_k)}{\sum_{k=1}^{N_f} |\hat{A}_T(f_k)|^2 I_T(f_k)} \tag{11}$$

will yields values $T_D \geq 0$. Note that power normalization is not required for $I_T(f)$ since T_D does not depend on scaling of $I_T(f)$.

This test statistic, which may be viewed as an anomaly detector, will indicate when the return from any pulse produces a spectrum significantly different from the background. No information, however, is obtained about the type of departure and hence of a particular chemical. A threshold crossing, which occurs if $T_D > \gamma$ for a threshold γ, indicates that the spectrum of the current pulse does not match the background spectrum. As an example, impurities in the surface will also cause a threshold crossing. Hopefully, however, these will be isolated occurrences and not produce a chemical present condition. If this is not the case, then a classification indicating impurities will be needed.

B. Overall Detection Algorithm

The test statistic given by (11) is computed for each pulse. A threshold crossing indicates a possible chemical detection in that pulse. In order to declare a chemical present, however, we expect a certain percentage of the pulse returns to have a chemical in them. This percentage is currently set to 10%. For example, if a chemical is present, then for 100 pulses, we expect 10 or more of them to produce threshold crossings, assuming the test statistic always produces a threshold crossing when a chemical is present in the pulse return. The remaining 90 test statistics will not have a threshold crossing since they are based on data for which the laser did not illuminate the chemical, as explained previously. With this assumption we can now set the desired threshold for T_D. We assume that a chemical is present if 10% *or more* of the test statistics in a given block of pulse data produce threshold crossings. These threshold crossings need not be sequential, but can be scattered anywhere within the block. For example, if the block consists of 100 successive pulse returns, then a chemical is declared to be present if at least 10 of the test statistics produce a threshold crossing. This block of 100 successive pulses is assumed to "slide along" in time. For the example described below, the blocks are overlapped by 50%, although other overlaps can be used.

Next in order to ensure a fixed false alarm rate, we need to set the threshold, which we call γ, for T_D appropriately. Thus, we must specify the probability of a threshold crossing for T_D, which is $P_{FA_p} = \Pr[T_D > \gamma|\mathcal{H}_0]$, and is the probability of false alarm for a *single pulse*. Then, once P_{FA_p} is found, the threshold γ can be specified. It is shown in Appendix B how P_{FA_p} can be found so that the overall false alarm rate is less than one false alarm per h hours.

First we find P_{FA_b}, which is the probability of a false alarm for a single block, and is given as the solution of

$$(1 - P_{FA_b})^L + L\,(1 - P_{FA_b})^{L-1} = 0.99 \tag{12}$$

where $L = 1800h$ is the number of blocks analyzed in h hours. This value for L assumes a pulse rate of 25 per second, a block size of 100 pulses, and a 50% block overlap. Each block is therefore 4 sec long with an overlap of 2 sec. This can be solved for P_{FA_b}. Once P_{FA_b} is found we can determine P_{FA_p} by solving the equation

$$P_{FA_b} = 1 - \sum_{i=0}^{9} \binom{100}{i} P_{FA_p}^{i} \left(1 - P_{FA_p}\right)^{i}. \tag{13}$$

This is just the calculation that a false alarm occurs in a block, which is defined as 10 or more threshold crossings out of 100 possible ones. For example, if $h = 2$ hours and therefore $L = 3600$, then from (12) we have that $P_{FA_b} = 5 \times 10^{-5}$. Using this value in the left-hand-side of (13) we can solve for P_{FA_p}, which is about $P_{FA_p} = 0.02$. The details are given in Appendix B. As a result we need to find the threshold γ so that the probability that $T_D > \gamma$ for a single pulse is 0.02.

Theoretically, the GLRT statistic T_D should have a chi-squared probability density function (PDF) with q degrees of freedom [7], which would allow us to determine γ. It has been found through analysis of field data, however, that this theoretical PDF is not sufficiently accurate. (This is why, as mentioned earlier, the Gaussian assumption for the fictitious time series is not always accurate.) As a result, it is necessary to estimate the PDF of T_D when background only is present and then use this to set the threshold. It is conceivable that this threshold will depend upon the background statistics, which are unknown. We next indicate how this is done on-line.

Assume that we have I independent and identically distributed test statistics T_{D_i} for $i = 1, 2, \ldots, I$. We can estimate on-line the right-tail probability of the PDF by using an AR model for the PDF [21], [22]. The procedure is as follows:

1. Normalize the test statistics by a constant equal to the maximum value of the T_{D_i}'s. If we denote this as $T_{\max} = \max_{i=1,\ldots,I} T_{D_i}$, then we form the new data set $\tilde{T}_{D_i} = T_{D_i}/T_{\max}$. Thus all values are now in the range $[0, 1]$ since $T_{D_i} \geq 0$.
2. Next we use the AR spectral estimator, but as a PDF estimator, with the "estimated autocorrelation" sequence (actually the estimated characteristic function)

$$r[k] = \frac{1}{I} \sum_{i=1}^{I} \exp\left(j2\pi k \tilde{T}_{D_i}\right) \tag{14}$$

for $k = 0, 1, \ldots, p$ and will in general be complex-valued. The AR parameters are estimated using (2) and (3) but with $r[-k] = r^*[k]$. The estimated PDF of T_D then becomes

$$p_{T_D}(t) = \frac{\sigma_u^2}{|1 + a[1] \exp(-j2\pi t/T_{\max}) + \cdots + a[p] \exp(-j2\pi pt/T_{\max})|^2} \tag{15}$$

for $0 \leq t \leq T_{\max}$, where σ_u^2 is real-valued and $\sigma_u^2 > 0$, and the $a[k]$'s are complex-valued.

3. Determine the threshold by numerical integration as the value of γ that solves

$$\int_{\gamma}^{\infty} p_{T_D}(t)dt = P_{FA_p}. \tag{16}$$

V. Experimental Detection Performance for Field Background Data

The following results make use of 10,000 pulses of concrete field background data to which chemical signatures obtained in the laboratory were added using a computer. The first 500 pulses of background data only are used for initialization so that the background spectrum can be estimated as needed for $|A_B(f_k)|^2$ in (11). Also, using the same 500 pulses the threshold γ is found for the detector using (14–16). The threshold is then fixed for the entire remaining 9500 pulses. From the results to be presented it is found that the threshold will have to be periodically updated. After the initialization period a chemical signature is added to the background at a rate of 10% in a random manner. As explained previously, a window of 25 previous pulses without threshold crossings is used to update the background spectrum.

As an example of the detection performance for concrete field background data we set the threshold so that the false alarm rate is at most 1 per 2 hours, as explained previously. Then, we plot the probability of detection P_D versus signal-to-noise ratio (SNR) for a single chemical. The SNR is defined as the *broadband* SNR. This is

$$SNR = 10 \log_{10} \frac{\sum_{k=1}^{N_f} \theta P_s(f_k)}{\sum_{k=1}^{N_f} P_B(f_k)} \tag{17}$$

where $P_B(f)$ is the PSD of the background, $P_s(f)$ is the known spectral signature for the chemical, and θ is a scaling factor that produces the desired SNR. We next plot the probability of detection P_{D_p} based on a *single pulse*, which is the probability of a threshold crossing, versus SNR. This is done by examining the pulses where we know that a chemical has been added throughout the 9500 pulses. The spectra of all the chemicals that are used in the simulations are plotted in Figure 3 (one or more chemicals are added to the background for either detection performance or classification performance). When chemical 15 is added to the background, the probability of detection is shown in Figure 4. It is seen that the probability of detection is perfect for SNRs in excess of -10 dB. If we instead add chemical 31, the results are as shown in Figure 5. Again the detection performance is nearly perfect at a fairly low SNR.

VI. Experimental False Alarm Rate Performance

Since the threshold is critical to maintain a reasonable false alarm rate, we performed an experiment to determine if the computed one was reasonable. For the same 10,000 pulses (6.67 minutes of data), we used the first 500 pulses for initialization. Then, for a concrete background (no added chemical) we implemented the detection algorithm previously described. A false alarm will occur if the number of threshold crossings for a block of 100 pulses exceeds 10. The same threshold as found from the first 500 pulses was used throughout the remaining 9500 pulses. It was found that for blocks that are 50% overlapped, as assumed in the analysis, there were 3 false alarms as shown in Figure 6. However, two of the false alarms are close together and so can be considered as the same one. Hence, there are 2 false alarms. This is still higher than predicted. In a two-hour period there would be on the average 36 false alarms, instead of the prediction of 1. This would imply that the background is not stationary over this time interval. Thus we should update the threshold periodically.

In this example we then updated the threshold for every 500 pulses. That is, if there is no detection declared for all blocks within 500 pulses, we update the threshold using all the 500 test statistic T_D's, and use the updated threshold for the next 500 pulses. Otherwise if there is detection of chemicals for some blocks in these 500 pulses, we use the test statistic T_D's in the blocks within these 500 pulses that do not declare detections to update the threshold. It was found that for blocks that are 50% overlapped, there were no false alarms for the remaining 9500 pulses. When we updated the threshold every 500 pulses, even for successive blocks, there were no false alarms either. This suggests our derivation of the required threshold needed to control the false alarm rate.

VII. Classification

For the purpose of this paper, we constrain ourselves to *single pulse classification*, that is, we perform a classification based on a single pulse that has a threshold crossing. Once the detection algorithm declares that some chemicals are present in a block of data (of say 100 pulses, for which 10 or more have had threshold crossings), we proceed with single pulse classification on those pulses that have threshold crossings within this block, and choose the chemicals that appear most often in single pulse classification. In subsection VII-A, it is assumed that only one of M chemicals may be present. In subsection VII-B, we consider the problem of mixtures of chemicals, i.e., the case when two or possibly three chemicals are present in a single scattered pulse. To do it, we initially assume that we know there are K out of M chemicals on the background in VII-B, so we just need to decide which combination of K chemicals is present. Then we will use a model order selection criterion to decide how many chemicals are present, i.e., the value of K, in subsection VII-C.

A. Classification if Only One of M Chemicals Is Present

To determine which chemical is present we compute M test statistics and choose the chemical with the largest value of the test statistic. The test statistic that is used is that associated with a *locally most powerful* (LMP) test [7]. It can also be interpreted as an estimate of the chemical amplitude normalized by its standard deviation. The overall classification procedure is as follows (see Appendix D for the derivation and detailed model):

1. For the pulse $I(f)$ that has a threshold crossing, estimate the background by using the previous 25 pulses that did not have threshold crossings. To do so first normalize the power in each of these pulses to have a total power of one and then average the spectra to yield $\bar{I}(f)$. Then, estimate the AR parameters to obtain $P_B(f)$ as given by (4). Next normalize $I(f)$ to make $\sum_{k=1}^{N_f} I(f) = \sum_{k=1}^{N_f} P_B(f)$. By doing this we make sure the pulse has the same power as the background. Since the chemical signature power is assumed to be small, this also guarantees the pulse has about the same background power, which satisfies the assumption of the M-ary hypothesis test as in (30) in Appendix D. (Note also that by this assumption, the background normalization needed to form $P_B(f)$ should not be affected by the chemical present.)
2. For each chemical signature $P_{s_i}(f)$, use the estimate of the background $P_B(f)$ and the pulse data to be classified $I(f)$ to compute the classification test statistic
$$T_{C_i} = \frac{\displaystyle\sum_{k=1}^{N_f} \frac{P_{s_i}(f_k)}{P_B(f_k)} \left(\frac{I(f_k)}{P_B(f_k)} - 1 \right)}{\sqrt{\displaystyle\sum_{k=1}^{N_f} \frac{P_{s_i}^2(f_k)}{P_B^2(f_k)}}} \tag{18}$$
Note that the chemical signature $P_{s_i}(f)$ need not be power normalized since T_{C_i} does not depend on the scaling of $P_{s_i}(f)$.
3. Repeat step 2 for $i = 1, 2, \ldots, M$.
4. Choose the chemical that produces the largest T_{C_i}.

Preliminary results indicate that even with a single pulse nearly a perfect classification can be made as described in Section VIII.

B. Classification if K out of M Chemicals Are Present

This problem is more complicated since we now need to pick K out of M chemicals instead of just picking one out of M. The total number of possible combinations is $\binom{M}{K}$. An asymptotic likelihood function method is proposed. The idea is that we first find the asymptotic maximum likelihood estimate (MLE) of the unknown powers for chemical signatures and plug it into the corresponding log-likelihood function of this hypothesis. The chemical combination that produces the largest log-likelihood is chosen. The classification procedure is as follows (see Appendix E for the derivation and detailed model):

1. The first step is the same as in the previous subsection. Obtain the average spectrum of the chemical plus background $\bar{I}(f)$.
2. For each chemical combination hypothesis, compute the asymptotic MLE of the chemical signature powers by
$$\hat{\boldsymbol{\theta}} = \mathbf{I}^{-1}(\mathbf{0}) \left. \frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\mathbf{0}} \tag{19}$$
where
$$\left. \frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \theta_i} \right|_{\boldsymbol{\theta}=\mathbf{0}} = \frac{N}{2} \sum_{k=1}^{N_f} \frac{P_{s_{k_i}}(f_k)}{P_B(f_k)} \left(\frac{I(f_k)}{P_B(f_k)} - 1 \right) \Delta f. \tag{20}$$
and
$$\mathbf{I}_{ij}(\mathbf{0}) = \frac{N}{2} \sum_{k=1}^{N_f} \frac{P_{s_{k_i}}(f_k) P_{s_{k_j}}(f_k)}{(P_B(f_k))^2} \Delta f. \tag{21}$$
If there is at least one negative element in $\hat{\boldsymbol{\theta}}$, set the log-likelihood of this hypothesis to $-\infty$. Otherwise plug $\hat{\boldsymbol{\theta}}$ into the following log-likelihood function
$$\begin{aligned} \ln p(\mathbf{x};\boldsymbol{\theta}) &= -\frac{N}{2} \sum_{k=1}^{N_f} \left[\ln \left(\sum\nolimits_{i=1}^{K} \theta_i P_{s_{k_i}}(f_k) + P_B(f_k) \right) + \frac{I(f_k)}{\sum_{i=1}^{K} \theta_i P_{s_{k_i}}(f_k) + P_B(f_k)} \right] \Delta f \\ &\quad -\frac{N}{2} \ln(2\pi). \end{aligned} \tag{22}$$
Also note that the chemical signatures $P_{s_{k_i}}$'s do not need to be power normalized since the ith element $\hat{\theta}_i$ of $\hat{\boldsymbol{\theta}}$ from (19) is proportional to $1/P_{s_{k_i}}$. Thus $\theta_i P_{s_{k_i}}$ in (22) does not depend on the scaling of $P_{s_{k_i}}$.
3. Repeat step 2 for all the $\binom{M}{K}$ hypotheses.
4. Choose the chemical combination that corresponds to the hypothesis having the largest log-likelihood. Note that the number of data samples N in the time domain is unknown. However, we can compare the log-likelihoods without knowing N. This is because that $\hat{\boldsymbol{\theta}}$ does not depend on N since N cancels in (19) and N is just a scaling factor in (22).

C. Model Order Selection on How Many Chemicals Are Present in the Mixture

We have considered the case when we know the number of chemicals that are present. But in practice, this information is unknown a-priori. Thus, we need to select the *model order*, i.e., how many chemicals are present. Again, we will use the EEF as the model order selection criterion. For each hypothesis, the EEF can be calculated by

$$EEF = \begin{cases} l_G(\mathbf{x}) - K\left[\ln\left(\frac{l_G(\mathbf{x})}{K}\right) + 1\right] & \text{if } \frac{l_G(\mathbf{x})}{K} > 1 \\ 0 & \text{if } \frac{l_G(\mathbf{x})}{K} \leq 1 \end{cases} \tag{23}$$

where K is the assumed number of chemicals deposited on the background and

$$l_G(\mathbf{x}) = 2\ln\frac{p(\mathbf{x};\hat{\boldsymbol{\theta}})}{p(\mathbf{x};\mathbf{0})}.$$

The log-likelihood functions $\ln p(\mathbf{x};\hat{\boldsymbol{\theta}})$ and $\ln p(\mathbf{x};\mathbf{0})$ can be found by plugging (19) and $\boldsymbol{\theta} = \mathbf{0}$ into (22), respectively. We choose the hypothesis with the largest EEF value.

Since EEF is increasing with $l_G(\mathbf{x})$, for the same model order K, the largest $l_G(\mathbf{x})$ corresponds to the largest EEF. So for each K, we just need to find $l_G(\mathbf{x})$'s for all the $\binom{M}{K}$ hypotheses, choose largest $l_G(\mathbf{x})$, and plug it into (23). Then, we compare the EEF's for different K's and choose the model with the largest EEF. We select the hypothesis with the largest $l_G(\mathbf{x})$ for the model order that has been chosen.

Again, we need the number of data samples N in the time domain in computing the EEF, since (22) depends on N. We will assume that same number of samples in the time domain as in the frequency domain. Since we have $N_f = 1024$ samples equally spaced on half a period in the frequency domain, we will use $N = 2N_f = 2048$. By simulation we have seen that the performance is excellent with $N = 2048$.

VIII. Experimental Classification Performance for Field Background Data

For the same data conditions as for the detection experiment, we isolate all the pulses that have had threshold crossings. The probability of a correct single pulse classification is found by

$$P_C = \frac{\text{number of correct classifications for the pulses that have the added chemical}}{\text{number of pulses that have the added chemical}}$$

First we consider the case when there is only one chemical present. Using a library of $M = 60$ possible chemicals we classify the pulses with threshold crossings as per the discussion in Section 6. The results for chemicals 15, 31 and 45 are shown in Figures 7, 8 and 9 respectively. Again nearly perfect results are obtained for an SNR in excess of -10 dB.

Next we added the two chemicals 15 and 16 to the background, each chemical with the same SNR. We assume that we know the number of chemicals present. In the simulation, we have found that the classifier will sometimes choose chemicals 16 and 29. In Figure 10, we see that the probability of choosing chemicals 15 and 16 does not go to 1 as SNR increases. But if we consider chemicals 15 and

29 to be the same, the performance is much improved. This is because the spectrum of chemical 15 is very similar to that of chemical 29 as shown in Figure 3. The correlation between the spectra of the two chemicals is 0.968, which means that they are approximately linearly dependent. In this case, it is hard to distinguish between these two chemicals. Two approaches are possible. We can either treat chemicals 15 and 29 as the same, or remove chemical 29 from the library. When the classifier chooses chemical 15 in the case chemical 29 is removed, a second stage classification can be performed to further discriminate between chemical 15 and chemical 29. The same approach is considered in [11] where one spectrum of the spectrum pairs whose correlations are greater than a threshold is removed. The performance is shown in Figure 11 when chemical 29 is removed from the library. The simulation results for the chemical 20 and 45 combination and for the chemical 31 and 45 combination are shown in Figure 12 and Figure 13 respectively. These combinations are easily classified.

Next we would like to ascertain how the EEF works for the case when the number of chemicals in the mixture is unknown. We assume that there are at most 3 chemicals present. Thus, we need to compare the EEF for $K = 1, 2, 3$. Chemicals 15, 56 and 58 with the same SNR are added to the background. The performance of the EEF is compared to that of the minimum description length (MDL) criterion. The MDL is based on coding arguments [23] and can also be derived by an asymptotic Bayesian procedure [24]. We still consider chemicals 15 and 29 as the same chemical because of the high correlation between them. The resulting probability of correct classification versus SNR is shown in Figure 14. The result for the chemical 15, 31 and 45 combination is shown in Figure 15. The result for the chemical 20 and 45 combination is shown in Figure 16. Comparing Figure 12 and Figure 16, we see that the former produces a slightly higher probability of correct classification. This is because for Figure 12, we assume that we know the number of chemicals, but for Figure 16, we need to estimate the number of chemicals.

As we have seen, some of the target chemicals in the library are highly correlated. As a result, we need to remove some of them from the library or else treat them as a group of similar chemicals. A further consideration is that a linear combination of some chemicals might appear similar to another single chemical. Then the classifier would not perform well if that single chemical were present, since we might choose the chemicals that form the equivalent linear combination instead. A future paper will address this issue.

IX. Conclusion

An AR model has been proposed for a chemical detector and a chemical classifier based on Raman spectra. The use of a detection procedure followed by a classification scheme is used to control the false alarm rate. This is an unsupervised approach which estimates on-line the information of the non-stationary background data. Experiments with field background data have shown excellent performance of both the detector and the classifier.

APPENDIX A
DERIVATION OF ESTIMATING THE AR MODEL ORDER

The basic assumption is that the spectral data obtained through the action of the Raman spectroscopy unit can be modeled as a periodogram of real-valued Gaussian data. This implies certain statistics of the spectral data, which although not completely satisfied, allows us to derive a detector that will perform well in practice. As an example of this modeling discrepancy, in analyzing field obtained spectral data it has been found that the probability density function of the spectral data is not chi-squared with two degrees of freedom, which is implied by the Gaussian model. Hence, algorithms which push this Gaussian assumption too far may not work as predicted. Fortunately, for the problem at hand the algorithms so derived appear to perform exceedingly well.

Assume that N samples $\{x[0], x[1], \ldots, x[N-1]\}$ in the time domain of the Gaussian AR random process are observed (this is fictitious). We assume that we have the same number of samples in the time domain as in the frequency domain (as in a discrete Fourier transform). Since N_f is the number of samples equally spaced on half a period in the frequency domain, we have $N = 2N_f$. We need to estimate the order of the AR process. This is a multiple hypothesis testing problem with

$$\begin{aligned}
\mathcal{H}_0 &: \quad a[1]=0, a[2]=0, \ldots, a[p_M]=0, \sigma_u^2>0 \\
\mathcal{H}_1 &: \quad a[1]\neq 0, a[2]=0, \ldots, a[p_M]=0, \sigma_u^2>0 \\
&\vdots \\
\mathcal{H}_{p_M} &: \quad a[1]\neq 0, a[2]\neq 0, \ldots, a[p_M]\neq 0, \sigma_u^2>0
\end{aligned}$$

where p_M is the largest candidate model order. That is, for the AR process with order p, only the first p AR parameters are nonzero. Let $p(\mathbf{x}; \mathbf{a}_p, \sigma_u^2, \mathcal{H}_p)$ denote the PDF under $\mathcal{H}_p$, where $\mathbf{x}$ denotes the random process data vector and $\mathbf{a}_p$ is the $p \times 1$ vector of the first p nonzero AR filter parameters. Note that under $\mathcal{H}_0$, the AR process with order 0 is white Gaussian noise, so we write the PDF under $\mathcal{H}_0$ as $p(\mathbf{x}; \sigma_u^2, \mathcal{H}_0)$.

To estimate the order, we resort to the exponentially embedded families (EEF) which is a model order selection criterion that has been recently proposed [19]. It has been shown that asymptotically, the EEF minimizes the divergence between the true PDF and the estimated one. For each hypothesis $\mathcal{H}_p$, its EEF can be calculated by

$$EEF(p) = \begin{cases} l_{G_p}(\mathbf{x}) - p\left[\ln\left(\frac{l_{G_p}(\mathbf{x})}{p}\right) + 1\right] & \text{if } \frac{l_G(\mathbf{x})}{p} > 1 \\ 0 & \text{if } \frac{l_{G_p}(\mathbf{x})}{p} \leq 1 \end{cases} \tag{24}$$

where $l_{G_p}(\mathbf{x})$ is the generalized likelihood ratio for $\mathcal{H}_p$ [7] with

$$l_{G_p}(\mathbf{x}) = 2\ln\frac{p(\mathbf{x};\hat{\mathbf{a}}_p, \hat{\sigma}^2_{u_p}; \mathcal{H}_p)}{p(\mathbf{x};\hat{\sigma}^2_{u_0}, \mathcal{H}_0)} \tag{25}$$

Here $\hat{\mathbf{a}}_p, \hat{\sigma}^2_{u_p}$ are the maximum likelihood estimators (MLE) of $\mathbf{a}_p$ and σ^2_u under $\mathcal{H}_p$, and $\hat{\sigma}^2_{u_0}$ is the MLE of σ^2_u under $\mathcal{H}_0$. The EEF criterion chooses the hypothesis with the largest EEF value.

The PDF can be written in the frequency domain (and *hence the time series data can be replaced by the spectral data*) as [20]

$$\ln p(\mathbf{x};\mathbf{a}_p, \sigma^2_u, \mathcal{H}_p) = -\frac{N}{2}\ln 2\pi - \frac{N}{2}\int_0^1 \left[\ln P_p(f) + \frac{I(f)}{P_p(f)}\right] df \tag{26}$$

where $I(f)$ is the periodogram data and $P_p(f)$ is the true power spectral density (PSD) of the AR process with parameters $\mathbf{a}_p, \sigma^2_u$. Since

$$P_p(f) = \frac{\sigma^2_u}{|A_p(f)|^2}$$

the log-PDF can be written as

$$\begin{aligned} \ln p(\mathbf{x};\mathbf{a}_p, \sigma^2_u, \mathcal{H}_p) &= -\frac{N}{2}\ln 2\pi - \frac{N}{2}\int_0^1 \left[\ln\frac{\sigma^2_u}{|A_p(f)|^2} + \frac{I(f)}{\frac{\sigma^2_u}{|A_p(f)|^2}}\right] df \\ &= -\frac{N}{2}\ln 2\pi - \frac{N}{2}\int_0^1 \left[\ln\sigma^2_u + \frac{|A_p(f)|^2 I(f)}{\sigma^2_u}\right] df \end{aligned}$$

since it can be shown that $\int_0^1 \ln|A_p(f)|^2 df = 0$ [16]. Next we maximize the log-PDF over σ^2_u to obtain the MLE as

$$\hat{\sigma}^2_{u_p} = \int_0^1 |A_p(f)|^2 I(f) df.$$

and substituting back into $\ln p(\mathbf{x};\mathbf{a}_p, \hat{\sigma}^2_{u_p}, \mathcal{H}_p)$ yields

$$\ln p(\mathbf{x};\mathbf{a}_p, \hat{\sigma}^2_{u_p}, \mathcal{H}_p) = -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln\int_0^1 |A_p(f)|^2 I(f) df - \frac{N}{2}.$$

Finally, we need to maximize this over $\mathbf{a}_p$ to obtain $\hat{\mathbf{a}}_p$. It can be shown that this maximization requires one to use the Yule-Walker equations to estimate the AR filter parameters. Denoting the resultant MLE of $A_p(f)$ under $\mathcal{H}_p$ as $\hat{A}_p(f)$, we have that

$$\ln p(\mathbf{x};\hat{\mathbf{a}}_p, \hat{\sigma}^2_{u_p}, \mathcal{H}_p) = -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln\int_0^1 |\hat{A}_p(f)|^2 I(f) df - \frac{N}{2}$$

Note that since we have white Gaussian noise under $\mathcal{H}_0$, we have $A_0(f) = 1$. We maximize the log-PDF over σ_u^2 for $A_0(f) = 1$ to yield

$$\hat{\sigma}_{u_0}^2 = \int_0^1 I(f)df.$$

and hence we have

$$\ln p(\mathbf{x}; \hat{\sigma}_{u_0}^2, \mathcal{H}_0) \quad = \quad -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln \int_0^1 I(f)df - \frac{N}{2}$$

As a result,

$$l_{G_p}(\mathbf{x}) = 2\ln \frac{p(\mathbf{x}; \hat{\mathbf{a}}_p, \hat{\sigma}_{u_p}^2; \mathcal{H}_p)}{p(\mathbf{x}; \hat{\sigma}_{u_0}^2, \mathcal{H}_0)} = -N \ln \frac{\int_0^1 |\hat{A}_p(f)|^2 I(f)df}{\int_0^1 I(f)df} \tag{27}$$

When this is discretized over the band $0 \leq f \leq 1/2$ we have

$$l_{G_p}(\mathbf{x}) = -N \ln \frac{\sum_{k=1}^{N_f} |\hat{A}_p(f_k)|^2 I(f_k)\Delta f}{\sum_{k=1}^{N_f} I(f_k)\Delta f} \tag{28}$$

Finally we choose the AR model with the largest EEF calculated by (24).

Appendix B

Derivation of Test Statistic for Detection

To begin, we assume that the background random process (in the time domain) is a real-valued Gaussian AR process with parameters $\{a_B[1], a_B[2], \ldots, a_B[p], \sigma_u^2\}$. Note that the parameters $\{a_B[1], a_B[2], \ldots, a_B[p], \sigma_u^2\}$ and also the order p is estimated using the sample average of the background spectral data as in (8). Under $\mathcal{H}_0$, which is background only, the AR filter parameters are assumed to be known but not the excitation noise variance. Under $\mathcal{H}_1$, the AR filter parameters and the excitation noise variance are both unknown. Let q be the estimated AR model order under $\mathcal{H}_1$ using the observed spectral data $I(f)$. Then, we set up the hypothesis test

$$\begin{aligned} \mathcal{H}_0 \quad &: \quad \text{AR parameters are } a_B[1], a_B[2], \ldots, a_B[p], \sigma_u^2 > 0 \\ \mathcal{H}_1 \quad &: \quad \text{AR parameters are } a[1], a[2], \ldots, a[q], \sigma_u^2 > 0 \end{aligned}$$

This effectively says that under $\mathcal{H}_0$ (no signal present) the spectrum is just the known background spectrum, although with an unspecified σ_u^2. Under $\mathcal{H}_1$ the *shape* of the spectrum is changed due to the change in the AR filter parameters. This is caused by the presence of a signal, which has been added to the background. We also assume that the fictitious N time samples $\{x[0], x[1], \ldots, x[N-1]\}$ of the Gaussian AR random process are observed. Let $p(\mathbf{x}; \mathbf{a}_B, \sigma_u^2, \mathcal{H}_0)$ denote the PDF under $\mathcal{H}_0$ and $p(\mathbf{x}; \mathbf{a}, \sigma_u^2, \mathcal{H}_1)$ denote the PDF under $\mathcal{H}_1$, where $\mathbf{x}$ denotes the random process data vector, $\mathbf{a}_B$ is the known $p \times 1$ vector of AR filter parameters and $\mathbf{a}$ is the unknown $q \times 1$ vector of AR filter parameters. The generalized likelihood ratio test statistic (GLRT) is [7]

$$l_G(\mathbf{x}) = \ln \frac{p(\mathbf{x}; \hat{\mathbf{a}}, \hat{\sigma}^2_{u_1}; \mathcal{H}_1)}{p(\mathbf{x}; \mathbf{a}_B, \hat{\sigma}^2_{u_0}, \mathcal{H}_0)} \tag{29}$$

where $\hat{\mathbf{a}}, \hat{\sigma}^2_{u_1}$ is the maximum likelihood estimator (MLE) of $\mathbf{a}$ and σ^2_u under $\mathcal{H}_1$, and $\hat{\sigma}^2_{u_0}$ is the MLE of σ^2_u under $\mathcal{H}_0$.

From similar derivations as in Appendix A, and denoting the resultant MLE of $A(f)$ under $\mathcal{H}_1$ as $\hat{A}_T(f)$, we have that

$$\begin{aligned} \ln p(\mathbf{x}; \mathbf{a}_B, \hat{\sigma}^2_{u_0}, \mathcal{H}_0) &= -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln \int_0^1 |A_B(f)|^2 I(f) df - \frac{N}{2} \\ \ln p(\mathbf{x}; \hat{\mathbf{a}}_1, \hat{\sigma}^2_{u_1}, \mathcal{H}_1) &= -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln \int_0^1 |\hat{A}_T(f)|^2 I(f) df - \frac{N}{2} \end{aligned}$$

and finally from (29)

$$l_G(\mathbf{x}) = \frac{N}{2} \ln \frac{\int_0^1 |A_B(f)|^2 I(f) df}{\int_0^1 |\hat{A}_T(f)|^2 I(f) df}.$$

When this is discretized over the band $0 \le f \le 1/2$ we have

$$l_G(\mathbf{x}) = \frac{N}{2} \ln \frac{\sum_{k=1}^{N_f} |A_B(f_k)|^2 I(f_k)}{\sum_{k=1}^{N_f} |\hat{A}_T(f_k)|^2 I(f_k)}.$$

and omitting the $N/2$ factor, we have finally (11).

APPENDIX C
DERIVATION OF PROBABILITY OF DETECTION STATISTIC THRESHOLD CROSSING FOR GIVEN FALSE ALARM RATE

We declare that a chemical has been detected if at least 10% of the pulses in a given block produce threshold crossings. As an example, we consider the block to consist of 100 pulses and hence a detection occurs if at least 10 threshold crossings are observed. Also, we assume an operational requirement of one false alarm per two hours of time. To compute the desired probabilities exactly is difficult due to the fact that the successive blocks, which differ by only one sample, are heavily dependent. As an approximation we assume that the blocks are overlapped by 50% (which may be necessary in practice to avoid excessive computation and is commonly done in practice) and therefore that the data in each block is *approximately* independent. Then, in two hours we have examined

$$L = (2 \times 3600 \times 25)/50 = 3600 \text{ blocks}$$

for a 10% threshold crossing rate. Hence, the probability of false alarm for each block is obtained as follows. Let P_{FA_b} be the probability of a false alarm for a single block. Then, the probability of *at most* one false alarm in L *independent* blocks is

$$\begin{aligned} P_1 = P[\text{at most one false alarm in } L \text{ blocks}] &= P[\text{no false alarms in } L \text{ blocks}] \\ &\quad + P[\text{one false alarm in } L \text{ blocks}] \\ &= (1 - P_{FA_b})^L + L P_{FA_b} (1 - P_{FA_b})^{L-1} \end{aligned}$$

since this is a binomial type of probability. We want the probability of at most one false alarm per two hours to be large, say 0.99. Hence, we need to solve for P_{FA_b} by finding that value that satisfies

$$(1 - P_{FA_b})^L + L P_{FA_b} (1 - P_{FA_b})^{L-1} = 0.99.$$

In general, for at most one false alarm per h hours we should use $L = 1800h$. For the example of $h = 2$ we plot the probability of at most one false alarm per two hours versus P_{FA_b} in Figure 17. It is seen that we should require that $P_{FA_b} = 4 \times 10^{-5}$, which is the probability of a false alarm for a single block of 100 pulses. Next, since we declare a chemical present if there are at least 10 threshold crossings out of a 100 possible ones, then the probability of a false alarm for a single block is

$$P_{FA_b} = \sum_{k=10}^{100} \binom{100}{k} P_{FA_p}^k (1 - P_{FA_p})^{100-k} = 1 - \sum_{k=0}^{9} \binom{100}{k} P_{FA_p}^k (1 - P_{FA_p})^{100-k}$$

where P_{FA_p} is the probability of a threshold crossing, i.e., probability of a false alarm for a single pulse. In Figure 18 we plot P_{FA_b} versus P_{FA_p}. For $P_{FA_b} = 4 \times 10^{-5} = -44$ dB, we require from Figure 18 that $P_{FA_p} = 0.02$. Hence, the threshold of the test statistic given by (11) should be set so that the probability of T_D exceeding this threshold γ is 0.02.

Appendix D
Derivation of LMP Test Statistic for Classification

It is assumed that one of M chemicals is present. The spectral data is assumed to be of the form $P(f) = \theta_i P_{s_i}(f) + P_B(f)$ if the ith chemical is present. As usual $P_B(f)$ is the PSD of the background, $P_{s_i}(f)$ is the known spectral signature for the ith chemical, and θ_i is an unknown scaling factor that accounts for the unknown power of the chemical. To decide which chemical is present we set up an

M-ary hypothesis test as

$$\begin{aligned}
\mathcal{H}_1 &: P(f) = \theta_1 P_{s_1}(f) + P_B(f) \\
\mathcal{H}_2 &: P(f) = \theta_2 P_{s_2}(f) + P_B(f) \\
\vdots &\;\vdots \\
\mathcal{H}_M &: P(f) = \theta_M P_{s_M}(f) + P_B(f).
\end{aligned}$$

The θ_i's are positive but otherwise unknown. They are assumed to be small so that an LMP approach can be used. An LMP classification test statistic decides chemical k is present if among

$$T_{C_i}(\mathbf{x}) = \frac{\left.\frac{\partial \ln p(\mathbf{x};\mathcal{H}_i)}{\partial \theta_i}\right|_{\theta_i=0}}{\sqrt{I_{F_i}(0)}} \tag{30}$$

for $i = 1, 2, \ldots, M$, $T_{C_k}(\mathbf{x})$ is the maximum value. In (30) $I_F(0)$ is the Fisher information for θ when evaluated at $\theta = 0$. To evaluate the test statistics, we first note that the log-PDF of the spectral data is given as in Appendix A as

$$\ln p(\mathbf{x}) = -\frac{N}{2}\ln 2\pi - \frac{N}{2}\int_0^1 \left[\ln P(f) + \frac{I(f)}{P(f)}\right] df.$$

Using $P(f) = \theta P_s(f) + P_B(f)$ we have

$$\ln p(\mathbf{x}) = -\frac{N}{2}\ln 2\pi - \frac{N}{2}\int_0^1 \left[\ln\left(\theta P_s(f) + P_B(f)\right) + \frac{I(f)}{\theta P_s(f) + P_B(f)}\right] df$$

and differentiating produces

$$\frac{\partial \ln p(\mathbf{x})}{\partial \theta} = -\frac{N}{2}\int_0^1 \frac{P_s(f)}{\theta P_s(f) + P_B(f)} - \frac{I(f)P_s(f)}{(\theta P_s(f) + P_B(f))^2} df \tag{31}$$

which when evaluated at $\theta = 0$ yields

$$\begin{aligned}
\left.\frac{\partial \ln p(\mathbf{x})}{\partial \theta}\right|_{\theta=0} &= -\frac{N}{2}\int_0^1 \frac{P_s(f)}{P_B(f)} - \frac{I(f)P_s(f)}{(P_B(f))^2} df \\
&= \frac{N}{2}\int_0^1 \frac{P_s(f)}{P_B(f)}\left(\frac{I(f)}{P_B(f)} - 1\right) df. \qquad (32)
\end{aligned}$$

To determine the Fisher information we differentiate (31) a second time to produce

$$\frac{\partial^2 \ln p(\mathbf{x})}{\partial \theta^2} = -\frac{N}{2}\int_0^1 -\frac{P_s^2(f)}{(\theta P_s(f) + P_B(f))^2} + 2\frac{I(f)P_s^2(f)}{(\theta P_s(f) + P_B(f))^3} df.$$

Taking the expected value and noting that $E[I(f)] = P(f) = \theta P_s(f) + P_B(f)$ produces

$$\begin{aligned}
E\left[\frac{\partial^2 \ln p(\mathbf{x})}{\partial \theta^2}\right] &= -\frac{N}{2}\int_0^1 -\frac{P_s^2(f)}{(\theta P_s(f) + P_B(f))^2} + 2\frac{(\theta P_s(f) + P_B(f))P_s^2(f)}{(\theta P_s(f) + P_B(f))^3} df \\
&= -\frac{N}{2}\int_0^1 -\frac{P_s^2(f)}{(\theta P_s(f) + P_B(f))^2} + 2\frac{P_s^2(f)}{(\theta P_s(f) + P_B(f))^2} df.
\end{aligned}$$

Setting $\theta = 0$ and taking the negative produces

$$I_F(0) = \frac{N}{2} \int_0^1 \frac{P_s^2(f)}{P_B^2(f)} df. \tag{33}$$

Therefore, the LMP statistic becomes from (32) and (33)

$$T_C = \frac{\sqrt{\frac{N}{2}} \int_0^1 \frac{P_s(f)}{P_B(f)} \left(\frac{I(f)}{P_B(f)} - 1 \right) df}{\sqrt{\int_0^1 \frac{P_s^2(f)}{P_B^2(f)} df}}. \tag{34}$$

When discretized over the band $0 \leq f \leq 1/2$, this becomes

$$\frac{\sqrt{\frac{N}{2}} \sum_{k=1}^{N_f} \frac{P_s(f_k)}{P_B(f_k)} \left(\frac{I(f_k)}{P_B(f_k)} - 1 \right) \Delta f}{\sqrt{\sum_{k=1}^{N_f} \frac{P_s^2(f_k)}{P_B^2(f_k)} \Delta f}}$$

and ignoring a scaling factor, which will not affect the maximum, we have finally

$$T_C = \frac{\sum_{k=1}^{N_f} \frac{P_s(f_k)}{P_B(f_k)} \left(\frac{I(f_k)}{P_B(f_k)} - 1 \right)}{\sqrt{\sum_{k=1}^{N_f} \frac{P_s^2(f_k)}{P_B^2(f_k)}}}$$

Appendix E

Derivation of The Asymptotic Likelihood Function Method for Classification of Mixture of Chemicals

We assume that K out of M chemicals are present and they are additive. Hence, the spectral data is of the form $P(f) = \sum_{i=1}^{K} \theta_{k_i} P_{s_{k_i}}(f) + P_B(f)$ if chemicals $k_1, k_2, \ldots, k_K$ are present. The total number of candidate hypotheses is $\binom{M}{K}$. Let the unknown parameters be $\boldsymbol{\theta} = [\theta_{k_1}, \theta_{k_2}, \ldots, \theta_{k_K}]^T$. Asymptotically [7],

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + \mathbf{I}^{-1}(\boldsymbol{\theta}_0) \left. \frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}_0}$$

or in our problem $\boldsymbol{\theta}_0 = \mathbf{0}$,

$$\hat{\boldsymbol{\theta}} = \mathbf{I}^{-1}(\mathbf{0}) \left. \frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta} = \mathbf{0}}. \tag{35}$$

For each candidate hypothesis,

$$\begin{aligned} \ln p(\mathbf{x}; \boldsymbol{\theta}) \quad = \quad & -\frac{N}{2} \int_0^1 \left[\ln \left(\sum_{i=1}^{K} \theta_{k_i} P_{s_{k_i}}(f) + P_B(f) \right) + \frac{I(f)}{\sum_{i=1}^{K} \theta_{k_i} P_{s_{k_i}}(f) + P_B(f)} \right] df \\ & -\frac{N}{2} \ln(2\pi) \end{aligned} \tag{36}$$

and

$$\frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_{k_i}} = -\frac{N}{2} \int_0^1 \left[\frac{P_{s_{k_i}}(f)}{\sum_{i=1}^{K} \theta_{k_i} P_{s_{k_i}}(f) + P_B(f)} - \frac{I(f) P_{s_{k_i}}(f)}{\left(\sum_{i=1}^{K} \theta_{k_i} P_{s_{k_i}}(f) + P_B(f) \right)^2} \right] df$$

$$\left.\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \theta_{k_i}}\right|_{\boldsymbol{\theta}=\mathbf{0}} = \frac{N}{2}\int_0^1 \frac{P_{s_{k_i}}(f)}{P_B(f)}\left(\frac{I(f)}{P_B(f)}-1\right)df. \tag{37}$$

The second derivative is

$$\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \theta_{k_i}\partial \theta_{k_j}} = -\frac{N}{2}\int_0^1 \left[-\frac{P_{s_{k_i}}(f)P_{s_{k_j}}(f)}{\left(\sum_{i=1}^K \theta_{k_i}P_{s_{k_i}}(f)+P_B(f)\right)^2} + \frac{2I(f)P_{s_{k_i}}(f)P_{s_{k_j}}(f)}{\left(\sum_{i=1}^K \theta_{k_i}P_{s_{k_i}}(f)+P_B(f)\right)^3}\right]df$$

and therefore, the Fisher information matrix is

$$\begin{aligned}
\mathbf{I}_{ij}(\boldsymbol{\theta}) &= -E\left[\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \theta_{k_i}\partial \theta_{k_j}}\right] \\
&= \frac{N}{2}E\left[\int_0^1 \left[-\frac{P_{s_{k_i}}(f)P_{s_{k_j}}(f)}{\left(\sum_{i=1}^K \theta_{k_i}P_{s_{k_i}}(f)+P_B(f)\right)^2} + \frac{2I(f)P_{s_{k_i}}(f)P_{s_{k_j}}(f)}{\left(\sum_{i=1}^K \theta_{k_i}P_{s_{k_i}}(f)+P_B(f)\right)^3}\right]df\right] \\
&= \frac{N}{2}\int_0^1 \left[-\frac{P_{s_{k_i}}(f)P_{s_{k_j}}(f)}{\left(\sum_{i=1}^K \theta_{k_i}P_{s_{k_i}}(f)+P_B(f)\right)^2} + \frac{2E\left(I(f)\right)P_{s_{k_i}}(f)P_{s_{k_j}}(f)}{\left(\sum_{i=1}^K \theta_{k_i}P_{s_{k_i}}(f)+P_B(f)\right)^3}\right]df \\
&= \frac{N}{2}\int_0^1 \left[-\frac{P_{s_{k_i}}(f)P_{s_{k_j}}(f)}{\left(\sum_{i=1}^K \theta_{k_i}P_{s_{k_i}}(f)+P_B(f)\right)^2} + \frac{2P_{s_{k_i}}(f)P_{s_{k_j}}(f)}{\left(\sum_{i=1}^K \theta_{k_i}P_{s_{k_i}}(f)+P_B(f)\right)^2}\right]df \\
&= \frac{N}{2}\int_0^1 \frac{P_{s_{k_i}}(f)P_{s_{k_j}}(f)}{\left(\sum_{i=1}^K \theta_{k_i}P_{s_{k_i}}(f)+P_B(f)\right)^2}df
\end{aligned}$$

or

$$\mathbf{I}_{ij}(\mathbf{0}) = \frac{N}{2}\int_0^1 \frac{P_{s_{k_i}}(f)P_{s_{k_j}}(f)}{\left(P_B(f)\right)^2}df. \tag{38}$$

When discretized over the band $0 \le f \le 1/2$, (36), (37) and (38) become

$$\begin{aligned}
\ln p(\mathbf{x};\boldsymbol{\theta}) &= -\frac{N}{2}\sum_{k=1}^{N_f}\left[\ln\left(\sum_{i=1}^K \theta_{k_i}P_{s_{k_i}}(f_k)+P_B(f_k)\right) + \frac{I(f_k)}{\sum_{i=1}^K \theta_{k_i}P_{s_{k_i}}(f_k)+P_B(f_k)}\right]\Delta f \\
&\quad -\frac{N}{2}\ln(2\pi)
\end{aligned} \tag{39}$$

$$\left.\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \theta_{k_i}}\right|_{\boldsymbol{\theta}=\mathbf{0}} = \frac{N}{2}\sum_{k=1}^{N_f} \frac{P_{s_{k_i}}(f_k)}{P_B(f_k)}\left(\frac{I(f_k)}{P_B(f_k)}-1\right)\Delta f. \tag{40}$$

$$\mathbf{I}_{ij}(\mathbf{0}) = \frac{N}{2}\sum_{k=1}^{N_f} \frac{P_{s_{k_i}}(f_k)P_{s_{k_j}}(f_k)}{\left(P_B(f_k)\right)^2}\Delta f. \tag{41}$$

Now we have the MLE of $\boldsymbol{\theta}$ from (35), (40) and (41). The asymptotic likelihood function approach then substitutes $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ into (39) and chooses the hypothesis that has the largest likelihood.

One important issue that should be described is our assumption that $\theta_{k_i} \geq 0$ for $i = 1, 2, \ldots, K$. However, the MLE of $\boldsymbol{\theta}$ without these nonnegative constraints may produce negative solutions. However, this can be easily resolved by the Kuhn-Tucker conditions.

From the Kuhn-Tucker conditions, we know that if the MLE without those positivity constraints has negative solutions, then the MLE under these constraints will have at least one $\theta_{k_i} = 0$ [25]. Then this hypothesis is reduced to at least the $(K-1)$th order model. Then, any other Kth order hypothesis that has the same chemical signatures as the reduced $(K-1)$th order hypothesis and one arbitrary other chemical signature would have likelihood not less than the reduced $(K-1)$th order hypothesis. For example, for the hypothesis $\mathcal{H}_1$ that has chemical signatures $P_1(f), P_2(f), \ldots, P_K(f)$, if the unconstrained MLE of $\boldsymbol{\theta}$ has negative solutions, then the MLE under positivity constraints would have at least one $\theta_i = 0$ by the Kuhn-Tucker conditions, say $\theta_1 = 0$. Thus, any other hypothesis that includes $P_2(f), P_3(f), \ldots, P_K(f)$ and any other chemical signature (say $P_{s_1}(f)$) would have likelihood not less than $\mathcal{H}_1$, since we would at least get the same likelihood by using the same constrained MLE for hypothesis $\mathcal{H}_1$. This argument implies that we can just ignore the hypothesis that yields an unconstrained MLE with at least one negative solution, and therefore, the greatest likelihood must correspond to the hypothesis with a nonnegative unconstrained MLE.

Since we have as many as $\binom{M}{K}$ candidate hypotheses, we do not have to consider the case when all hypotheses have at least one negative solution in unconstrained MLE. In this case, it can be considered that there are less than K chemicals present, and we should decrease the value of K.

References

[1] S. Kay, C. Xu, and D. Emge, "Chemical detection and classification in raman spectra," in *Proceedings of the SPIE*, Mar. 2008, vol. 6969, pp. 4–12.

[2] K. Kneipp, H. Kneipp, I. Itzkan, R.R. Dasari, and M.S. Feld, "Ultrasensitive chemical analysis by raman spectroscopy," *Chemical Reviews*, vol. 99, pp. 2957C2975, 1999.

[3] R.L. Frost, D.A. Henry, and K.L. Erickson, "Raman spectroscopic detection of wyartite in the presence of rabejacite," *Journal of Raman Spectroscopy*, vol. 35, pp. 255–260, 2004.

[4] N. Hayazawa, M. Motohashi, Y. Saito, and S. Kawata, "Highly sensitive strain detection in strained silicon by surface-enhanced raman spectroscopy," *Applied Physics Letters*, vol. 86, pp. 263114 – 263114–3, 2005.

[5] A. Portnov, S. Rosenwaks, and I. Bar, "Detection of particles of explosives via backward coherent anti-stokes raman spectroscopy," *Applied Physics Letters*, vol. 93, pp. 041115 – 041115–3, 2008.

[6] D. Manolakis, D. Marden, and G.A. Shaw, "Hyperspectral image processing for automatic target detection applications," *Lincoln Laboratory Journal*, vol. 14, no. 1, pp. 79–116, 2003.

[7] S. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1998.

[8] L.L. Scharf and B. Friedlander, “Matched subspace detectors,” *IEEE Trans. Signal Process.*, vol. 42, no. 8, pp. 2146–2157, Aug. 1994.

[9] W. Wang and T. Adali, “Constrained ica and its application to raman spectroscopy,” in *Proc. Antennas and Propagation Society International Symposium*, Jul. 2005, pp. 109–112.

[10] W. Wang, T. Adali, and D. Emge, “Unsupervised detection using canonical correlation analysis and its application to raman spectroscopy,” in *Proc. IEEE Workshop on Machine Learning for Signal Processing*, Aug. 2007.

[11] W. Wang, T. Adali, and D. Emge, “Subspace partitioning for target detection and identification,” *IEEE Trans. Signal Process.*, vol. 57, no. 4, pp. 1250–1259, Apr. 2009.

[12] M.S. Alam, M. Nazrul Islam, A. Bal, and M.A. Karim, “Hyperspectral target detection using gaussian filter and post-processing,” *Optics and Lasers in Engineering*, vol. 46, pp. 817–822, Nov. 2008.

[13] T.H. Chyba, N.S. Higdon, W.T. Armstrong, C.T. Lobb, P.L. Ponsardin, D.A. Richter, B.T. Kelly, Q. Bui, R. Babnick, M.K. Boysworth, A.J. Sedlacek, and S.D. Christesen, “Field tests of the laser interrogation of surface agents (lisa) system for on-the-move standoff sensing of chemical agents,” in *Proc. Int. Symp. Spectral Sensing Research*, 2003.

[14] W. Knight, R. Pridham, and S. Kay, “Digital signal processing for sonar,” in *Proceedings of the IEEE*, Nov. 1981, pp. 1451–1506.

[15] R.G. Wiley, *ELINT: The Interception and Analysis of Radar Signals*, Artech House, Boston, MA, 2006.

[16] S. Kay, *Modern Spectral Estimation: Theory and Application*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[17] D. Bowyer, P. Rajasekaran, and W. Gebhart, “Adaptive clutter filtering using autoregressive spectral estimation,” *IEEE Trans. Aerosp. Electron. Syst.*, pp. 538–546, Jul. 1979.

[18] S. Kay and J. Salisbury, “Improved active sonar detection using autoregressive prewhiteners,” *J. Acoustical Soc. of America*, pp. 1603–1611, Apr. 1990.

[19] S. Kay, “Exponentially embedded families - new approaches to model order estimation,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 41, pp. 333–345, Jan. 2005.

[20] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[21] A. Pages-Zamora and M.A. Lagunas, “New approaches in non-linear signal processing: Estimation of the probability density function by spectral estimation methods,” in *IEEE Workshop on Higher Order Statistics*, 1995.

[22] S. Kay, “Model based probability density function estimation,” *IEEE Signal Process. Lett.*, pp. 318–320, Dec. 1998.

[23] J. Rissanen, “Modeling by shortest data description,” *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[24] G. Schwarz, “Estimating the dimension of a model,” *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[25] C.L. Lawson and R.J. Hanson, *Solving Least Squares Problems*, SIAM, 1995.

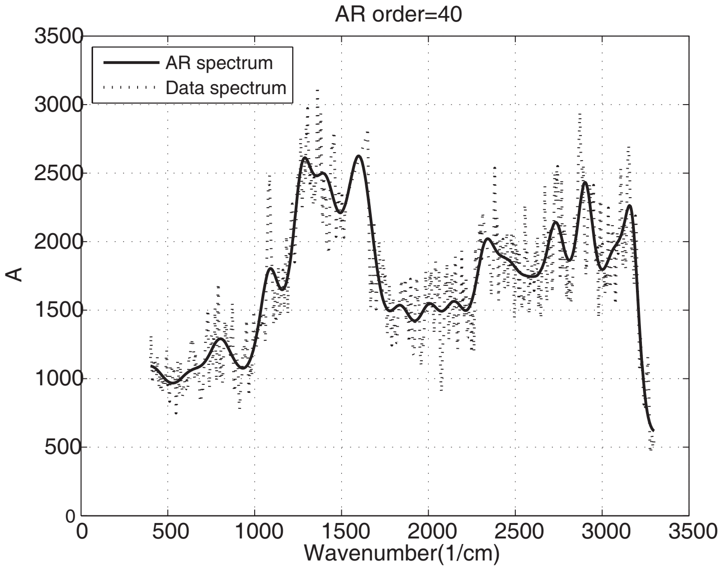

Fig. 1. AR spectral estimate and background spectral data for asphalt surface ($F_c = 3300$).

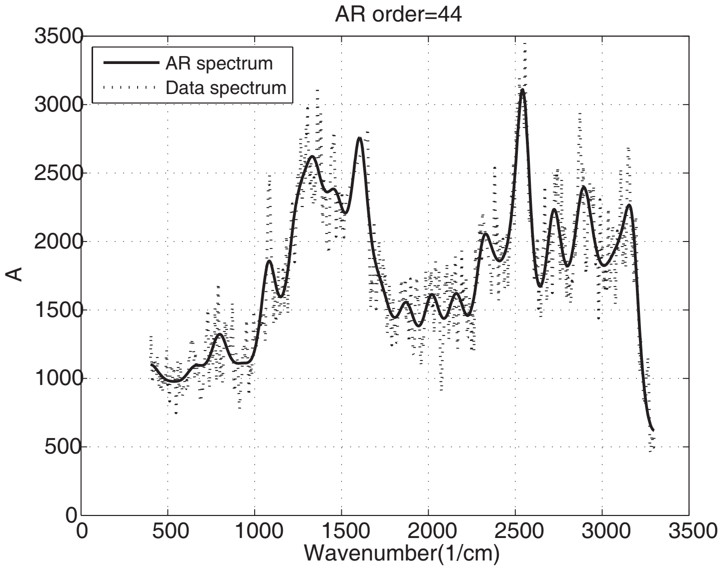

Fig. 2. AR spectrum for asphalt surface plus an artificial signal ($F_c = 3300$).

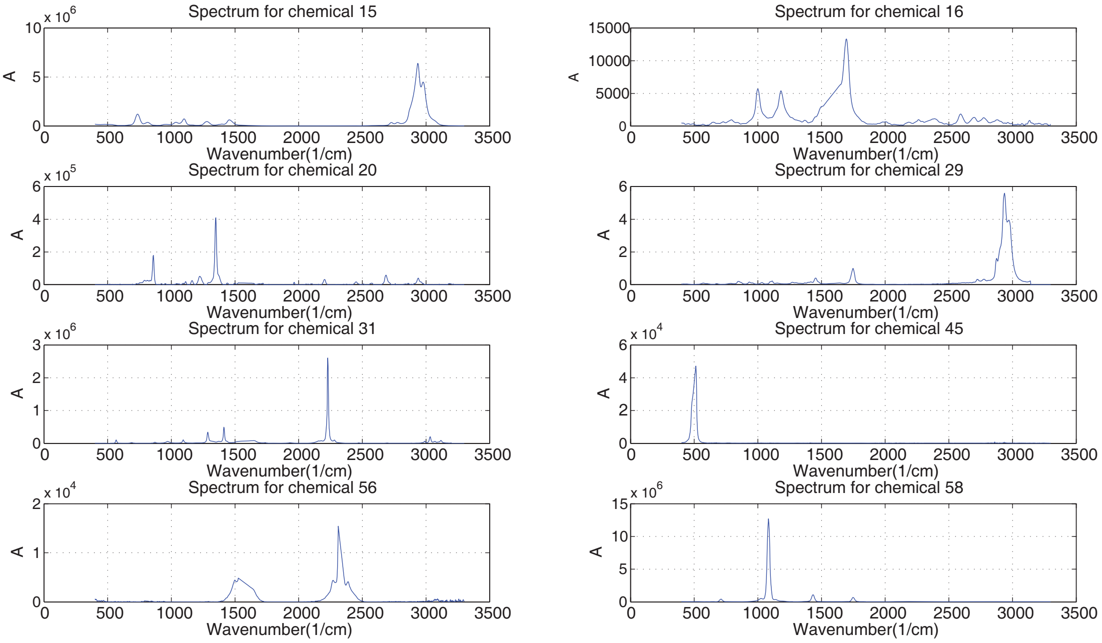

Fig. 3. Spectra of the chemicals that are used in simulations.

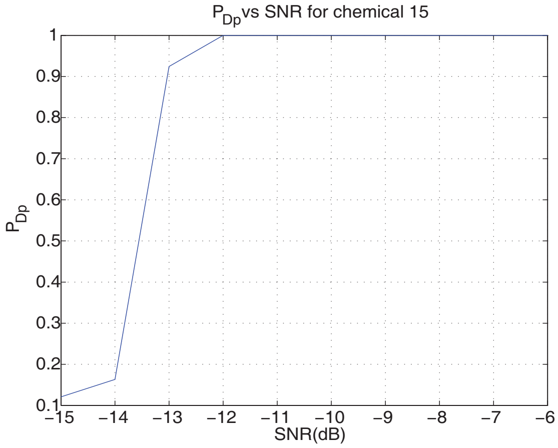

Fig. 4. Probability P_{D_p} of detecting chemical 15 versus SNR based on a single pulse.

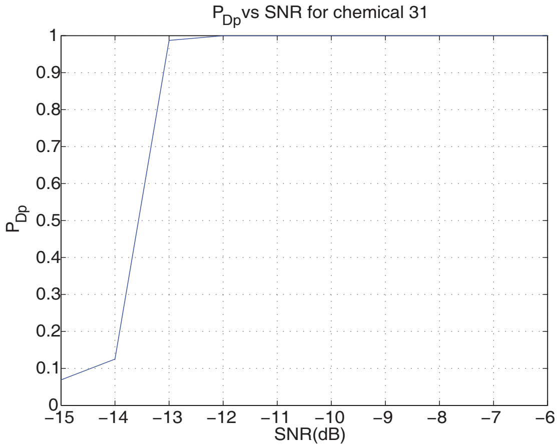

Fig. 5. Probability P_{D_p} of detecting chemical 31 versus SNR based on a single pulse.

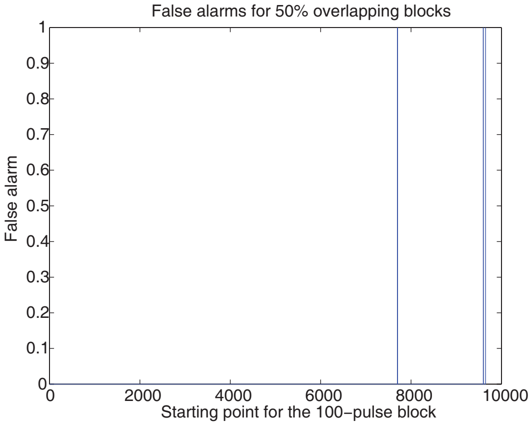

Fig. 6. False alarms for a concrete background and fixed threshold.

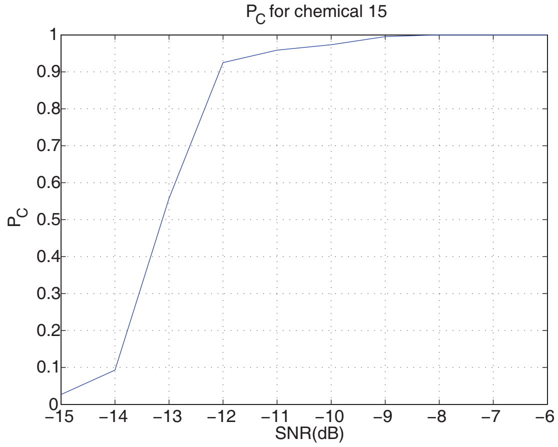

Fig. 7. Probability of correct single pulse classification versus SNR. Chemical 15 is present.

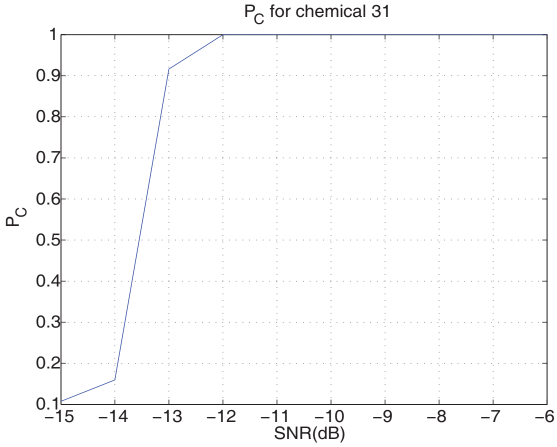

Fig. 8. Probability of correct single pulse classification versus SNR. Chemical 31 is present.

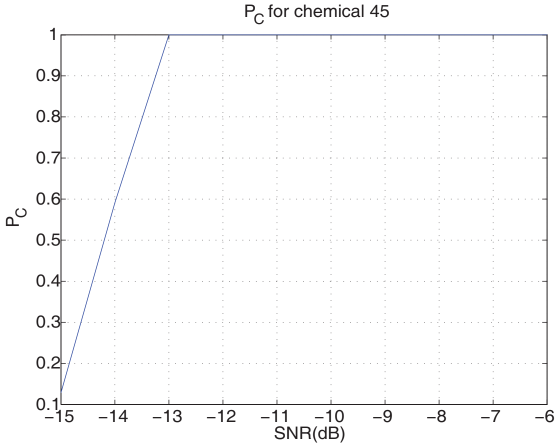

Fig. 9. Probability of correct single pulse classification versus SNR. Chemical 45 is present.

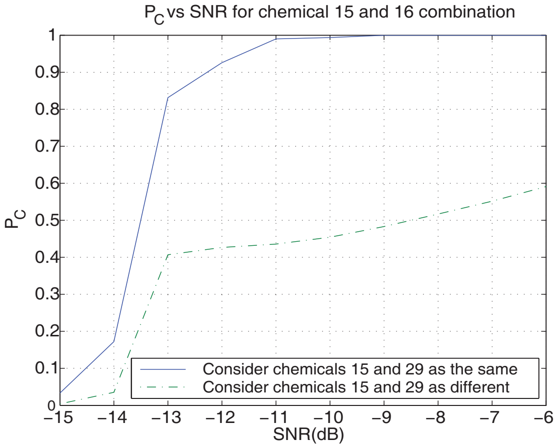

Fig. 10. Probability of correct single pulse classification versus SNR. Chemicals 15 and 16 are present.

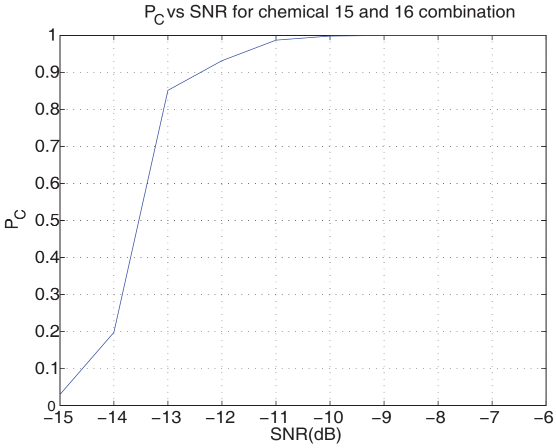

Fig. 11. Probability of correct single pulse classification versus SNR. Chemicals 15 and 16 are present. Chemical 29 is removed from the library.

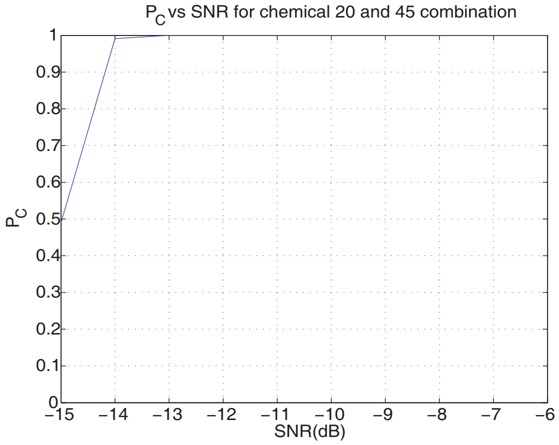

Fig. 12. Probability of correct single pulse classification versus SNR. Chemicals 20 and 45 are present.

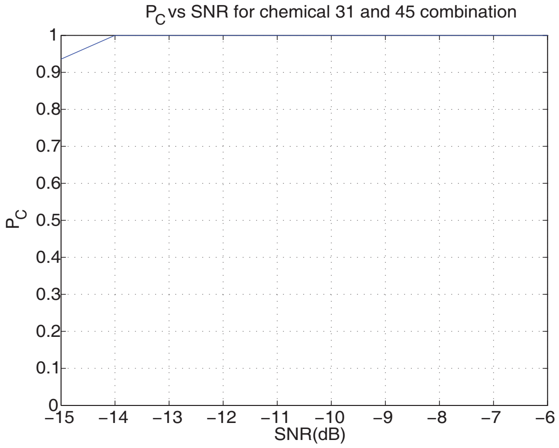

Fig. 13. Probability of correct single pulse classification versus SNR. Chemicals 31 and 45 are present.

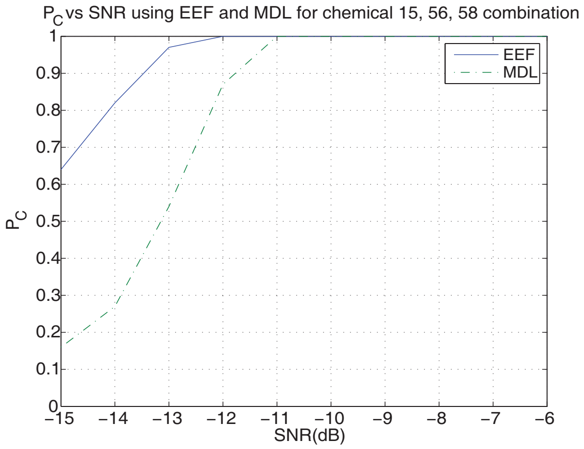

Fig. 14. Probability of correct single pulse classification versus SNR using EEF and MDL. Chemicals 15, 56 and 58 are present.

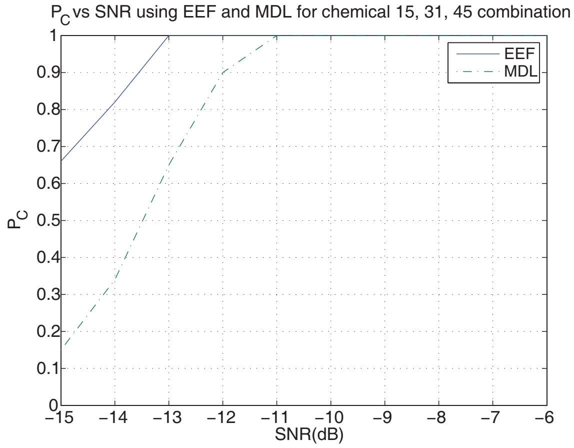

Fig. 15. Probability of correct single pulse classification versus SNR using EEF and MDL. Chemicals 15, 31 and 45 are present.

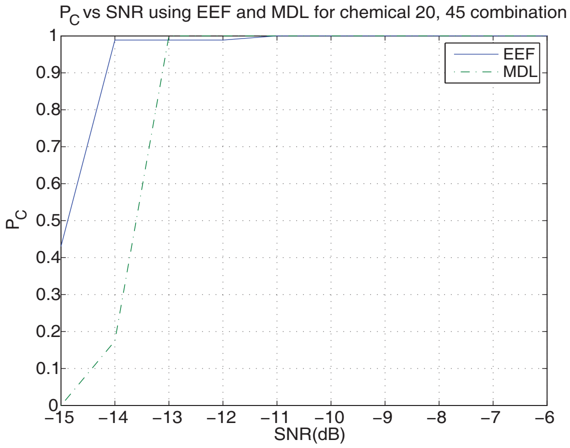

Fig. 16. Probability of correct single pulse classification versus SNR using EEF and MDL. Chemicals 20 and 45 are present.

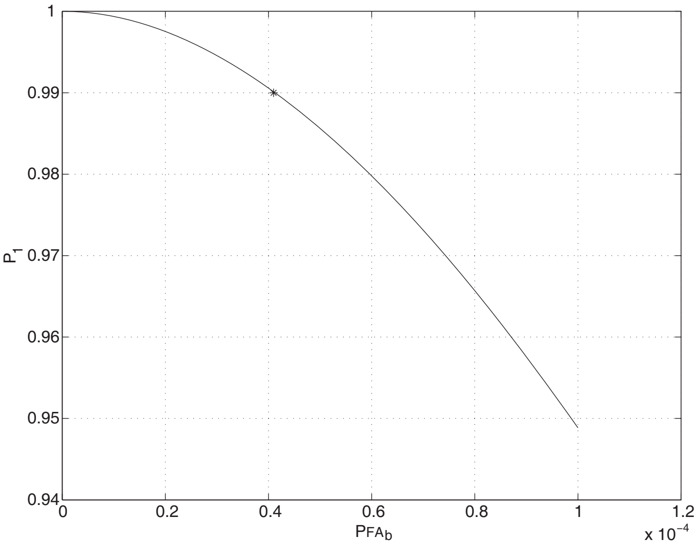

Fig. 17. Probability P_1 of at most one false alarm per two hours versus P_{FA_b}.

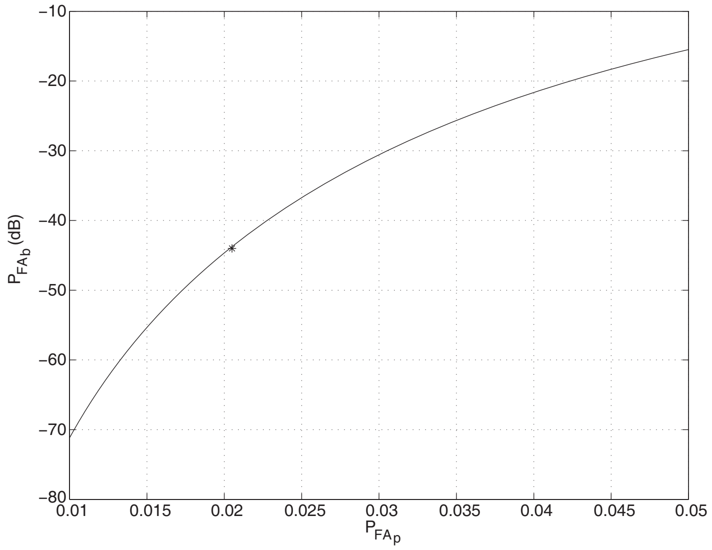

Fig. 18. Probability of at most one false alarm per two hours versus P_{FA_p}.