Naval Research Laboratory

Washington, DC 20375-5320

NRL/MR/6180--13-9509

Information Measures for Multisensor Systems

Christian P. Minor
Joseph C. Gezo

Nova Research, Inc.
Alexandria, Virginia

Kevin J. Johnson

Navy Technology Center for Safety and Survivability
Chemistry Division

December 11, 2013

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)* 11-12-2013	**2. REPORT TYPE** Memorandum Report	**3. DATES COVERED** *(From - To)* January 2013 – April 2013

4. TITLE AND SUBTITLE

Information Measures for Multisensor Systems

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Christian P. Minor,[1] Joseph C. Gezo,[1] and Kevin J. Johnson

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

61-9010-0-2-5

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Research Laboratory, Code 6181
4555 Overlook Avenue, SW
Washington, DC 20375-5320

8. PERFORMING ORGANIZATION REPORT NUMBER

NRL/MR/6180--13-9509

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Defense Threat Reduction Agency (DTRA)
Attn: Ngai Wong
8725 John J. Kingman Road
Ft. Belvoir, VA 22060-6201

10. SPONSOR / MONITOR'S ACRONYM(S)

DTRA

11. SPONSOR / MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

[1]Nova Research, Inc., 1900 Elkin Street, Suite 230, Alexandria, VA 22308

14. ABSTRACT

The purpose of this report is to demonstrate the utility of an information-theoretic approach to next generation chemical detection. Recent research at the Naval Research Laboratory (NRL) has yielded probabilistic models for spectral data that enable the computation of information measures such as entropy and divergence, with the goal of developing feature sets to increase the sensitivity and selectivity of multivariate chemical sensors of several modalities. Results are presented for several types of spectral data in multisensor systems, as well as strategies for using information measures with other data sources. Binary, univariate, and multivariate sensors can all be modeled from an information-theoretic perspective, making it well-suited for the challenges of next generation chemical detection.

15. SUBJECT TERMS

Chemical sensing	Information entropy	Infrared spectroscopy
Information theory	Information divergence	Multisensor
Spectral data	Mass spectrometry	

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Kevin J. Johnson
a. REPORT Unclassified Unlimited	**b. ABSTRACT** Unclassified Unlimited	**c. THIS PAGE** Unclassified Unlimited	Unclassified Unlimited	23	**19b. TELEPHONE NUMBER** *(include area code)* (202) 404-5407

Contents

Figures

Executive Summary

The purpose of this report is to demonstrate the utility of an information-theoretic approach to next generation chemical detection. Research and development of chemical sensing systems that are effective for a variety of field environments outside the laboratory is an ongoing challenge. Given the requirements for discerning potentially hundreds to thousands of target analytes from an even larger number of background contaminants, sensors capable of generating multivariate data – such as spectral sensing systems – have become a central focus for new system designs due to their prior successes in the laboratory.

Spectral data capture a wealth of chemical information in a variety of sensing modalities. While sharing a common multivariate structure, spectral data often exhibit characteristics that can be challenging to model. Traditional statistical methods often focus on typical or most frequent values in the data. But spectral data tend to be sparse, characterized instead by large values in peaks that occur infrequently. It is sparseness that gives spectral data their robustness for chemical identification.

Information theory provides an alternate framework for modeling sparse spectral data. Information measures are methods that can quantify how effective chemical spectra are at discriminating between hypotheses such as potential target analytes, even if those spectra arise from different sensing modalities. Providing additional flexibility, information measures operate on probability distributions that can be developed directly from data themselves, when available, or from statistical models or simulations of sensors when not.

The application of information theory to problems in analytical chemistry is not new. However, its inclusion in chemometrics for modeling large scale sets of multivariate spectral data that have only recently become available is novel. Recent research at the Naval Research Laboratory (NRL) has yielded probabilistic models for spectral data that enable the computation of information measures such as entropy and divergence, with the goal of developing feature sets to increase the sensitivity and selectivity of multivariate chemical sensors of several modalities. Results are presented for several types of spectral data in multisensor systems, as well as strategies for using information measures with other data sources. Binary, univariate, and multivariate sensors can all be modeled from an information-theoretic perspective, making it well-suited for the challenges of next generation chemical detection.

Introduction

Accurate, reliable, and consistently effective chemical sensing outside the laboratory remains a challenge for the diverse needs of the joint armed forces. Physical requirements for the ideal sensor vary widely among the services, as do performance requirements. Optimal performance is also needed for a variety of environments with diverse backgrounds and potentially contaminating interferants that may vary widely as well[1]. As a consequence, next generation chemical detection remains an area of active research and development[2].

One of the most common applications for chemical sensing in the field (i.e., outside the laboratory) is distinguishing target analytes in the presence of background chemical compounds that may be entirely unknown or differ markedly across the many environments where the sensors are deployed. In the laboratory, multivariate spectral instruments are often relied on by trained analytical chemists for their ability to elucidate chemical compositions of samples where the chemical constituency may be entirely unknown. Adapting spectral instruments for field sensing applications or autonomous sample analysis often means replacing the trained chemist with algorithms that compare sample spectra with a library of spectra from known target analytes and common background compounds. Typically, a distance metric is used to quantify the separation between the sample spectrum and a candidate library spectrum. The candidate with the minimum distance above a certain threshold or match factor is selected. Examples of this approach are used in the industry standard NIST software for mass spectra analysis[3].

Traditional statistical methods for data modeling and analysis have focused on finding the most likely or most common value in fitting data to a probability distribution. For many sparse data types, such as those generated in mass (MS) and ion mobility (IMS) spectrometry, infrared spectroscopy (IR), and gas chromatography (GC) data, the most common values do not typically characterize the data. Rather, it is the values of the uncommon or infrequent peaks that uniquely define spectral data. Figure 1 shows an example of a GC-MS total ion chromatogram (TIC) of a Navy diesel fuel together with examples of spectral data for cycloheptane. Fitting a data model to the peaks means fitting the sparse large values in the tail of the data distribution, which is challenging as they lie far from the most common values near zero.

A further challenge for autonomous spectral matching is that the relative intensity values of peaks in spectral data are not generally preserved and vary strongly with concentration of the analyte in the sample. Figure 2 shows two examples of 1,3,5-trimethylbenzene spectra obtained with a laboratory GC-MS instrument (left) and GC-IR instrument (right) overlaid with the corresponding intensity profile from the NIST and PNNL libraries, respectively. The relative peak heights between the sample and library spectra are not preserved. In Figure 2 (left), the maximum peaks appear in different m/z bins, even though the general profiles of the two spectra are quite similar. Peak heights increase with increasing analyte concentration, but saturate at a maximum value in most spectral instruments. In Figure 2 (right), the IR spectra are shifted with respect to each other, and this shift is non-linear with respect to bin number.

Information theory[4] provides content-agnostic methods that can be used to characterize sparse spectral data and determine its most significant components. These tools can quantify how effective sensor spectra are at discriminating between hypotheses (information divergence), how informative a measurement is expected to be (entropy), how informative a particular outcome of that measurement is (self-information), and how closely associated two outcomes are (pointwise mutual information). As a group, these tools are information measures that operate on probability distributions that can be developed directly from the data or from statistical models. In addition, the universality of an information-theoretic approach can be used to determine the information gains from new features or alternative models with wide ranging impact for systems that rely on data fusion such as chemical sensing, machine vision, and robotics.

Information theory was first applied to the problems of analytical chemistry in the 1970s[5]. The field of chemometrics subsequently developed analysis tools for multivariate spectral data[6]. The purpose of this report is to demonstrate how an information-theoretic approach to the analysis of multivariate spectral data is an effective strategy for addressing the unique challenges of chemical detection in the field. The report is organized as follows: the Background section establishes why spectral-based sensing is considered the most relevant for modality for next generation chemical detection. The Method section introduces the mathematical framework of several information measures, biasing schemes, and parameter estimation. The Experiment section describes the spectral data sets, performance metrics, and sensor simulations. Next, results and discussion are presented, followed by a summary of conclusions.

Manuscript approved October 22, 2013.

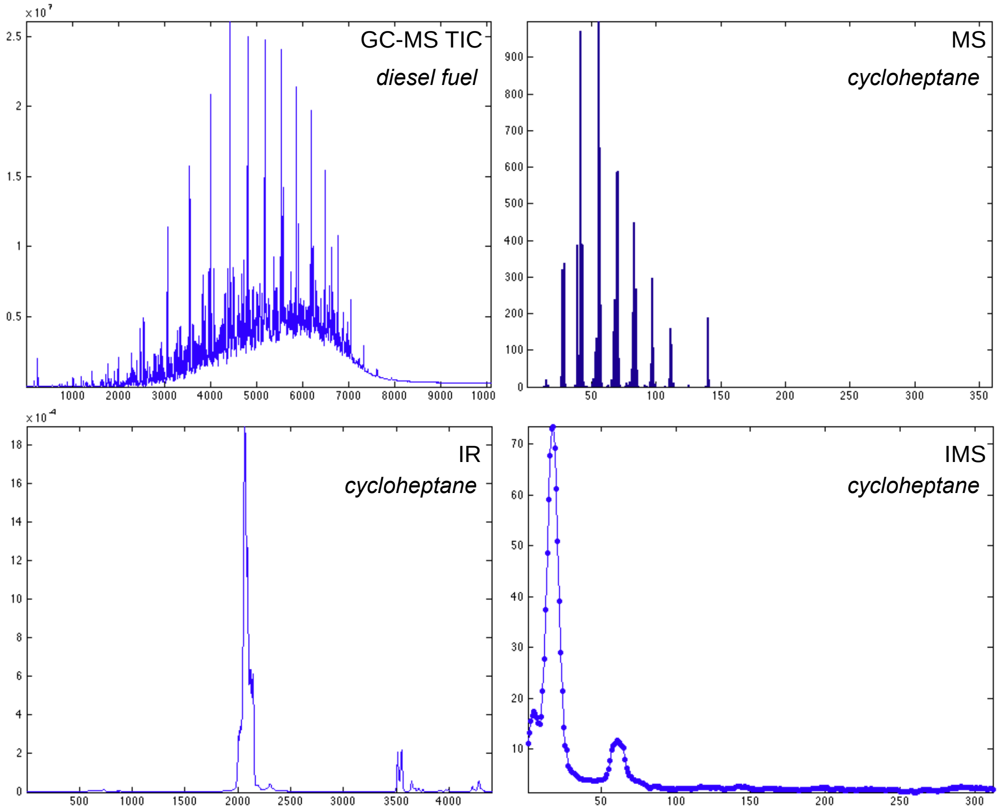

Figure 1 Examples of spectral data types: gas chromatograph mass spectrometry (GC-MS) total ion count, mass spectrometry (MS), infrared spectroscopy (IR), and ion mobility spectrometry (IMS).

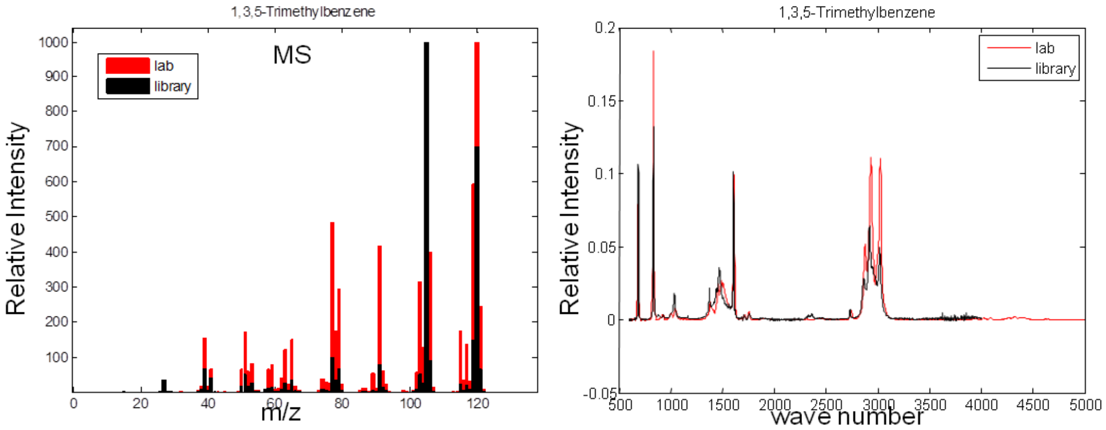

Figure 2 Comparison of mass (left) and IR (right) spectra with NIST and PNNL library spectra for 1,3,5-trimethylbenzene, respectively.

Background

Research efforts at NRL[7-10] have focused on the development of effective data fusion techniques for multisensor systems for Navy and other DoD and DHS sensing applications. These applications often require the integration of diverse sensing methods to meet coverage and performance goals. Targets are frequently rare events and potentially catastrophic, which adds an additional challenge for hypothesis tests that rely on prior probabilities. Backgrounds are often uncertain and vary according to whichever environment the system is placed. To be effective in field environments then, multisensor systems must be robust against these uncertain and potential contaminating backgrounds while maintaining efficient detection rates. The multivariate nature of spectral-based sensing systems (e.g., GC-MS, GC-IR) provides an inherent protection against unknown interferants. In analytical chemistry, these types of instruments are called second order instruments as they can detect targets and maintain calibration even in the presence of unknown interfering compounds[11].

Figure 3 illustrates the robustness present in infrared spectroscopy, for example. Here, principal component analysis[12] (PCA) was performed on a set of 540 unique IR spectra. PCA is a standard data reduction technique in chemometrics that functions by projecting high-dimensional data onto a lower-dimensional subspace that maximally describes the sample-to-sample variance present in the data. The cumulative variance represented by the IR spectra – one measure of their information content – reaches the 90% level with less than 50 of the 540 factors, which indicates a high degree of redundancy is present in this set of IR spectra. This redundancy gives spectral data an ability to represent a large number of chemical compounds with unique signatures while maintaining robustness to unknown interferants.

Under DTRA sponsorship, NRL is actively working to bring the robustness of spectral sensors to field-able multisensor instruments through an information-theoretic approach to sensor data analysis and fusion, as well as explicitly incorporating hypothesis testing that combines independent evidence for, evidence against, and uncertainty to target analyte detection.

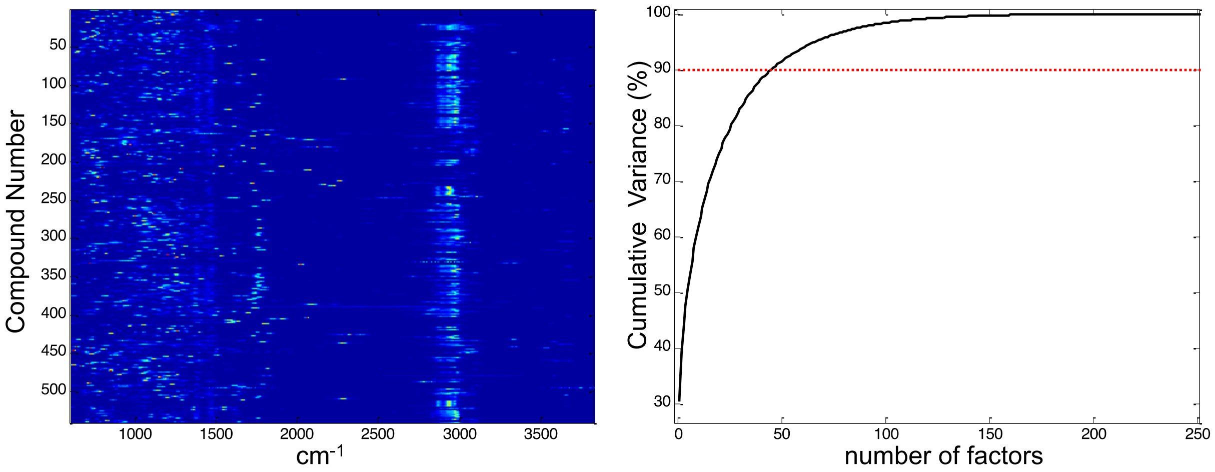

Figure 3 (Left) A set of 540 unique IR spectra. (Right) Cumulative variance (solid curve) represented as a function of the number of factors after principal components analysis; line of 90% of variance explained (dotted line).

Method

Information measures

Information measures are content-agnostic tools that can be used to characterize measurements and determine their most significant components. They operate on probability distributions that can be computed directly from data, or from simulations of sensors, or from statistical models.

Self-information is a property of the outcome of a measurement. Also called the “surprisal”, it quantifies how unusual the outcome is. For a measurement $\vec{x}_0$ sampled from a distribution $p(\vec{x})$, the self-information is:

$$I(\vec{x}_0) = -log_2[p(\vec{x}_0)] \tag{1}$$

This gives a measure of how informative the outcome is. If $p(\vec{x}_0) = 1$, then the result is deterministic, giving no information: $I(\vec{x}_0) = 0$. An outcome with a self-information of *n* bits gives the same information as 1 out of 2^n equally probable events.

The Shannon entropy (H) is a property of the sensor’s distribution in the measurement space – it is the average self-information:

$$H(X) = \mathbb{E}_X[I(\vec{x})] = -\sum_{\vec{x}\in X} p(\vec{x})log_2[p(\vec{x})] \tag{2}$$

It tells the average number of bits of information given by a measurement sampled from $p(\vec{x})$. The entropy $H(X) = 0$ when only one distinct outcome is possible (e.g., all samples are concentrated in one location in the measurement space), and is maximized, $H(X) = log_2(n)$, when each outcome is equally probable (e.g., n samples are spread uniformly over the measurement space). Together these information measures can be used to quantify how useful specific sensing measurements are; for example, determining which m/z bins in a mass spectrum are most important for discriminating between a particular sample and a library of target analyte spectra (self-information), or which are most important on average for discriminating between members of the library of targets and typical background spectra (entropy).

The Kullback-Leibler divergence (D_{KL}) is a tool for comparing two distributions:

$$D_{KL}[p(x)||q(x)] = \sum_{x\in X} p(x)log_2\left[\frac{p(x)}{q(x)}\right] \tag{3}$$

The Kullback-Leibler divergence is zero when $p(x) = q(x)$ everywhere, and increases as $q(x)$ becomes a worse approximation to $p(x)$. D_{KL} is specifically useful for data fusion applications when $p(x)$ is set to $p(\vec{x}|H_1)$ and $q(x)$ is set to $p(\vec{x}|H_0)$, where H_0 and H_1 are hypotheses being tested (e.g., “analyte X is present at or above concentration Y” versus “background is present”). In this case, D_{KL} measures the expected number of bits of information, per measurement, for discriminating in favor of H_1 when H_1 is true. Since the end-goal of any analytical task is to differentiate between a set of hypotheses, a statistical measure of the system’s capability to do just that is a robust, modality-independent predictor of performance.

The D_{KL} is an example of one of several possible divergence measures that can be constructed for the probability distributions $p(x)$ and $q(x)$. An alternative is the Jensen-Shannon (JS) divergence, D_{JS}:

$$D_{JS}[p(x)||q(x)] = \frac{1}{2}(D_{KL}[p(x)||m(x)] + D_{KL}[q(x)||m(x)]) \tag{4}$$

Here, $m(x) = \big(p(x) + q(x)\big)/2$ is the average of the two distributions. This quantity is closely related to the Kullback-Leibler divergence; if a measurement is drawn at random from $p(x)$ or $q(x)$, $D_{JS}[p(x)||q(x)]$ gives the expected number of bits of information, per measurement, for deciding which distribution was chosen. If $p(x)$ is set to $p(\vec{x}|H_1)$ and $q(x)$ is set to $p(\vec{x}|H_0)$, this gives the expected amount of information for discriminating between the hypotheses H_0 and H_1 when a sample is drawn at random from one of them. The JS divergence has the advantage that it is bounded both above and below and remains well-defined even for regions of measurement space where the probability distributions may be zero-valued (unlike D_{KL}, which requires absolute continuity).

The pointwise mutual information, I_p, is a property of two outcomes of a measurement:

$$I_p(x, y) = log_2\left[\frac{p(x, y)}{p(x)p(y)}\right] = I(x) - I(x|y) \tag{5}$$

The I_p is an information-theoretic generalization of the concept of statistical correlation. It computes the change in mutual information in result x, given that result y is known (and vice-versa). If the two outcomes are uncorrelated (that is, the outcome y doesn't affect the probability for x), then the I_p is equal to zero. Positive or negative values will be taken if the two outcomes occur together more or less often than they would if they were uncorrelated, respectively. I_p can be normalized by $I(x,y)$, so that its range extends from -1 (i.e., outcomes x and y never occur together) to +1 (i.e., outcomes x and y always occur together). Tests of correlation are important tools for validating data fusion algorithms, as many popular methods assume statistical independence (e.g., naïve Bayes, Dempster-Shafer Theory). By using a pointwise test, the regions of measurement space where statistical independence is valid can be ascertained.

Distance metrics

Four distance metrics were evaluated as features in this work. They were the cosine distance, Pearson's correlation distance, the cityblock or L_1 distance, and the euclidean or L_2 distance. Note that the cosine metric is used in the NIST spectral matching algorithm[13]. Table 1 summarizes the distance metrics.

Table 1 – Formula used for computing distance between sample and candidate library spectra.

Distance metric	Formula
cosine	$1 - \frac{\vec{p} \cdot \vec{q}}{\lVert\vec{p}\rVert\lVert\vec{q}\rVert}$
correlation	$\frac{cov(p, q)}{\sigma_p \sigma_q}$
cityblock (L_1)	$\sum_n \lvert p(x_n) - q(x_n)\rvert$
euclidean (L_2)	$\left(\sum_n \big(p(x_n) - q(x_n)\big)^2\right)^{1/2}$

Bias functions

A linear bias function was constructed to operate on mass spectral data where the data in high m/z bins were weighted more strongly than low m/z bins. A weight parameter, w, was multiplied against the vector of m/z bin values [1, 2, 3, … n]. The weight parameter thus acted as the slope of line that could be varied to implement the desired amount of bias to the intensity values of any given spectrum.

A non-linear bias function was constructed from the sigmoid function, a standard function frequently employed for logistic regression. The shifted sigmoid function

$$\sigma_i(x_i, \theta) = 2 \cdot \left(\frac{1}{(1 + e^{-\theta \cdot x_i})} - \frac{1}{2}\right) \tag{6}$$

can be defined using one parameter θ to weight all intensities equally, or as a vector $\boldsymbol{\theta}$ that can weight the intensity in each bin independently.

Parameter estimation

Information measures provide a means to estimate model parameters, for instance, optimizing values for denoising spectral signals. In both cases, the JS divergence was used to locate the optimal value. A data set of 540 noised infrared spectra was generated by adding Gaussian noise independently to each wavenumber bin of individual spectra. The noise was generated using a mean of 0 and a standard deviation of 5% of the maximum peak height in any given spectrum. Two approaches to denoising infrared spectra were implemented. One was thresholding; data below a selected threshold were truncated to zero. Selection of the optimal threshold was a trade off between removing noise (where a higher threshold was better) and preserving spectral information (lower threshold better).

A low pass filter was used as an alternative denoising method. A Fourier transform was applied to the noised infrared spectra. Frequency data were truncated above a selected cut off frequency. Selection of the optimal cut off frequency was again a trade off between removing noise (where a lower frequency was better) and preserving spectral information (higher frequency better).

Experiment

Sources of large sets of high-quality spectral data are available. One of these is the National Institute of Standards and Technology (NIST) 08 mass spectrometry library, a collection of reference electron-ionization mass spectra (EI-MS) of about 192,000 compounds. Another is the Department of Energy / Pacific Northwest National Laboratory (PNNL) Fourier-transform infrared (FTIR) spectroscopy library, a collection of high-resolution infrared spectra from more than 500 compounds including chemical agents, agent simulants, and common toxic industrial chemicals[14-15]. The mass spectral data were provided at unit mass-to-charge (m/z) resolution with intensities in a range of 0 to 1000 counts. Infrared spectra were provided at a resolution of 0.1 wave numbers (cm^{-1}) over a range of approximately 500 to 4000 cm^{-1} with intensities from 0.0 to 0.0569 absorption units. These two databases were cross-referenced to determine the compounds common to both data sets, which form a joint database (JDB) library of 540 unique compounds.

Several test data sets were constructed from these high-quality data sets to challenge matching algorithms with a range of different noisy and uncertain spectra with known and unknown classifications. One test set (TS-JDB) of 5500 mass spectra was constructed from the JDB library of 540 selected spectra. It consisted of three sets of 1000 randomly selected example spectra corrupted by Gaussian noise at levels of 0.3, 0.5, and 1.0 standard deviations as scaled by the maximum peak intensity. An additional 1000 random selected spectra were randomly permuted to generate spectra that were non-physical but preserved the entropy of the source spectra. Another 1000 spectra were constructed to mimic co-eluting compounds by adding two randomly selected spectra in a random proportion. Co-eluted spectra that were added with a proportion of 0.3 or less were considered as examples of the more prevalent compound; the remaining spectra were classified as non-matchable. Finally, 500 randomly selected and unaltered spectra were included as a truth set.

Other test sets (TS-NIST) of noisy mass spectra were constructed by adding random peaks to randomly selected spectra from the full NIST database to simulate altered relative peak intensities. For each sample spectrum, between 1 and 20 m/z bins were randomly selected, and their abundance values were increased by a random value between 100 and 500 (i.e., 10% - 50% as the spectra are pre-scaled to a maximum abundance of 1000 units). Uniform distributions were used for generating all random values.

Test sets (TS-PNNL) of noisy IR spectra were prepared from sample FTIR spectra by adding Gaussian white noise to spectra from the JDB database. Each wavenumber bin was shifted by an independently chosen random value to simulate the bin shifting frequently observed in IR spectra. Shift amounts were sampled from a normal distribution with zero mean and standard deviation equal to 5% of the spectrum's maximum value.

Correct matches (i.e., classifications) served as the metric for evaluating algorithm or feature performance. Spectra from a test set were compared individually to candidate spectra chosen from a subset of spectra selected from either of the high-quality MS or IR data sets. Comparisons were achieved using a distance metric or probabilistic measure. The candidate spectrum with the smallest distance value compared to the test set spectrum was selected as the "match". Tallies of correct and incorrect matches were used to generate confusion matrices and ROC curves to summarize performance[16]. The area under the ROC curve (AUC) metric was also used in performance comparisons[17].

Additionally, systems of one, two, and three univariate sensors were simulated using Gaussian sensor response distributions for target analyte and background signal. The degree of overlap was varied between the sensor responses and Gaussian noise was added. The performance of the sensor systems was evaluated using both AUC and information divergence metrics.

Results and Discussion

Self information

Self-information was used as a measure to quantify how informative each m/z bin is in a mass spectrum, given a corpus of mass spectra such as the NIST library. Figure 4 shows the mass spectrum of dimethoxy-benzamide, along with the self-information of each bin. For this calculation, the m/z bins were assumed to be statistically independent; a peak at one bin did not affect the chances of a peak occurring at another bin (a rather stringent assumption that is further discussed and tested below). With this simplifying assumption, the probability of a peak at each bin was straightforward to compute for the NIST database. While the qualitative general knowledge that higher m/z bins are more informative was reflected here, a quantitative interpretation is available as well. With nearly 192,000 spectra, distinguishing a spectrum in the NIST database takes (approximately) a minimum of $log_2(192{,}000) \approx 17.5$ bits of information, the equivalent of one bit per spectrum.

In Figure 4, the highest self-information m/z bin – near 360 – for dimethoxy-benzamide has only 5.3 bits. While that bin provides the most information for distinguishing that chemical, it is not sufficient to uniquely identify the compound on its own (5.3 << 17.5). However, the self-information $I(x)$ is extensive. Any feature-extraction/matching algorithm for distinguishing this chemical from those in the NIST database must incorporate enough peaks so that the sum of their self-information is above 17.5. For example, the five highest peaks in self-information are sufficient to meet the criteria; the five highest peaks in the spectrum (~ 13 bits) are not. Note that their locations in m/z differ. Thus, the most important peaks for distinguishing the chemical against unknown background are not necessarily those of the most common mass fragments.

Information divergence and AUC

Recent results using simulated sensor data have demonstrated a connection between measures of information divergence such as D_{KL} and D_{JS} and the area under a ROC curve (AUC) performance metric. The information divergence of a sensor's measurements is a statistical property of the sensor, whereas the AUC metric is an empirical property describing the sensor's performance that encompasses those measurements. Figure 5 shows a logit plot of the AUC against D_{JS} for simulated arrays of one, two, and three univariate sensors with random parameters and Gaussian noise. The data from all three array sizes collapse onto a single curve – so a lone sensor with D_{JS} = 0.7 bits will have approximately the same performance as an array of sensors with D_{JS} = 0.7 bits. That is, an array with a small number of high-performing sensors may outperform an array with more, but lower-performing, sensors. D_{JS} provides a modality-independent measure for comparing them. The logit scale is used to emphasize details in the region of high performance (areas of 0.95 < x < 1.0).

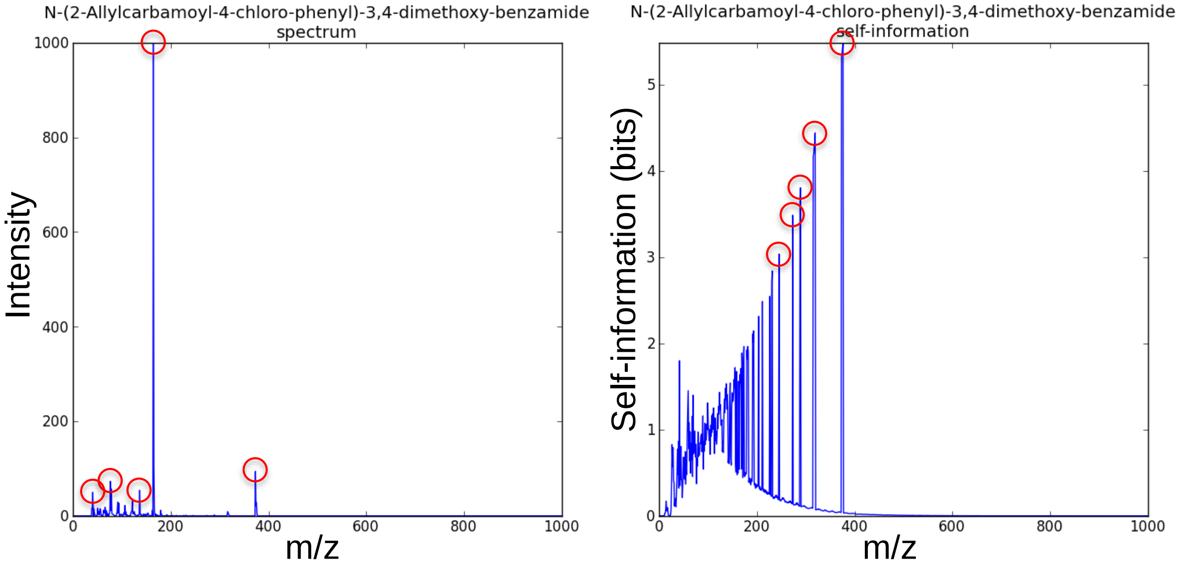

Figure 4 Mass spectrum (left) and self-information (right) of dimethoxy-benzamide. The five largest valued peaks are circled (red) in both plots.

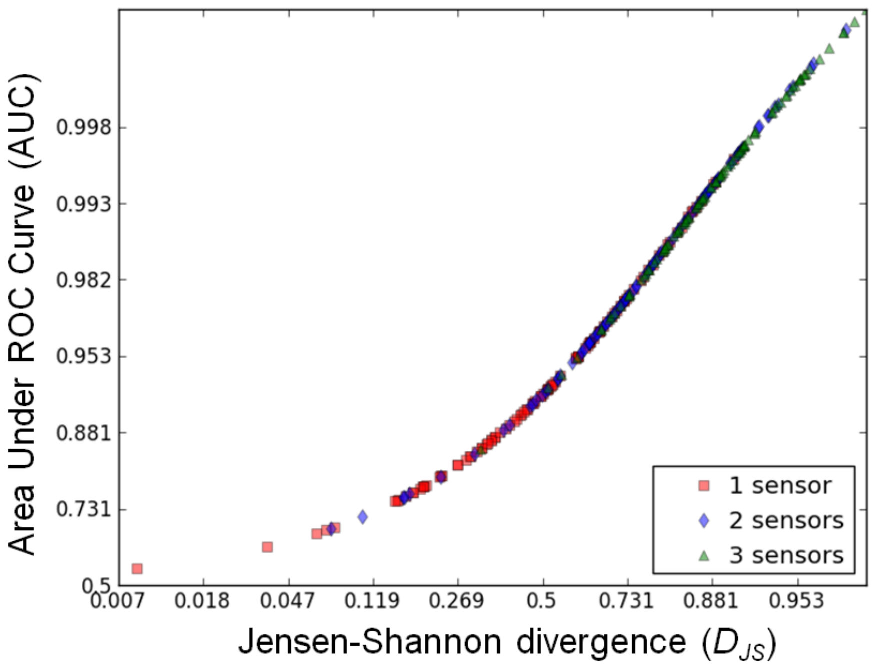

Figure 5 Logit-logit plot of area under ROC curve (AUC) versus D_{JS} for simulated 1-, 2-, and 3-sensor systems.

Feature selection

The Jensen-Shannon divergence, D_{JS}, may similarly be used as a measure of the effectiveness of a feature-extraction method when working with high-dimensional spectral data. Different features can be quantitatively compared based on the Jensen-Shannon divergence of the distributions they create. In addition, the free parameters of a feature extraction algorithm can be optimally selected by maximizing D_{JS}. Several features based on distance metrics and modifications of distance metrics were evaluated for mass spectral data. Simulated "sample" spectra from test set TS-NIST were compared to matching and nonmatching NIST library spectra using these features. From these data, smooth distributions of the distances between sample and matching library spectra, and between sample and nonmatching library spectra, were built using Gaussian kernel density estimation. Rather than selecting the next nearest matching spectra for the nonmatching distance, the nonmatching library spectra were selected at random from the list of all nonmatching chemicals for each sample spectrum separately. The Jensen-Shannon divergence between these two distributions was then computed and used to evaluate the performance gain (or loss) with that feature.

Different distance metrics capture different information about spectral data. Figure 6 shows one measure of information obtained from a set of 540 mass spectra and the distance metrics given in Table 1. Points in these scatter plots represent the separation between pairs of different spectra as computed with different distance metrics; for example, Figure 6 (upper left), with correlation distance (y-axis) and cosine distance (x-axis). Distance metrics that provide similar information will generate points in such a scatter that lie along the diagonal bisection line, as with the correlation and cosine metrics in Figure 6 (upper left). Metrics that measure more complementary information will scatter farther from this line; for example, Figure 6 (upper right, lower left, and lower right).

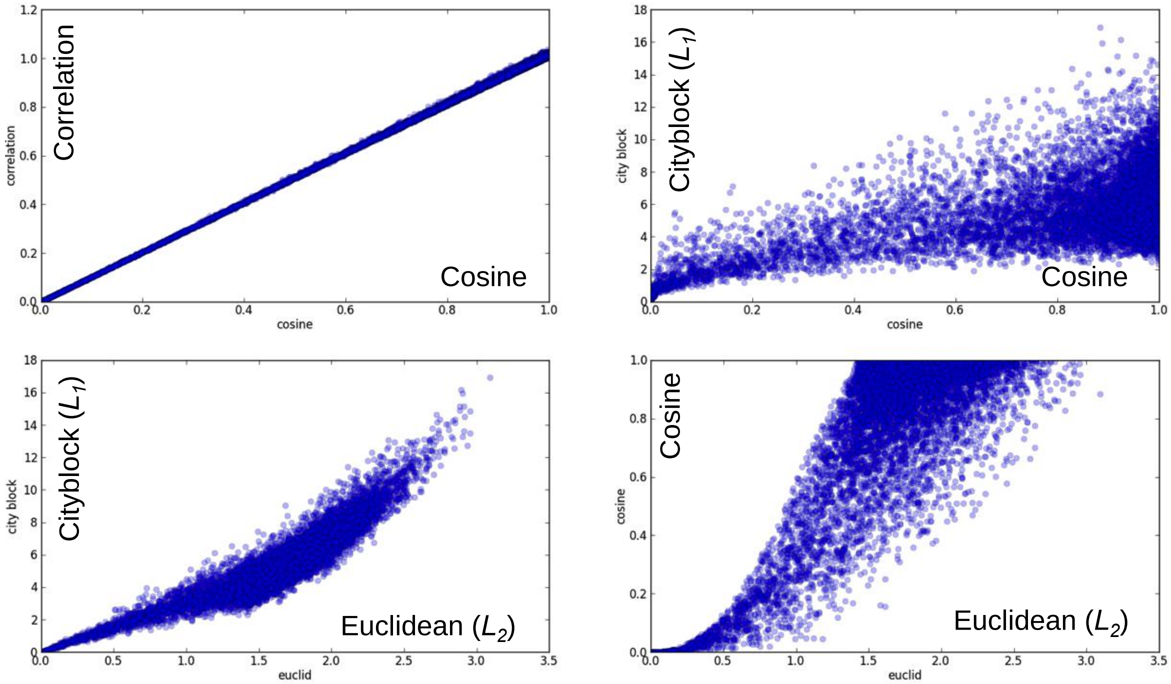

Figure 6 Differences in correlations for different distance metrics, as labeled. The amount of scatter is a measure of complementary information.

Linear bias

The first set of feature extraction methods tested involved application of a linear weight to the MS data before comparing them with a distance metric. Since higher-m/z bins are more informative by proximity to the molecular ion, it was reasoned that biasing the analysis toward them could improve matching performance. Each bin was multiplied by a linear weight $w = (1 + ax)$, where x was the m/z value of the bin and a was the scaling factor to be varied. Figure 7 (left) shows the Jensen-Shannon divergence (on a logit scale) for the four distance metrics of Table 1 as a function of the linear scaling factor a. In the $a \approx 0$ case (far left), distances were measured between nearly unweighted spectra. As the scaling factor was increased, D_{JS} decreased and then saturated for each distance metric. The relative performance of the different metrics remained approximately unchanged. This result suggests, then, that the linear bias method would actually *decrease* performance.

Non-linear filtering

The second set of methods tested involved the application of a nonlinear sigmoid (logistic) filter. Since real spectra can show significant variability in peak height (see, for example, Figure 2), it was reasoned that suppressing those differences with a logistic function could improve matching performance. The abundance of each bin in the spectrum was rescaled by the sigmoid function, $\sigma_i(x_i, \theta)$, of eq. (6). Here $\boldsymbol{x}$ was the spectral abundance vector and θ was the variable scale factor. The sigmoid was shifted and scaled from the standard logistic function, so that when $\theta \ll 1$ the distance metrics of the unweighted spectra would be recovered (up to a constant factor, which did not affect D_{JS} values). Figure 7 (right) shows the Jensen-Shannon divergence (on a logit scale) for the four distance metrics as a function of the scale factor θ. These results are considerably more interesting than the linear weight case. As θ increased from a very small value, D_{JS} decreased for all metrics, only to reach a minimum near $\theta \approx 10^{-2}$ before increasing. For higher values of θ, the divergences of the cityblock and euclidean metrics (i.e., the L_1 and L_2 norms) were substantially better than for the unweighted data case. However, the two metrics with the best performance for unweighted data (correlation and cosine distances) saw no improvement.

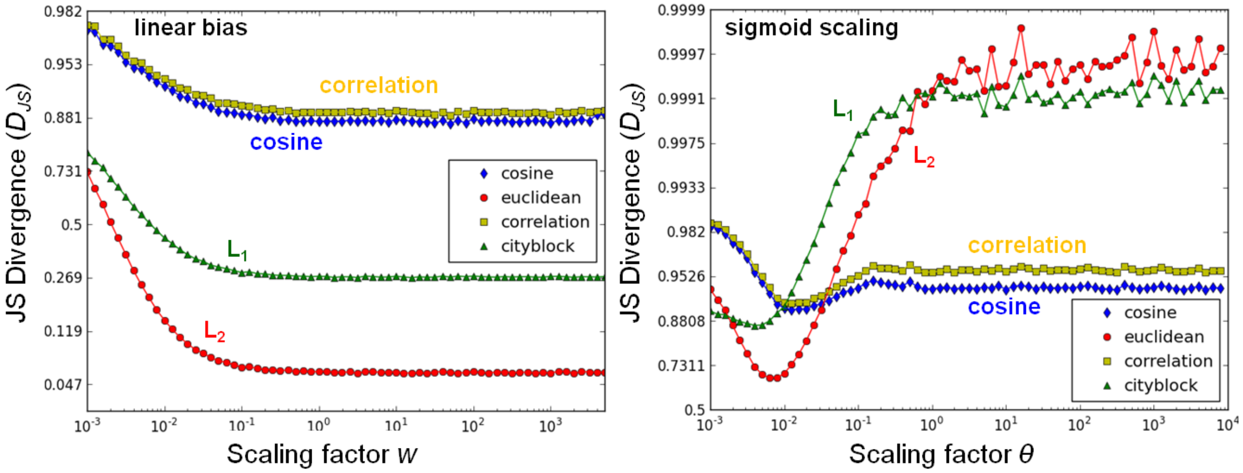

Figure 7 Logit plot of D_{JS} versus linear (left) and sigmoid (right) scale factors for four distance metrics.

The JS divergence can be used to show the effects of non-linear filtering for each distance metric separately. Figure 8 shows the separation distributions for matched spectra (blue) versus randomly selected mismatches (green) for four distance metrics when the scaling factor $\theta = 0$. The distributions are histograms of the distance values computed with each metric. Less overlap between the distributions indicates better distinguishability of spectra and a commensurable increase in D_{JS}. Figure 9 shows the separation distributions for scaling factor $\theta = 1000$. The distributions have generally narrowed and shifted to the left after application of the sigmoid filter.

Figure 10 shows the effects of the sigmoid filter on the correlation and euclidean distance metrics in scatter plots for three values of parameter θ: 0.01, 1, and 1000 for matched and unmatched spectra. As the scaling parameter increased, the two distributions changed shape; at low values (far left) they overlapped, and at high values (far right) a clear separation can be seen in one direction, but not the other. Figure 11 revisits the aforementioned link between AUC and D_{JS}, using the results of the linear-weighting and sigmoid filter experiments. For both experiments, the data from all four distance metrics approximately collapse onto a single curve. A logit-logit scale is used to show that this relationship extends to extremely high performances in the range of AUC > 0.95.

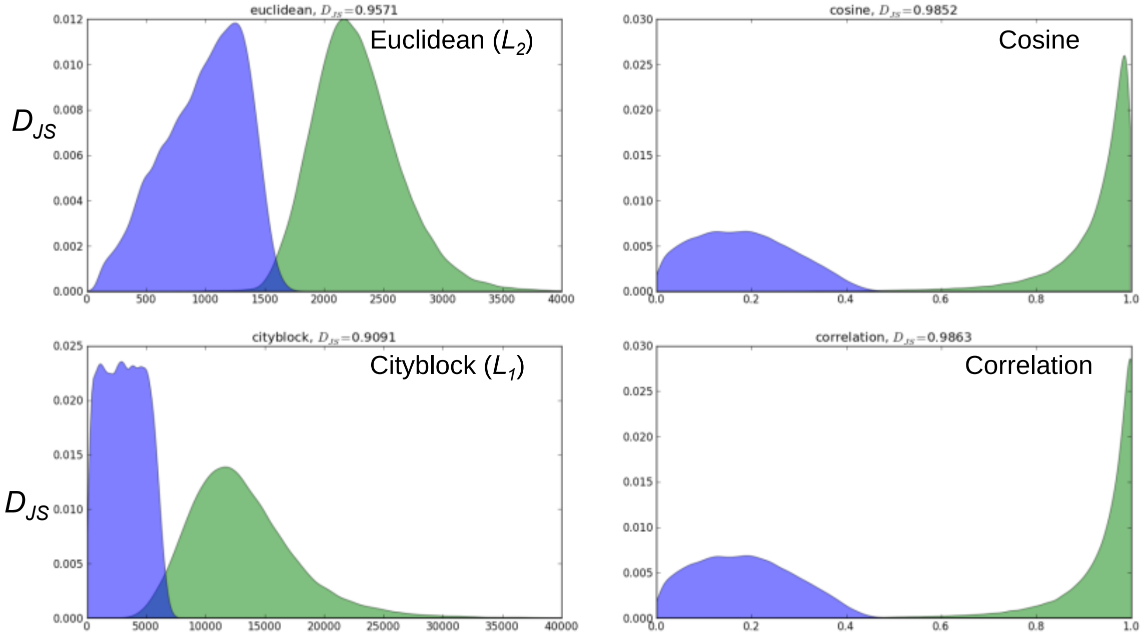

Figure 8 Scaling factor $\theta = 0$ – separation between matched (blue) and mismatched (green) spectral distributions for four distance metrics.

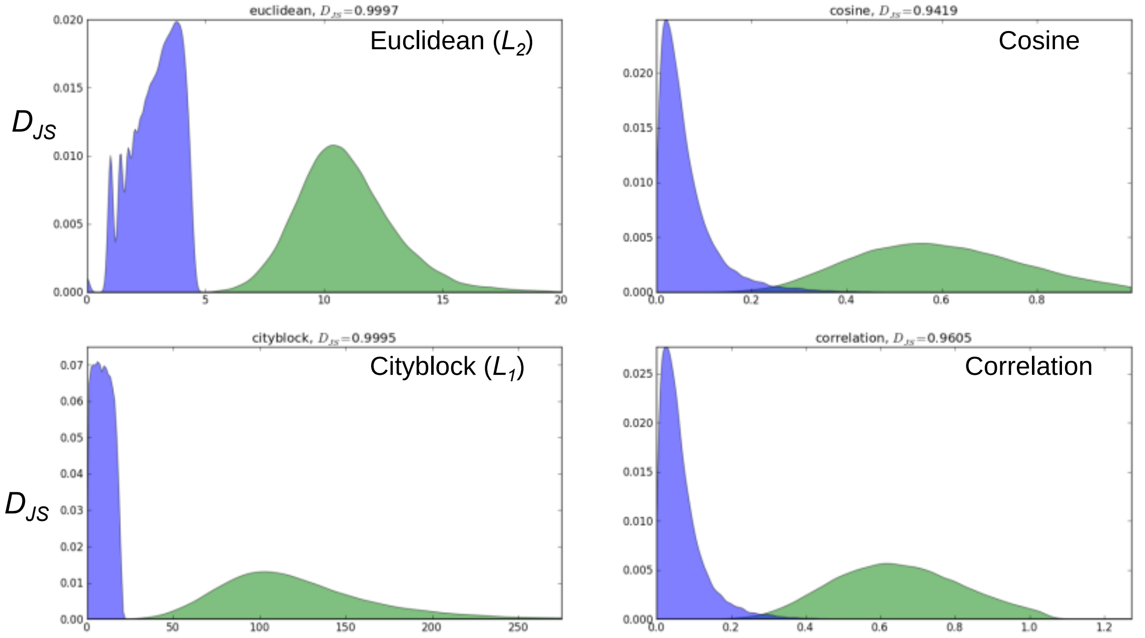

Figure 9 Scaling factor $\theta = 1000$ – separation between matched (blue) and mismatched (green) spectral distributions for four distance metrics.

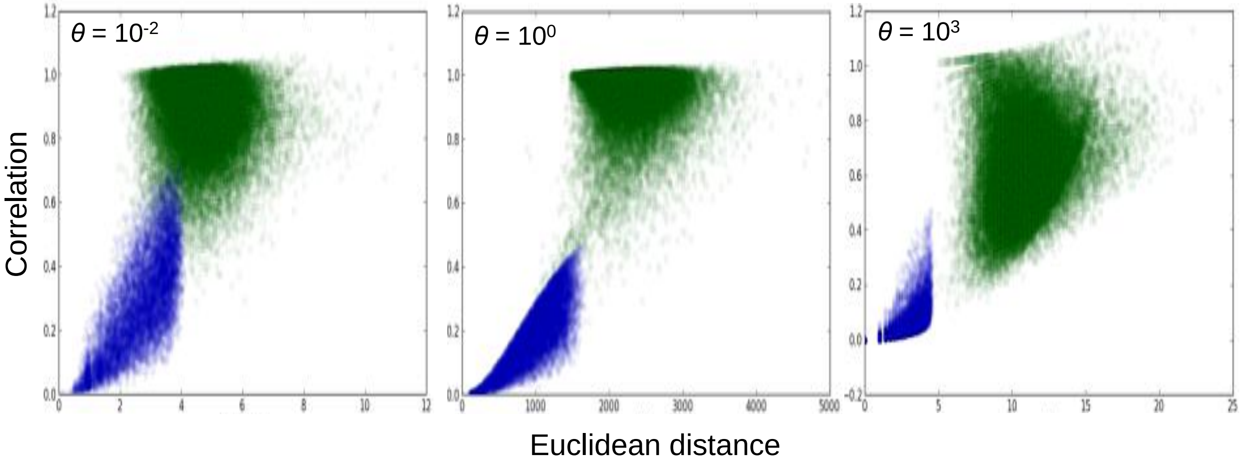

Figure 10 Correlation vs euclidean distance for matched spectra (blue) and unmatched spectra (green) for three different scale factors.

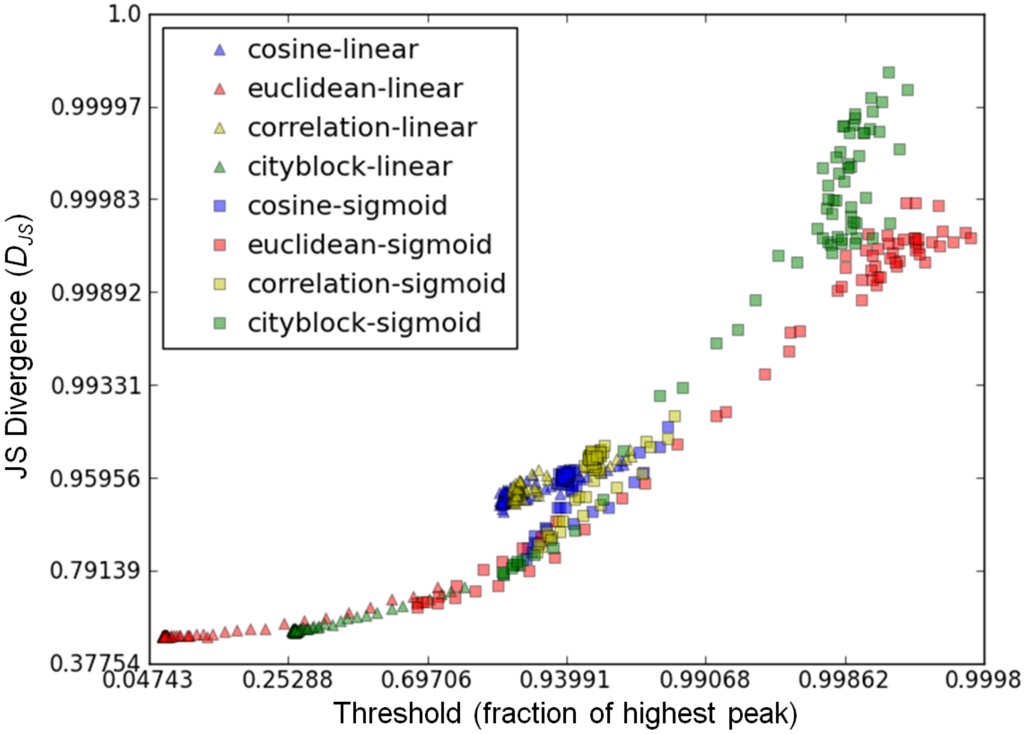

Figure 11 Logit-logit plot of area under ROC curve (AUC) versus D_{JS} for linear and sigmoid features.

Parameter estimation for optimal denoising

The noisy IR test data set, TS-PNNL, was used to test the modality-independence of performing feature-extraction optimization with the information measure D_{JS}. Distances were calculated after intensities in all bins below a variable threshold were set to zero (i.e., hard thresholding). Figure 12 illustrates the thresholding procedure for a sample noised IR spectrum. The position of the threshold represented a clear optimization problem: a threshold set too low let in most of the noise, which then broadened the matching and nonmatching distributions and decreased performance. A threshold set too high degraded performance by discarding too much useful distinguishing information. Figures 13 and 14 show changes in the separation distance between the distributions before (Figure 13) and after (Figure 14) thresholding with a given value. Figures 13 and 14 again demonstrate that different distance metrics measure different features of multispectral data. The optimal threshold value was determined using the Jensen Shannon divergence. Figure 15 shows D_{JS} as a function of threshold value. A clear maximum value can be determined for each curve. These values represent the best performance over all spectra for the given distance metrics.

The same noisy IR test data set was used for the alternative denoising method employing a low pass filter. In this application, a Fourier transform was applied to the noised infrared spectra. A low pass filter was implemented by truncating frequency data below a selected cut off frequency. Figure 16 shows application of a given low pass filter to a sample noised IR spectrum. Selection of the optimal cut off frequency was again a trade off between removing noise (where a lower frequency was better) and preserving spectral information (higher frequency better). Again, the optimal cut off frequency was determinable from the Jensen Shannon divergence. Figure 17 shows D_{JS} as a function of cut off frequency for the four distance metrics. Again, the maximum values can be determined for each curve and represent the best performance over all spectra for the given distance metric.

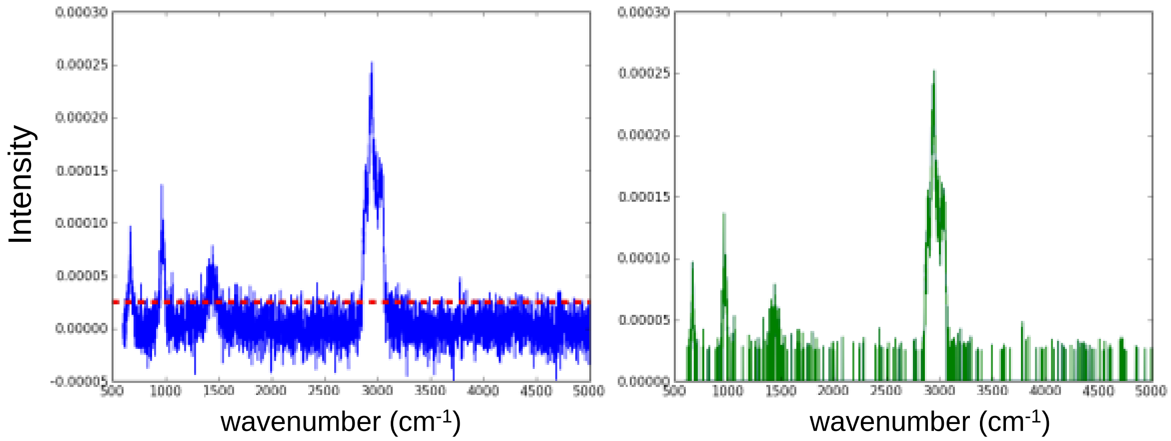

Figure 12 Application of thresholding: (left) noised IR spectrum (curve), threshold (dashed line), (right) IR spectrum after denoising.

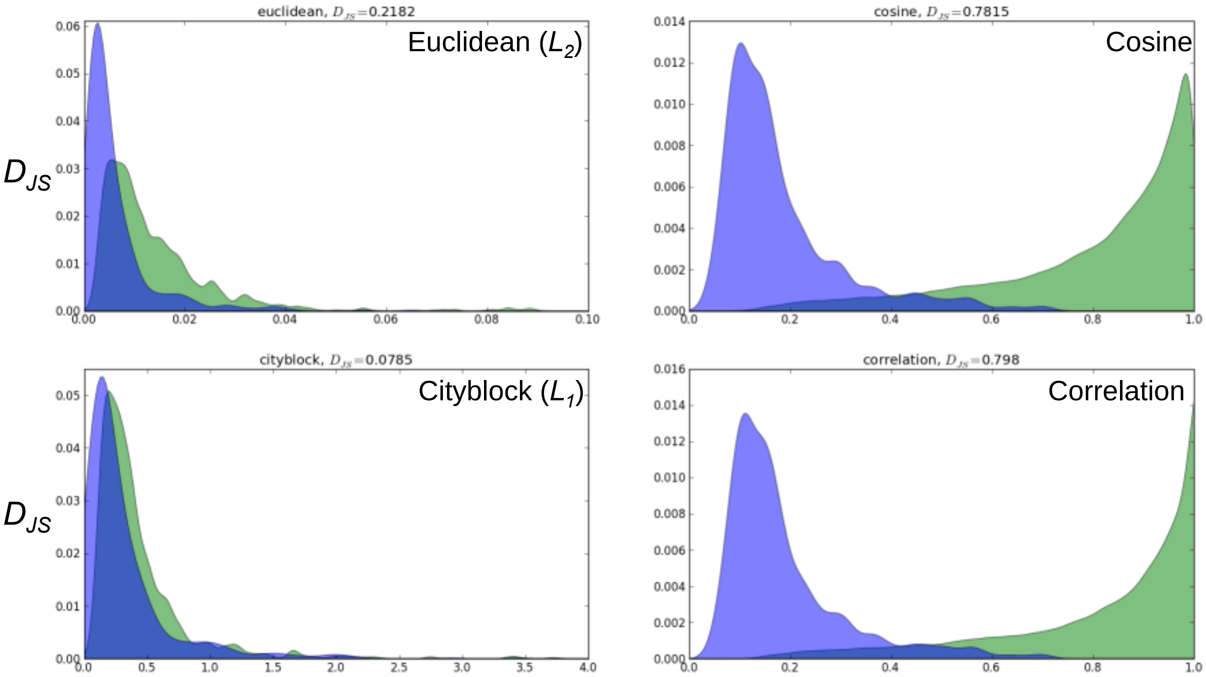

Figure 13 Distributions of matched (blue) and mismatched (green) noise IR spectra before thresholding for four distance metrics.

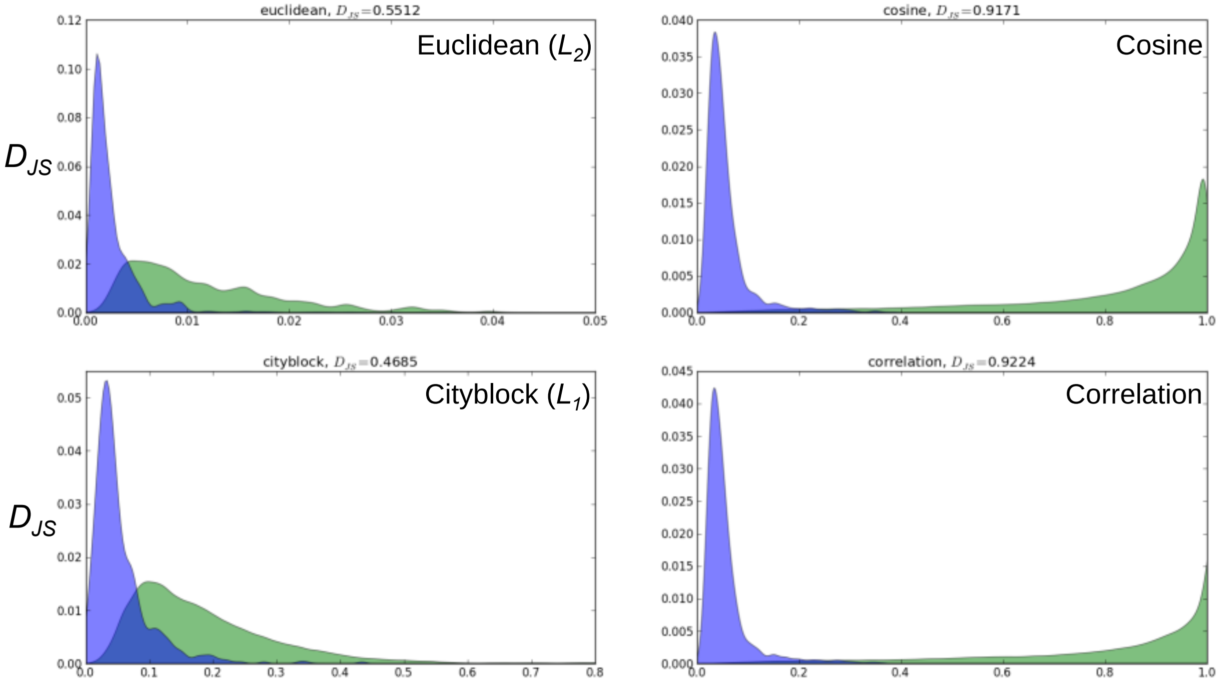

Figure 14 Distributions of matched (blue) and mismatched (green) noise IR spectra after thresholding for four distance metrics.

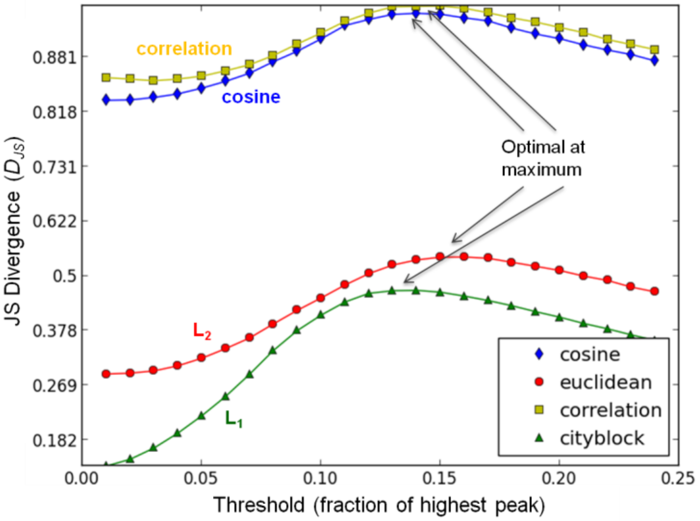

Figure 15 D_{JS} for selected thresholds over noised IR data for four different distance metrics.

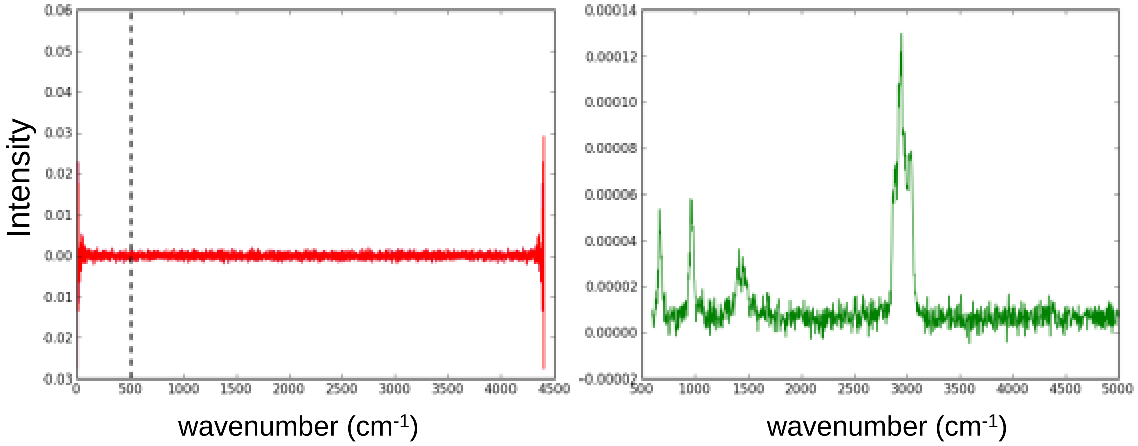

Figure 16 Application of low pass filter in Fourier space: (left) Fourier transformed noised IR spectrum (curve), cut off frequency (dashed line), (right) IR spectrum after denoising with low pass filter.

Pointwise mutual information

Finally, Figure 18 shows the normalized pointwise mutual information computed for the full NIST mass spectral database. The points in the thermal plot represent how the likelihood of a peak at one m/z bin affects the likelihood of a peak at another bin. If the two bins are uncorrelated, the value is near zero; points on the diagonal are unity, by definition. Dark blue areas along the *x* and *y* axes are likely artifacts of the low peak count when m/z < 50. The high mutual information, especially in the upper-right corner, suggests that the independent-bin approximation may be a poor one. That is, the presence of one peak at a high m/z bin strongly increases the likelihood of another high m/z peak, since one high-mass molecule is likely to have several high m/z peaks. Figure 18 also shows strong correlations parallel to the diagonal, spaced about Δm/z=14 units apart. This is due to a common fragmentation pattern observed in electron ionization mass spectrometry involving loss of successive methylene subunits (-CH_2-) of larger molecules. While these correlation results are not exactly novel, they do demonstrate the utility of pointwise mutual information for elucidating unknown structure in a data set.

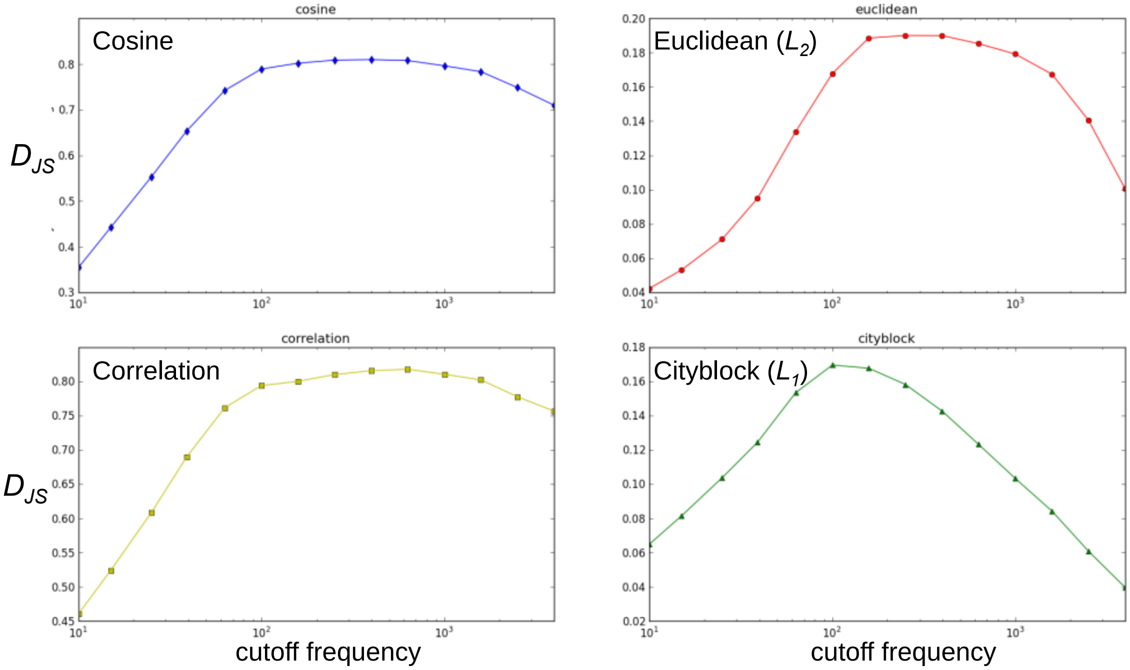

Figure 17 D_{JS} for selected cut off frequencies over noised IR data for four different distance metrics.

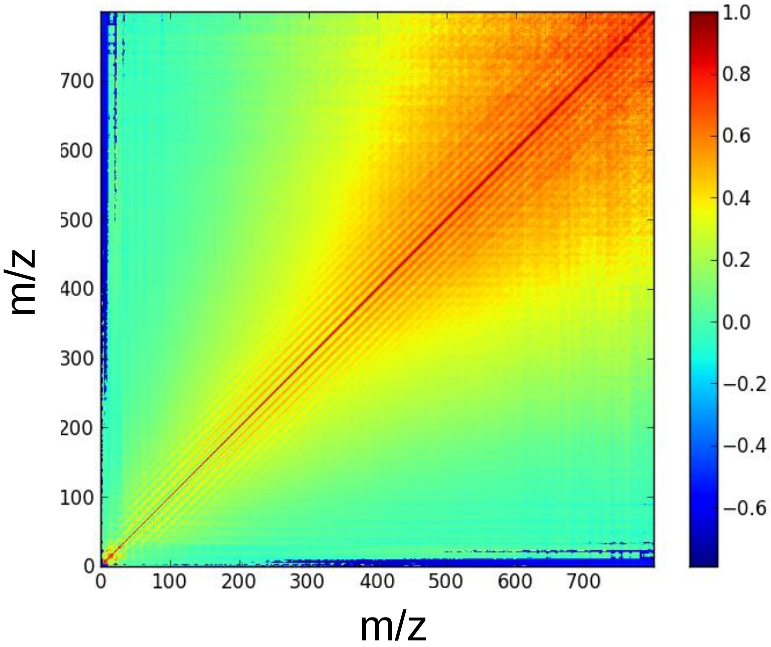

Figure 18 Correlations in the NIST data computed from using pointwise mutual information.

Conclusions

Chemical sensing data such as those from spectral instruments are typically sparse and characterized by a small number of large valued peaks. Traditional statistical methods that focus on characterizing the most frequent measurements using, for example, a mean or standard deviation for modeling data, have not performed well with sparse spectral data. Further, relative peak heights and peak locations in spectral data can vary significantly as they depend strongly on the input analyte concentration as well as the measurement process. Algorithms that match sample spectra to a library of candidate targets for chemical identification must be robust with respect to these characteristics of chemical sensing data. Distance metrics – by definition – treat each element of a spectrum as equally informative and thus are highly susceptible to variations in relative peak heights and locations. While an improvement over traditional statistical modeling, distance metrics have met with only limited success in large-scale autonomous identification of chemicals from sensor data.

An information-theoretic approach that focuses on differentiating significant from insignificant data in chemical spectra (i.e., relevant signal from irrelevant background) was shown as a potentially effective alternative to these methods. Information measures form an effective set of tools for analyzing and characterizing spectral data and multisensor systems: individual measurements with self-information, typical measurements with entropy, correlations with mutual information, and relative performance with divergence. Much of their utility arises from their independence of the actual information content; modeling how chemical sensors generate informative data is not required, but could potentially yield additional information for distinguishing chemicals.

Information measures also provide a means to quantify correlations and performance gains for different types of data, features, and fusion algorithms. Divergence metrics are directly relatable to ROC curves and the AUC performance metric, and thus, are able to measure the complementarity of sensor responses and data features. Binary, univariate, and multivariate sensors can all be modeled within this framework, making it well suited for addressing the unique challenges of next generation chemical detection in the field.

Acknowledgement

The authors gratefully acknowledge the support of the Defense Threat Reduction Agency under program CA09DET503G and the U.S. Naval Research Laboratory.

References

[1] U.S. Army Standardization and Specification Team, "Performance specification for the next generation chemical detector (NGCD)," JPM, NBC Contamination Avoidance, PRF EA-D-10029, CAGE Code 81361, June 4th, (2013).

[2] Joint Science and Technology Office for Chemical and Biological Defense, "Service call for proposals: FY 14/17 non-traditional agents new initiatives (JSTO-CBD FY14-17-NTA-BAA)," Defense Threat Reduction Agency, Chemical and Biological Defense Program, (2013). And Broad Agency Announcement (HDTRA1-14-17-NTA-BAA), http://www.fedbizopps.gov, (2013).

[3] NIST/EPA/NIH, "Mass spectral library with search program", (2008), <http://www.nist.gov/srd/nist1a.htm>.

[4] Kullback, S., *Information Theory and Statistics*. Courier Dover Publications, New York, (1968).

[5] Eckschlager, K. and Štěpánek, V., *Analytical Measurement and Information*. Research Studies Press, John Wiley and Sons, New York, (1985).

[6] Malinowski, E., *Factor Analysis in Chemistry, 2nd edition*. John Wiley and Sons, New York, (1991).

[7] Johnson, K.J, Minor, C.P., Guthrie, V.N., Rose-Pehrsson, S.L., "Intelligent data fusion for wide-area assessment of UXO contamination," *Stochastic Environmental Research and Risk Assessment (SERRA)*, 23(2), 237-252, (2009).

[8] Minor, C.P., Johnson, K.J., Rose-Pehrsson, S.L., Owrutsky, J.C., Wales, S.C., Steinhurst, D.A. and Gottuk, D.T., "A full-scale multisensor system for damage control and situational awareness," *Fire Technology*, 46(2), 437-469, (2009).

[9] Johnson, K.J., and Minor, C.P., "Practical considerations in bayesian fusion of point sensors," *Proc. SPIE* 8407, 84070X, (2012).

[10] Minor, C.P., Johnson, K.J., and Brooke, H., "Fusion of disparate spectra for chemical identification," *Proc. SPIE* 8064, 80640J, (2011).

[11] Booksh, K.L., and Kowalski, B.R., "Theory of analytical chemistry," *Anal. Chem.*, 66, 782A – 791A, (1995).

[12] Duda, R. O., Hart, P. E., and Stork, D. G., *Pattern Classification*. New York, John Wiley & Sons, Inc., 2001.

[13] Stein, S.E., and Scott, D.R., "Optimization and testing of mass spectral library search algorithms for compound identification", *J. Am. Soc. Mass Spectrom.*, 5, 859-866, (1994).

[14] Sharpe, S., Sams, R., and Johnson, T., "An infrared spectral library for atmospheric environmental monitoring," *SPIE Newsroom*, April 12th, (2006).

[15] Department of Energy (DOE) / Pacific Northwest National Laboratory (PNNL), "Spectral library of quantitative infrared absorption spectra", (2010), <http://nwir.pnl.gov/>.

[16] Fawcett, T., "An introduction to ROC analysis," *Pattern Recognition Letters*, Volume 27, Issue 8, June 2006, Pages 861-874.

[17] Bradley, A. P., "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, Volume 30, Issue 7, July 1997, Pages 1145-1159.