Extended cross correlation: A technique for spectroscopic pattern recognition

Matthew P. Jacobson, Stephen L. Coy, and Robert W. Field
Massachusetts Institute of Technology, Department of Chemistry and G. R. Harrison Spectroscopy Laboratory, Cambridge, Massachusetts 02139

(Received 24 April 1997; accepted 20 August 1997)

Recent improvements in spectrum excitation, recording, and processing capabilities have led to enormous enhancement in the quality and quantity of spectroscopic data sets. We describe here a pattern recognition technique, extended cross correlation (XCC), that is well suited to take advantage of large, high quality data sets. In particular, spectra are used to decode each other without any knowledge of or assumptions about the patterns that are sought. This paper describes the motivation for and construction of the XCC, and illustrates one of its simplest applications: To identify, in spectra of mixtures of chemical species, which peaks correspond to which chemical species. This application of the XCC is illustrated with both synthetic data and experimental data on mixtures of ammonia isotopic species. © *1997 American Institute of Physics.* [S0021-9606(97)01944-2]

I. INTRODUCTION

Spectra contain an enormous quantity of information. The task of extracting this information is made difficult by extrinsic (resolution, signal-to-noise, spectral coverage) and intrinsic (unknown or overlapping patterns) factors. We propose here a pattern recognition-based rather than model-based method for recovering information from spectra.

One traditional approach to understanding the information encoded in spectroscopic data has been first to assign approximate quantum numbers to the upper and lower energy levels involved in each observed transition and then to relate the positions and intensities of the assigned transitions to a quantum-mechanical effective Hamiltonian model that allows insight into the system being studied. In complex or congested spectra, however, the process of assignment may be difficult, tedious, or either impossible or ill advised. In such situations, it would be desirable to identify diagnostically important, but *a priori* unknown, patterns that are obscured by the complexity of the nascent spectroscopic data. For this purpose, we have developed two closely related pattern recognition techniques, which we refer to as extended auto correlation (XAC) and extended cross correlation (XCC).

The XAC method has been described in detail previously;[1] its purpose is to locate multi-element patterns that are repeated in an unspecified way within a single spectrum. The XCC, which is the focus of this paper, is designed to recognize patterns that are repeated in multiple spectra. To provide a concrete understanding of the type of pattern recognition for which the XCC is useful, we provide three examples of spectroscopic data in which patterns can be identified in multiple spectra:

1. *Spectra of an unknown mixture of chemical species.* Certain chemical species (e.g., transient molecules, single isotopomers) are difficult to isolate. If such a species is desired to be characterized spectroscopically, then frequently one must be content with obtaining spectra of mixtures of chemical species, one of which is the species of interest. A number of approaches are possible to determine which features in the spectra of the mixtures correspond to the species of interest. One straightforward method is to obtain spectra of several mixtures, each of which contains the species of interest in a different fractional abundance. Peaks in the spectra that belong to the species of interest will have intensities which vary with its fractional abundance; the relative intensity of a peak in the various spectra can be used to assign it to a chemical species. This process of assigning spectral features to distinct chemical species represents a type of pattern recognition: The spectrum of one chemical species represents a pattern which is searched for in several spectra.

2. *Dispersed fluorescence spectra of acetylene.* It is now well established that the vibrational structure of the acetylene $\tilde{X}$ state is characterized by a polyad structure.[2–5] That is, the eigenstates of the $\tilde{X}$ state can be described to a good approximation by an effective Hamiltonian which is block diagonal. Each of the blocks in the effective Hamiltonian is called a polyad and can be labeled by a set of three approximately conserved quantum numbers which are called polyad numbers. Dispersed fluorescence spectra have been recorded from several different vibrational levels of the $\tilde{A}$ state of acetylene, and it can be demonstrated that each of these spectra can be described in terms of the illumination of exactly one bright state per symmetry-accessible polyad, and that each of the vibrational intermediate states illuminates the same set of bright states. Under these conditions, each polyad that is experimentally observable has an identical appearance in each of the dispersed fluorescence spectra. That is, each of the eigenstates that belong to the same polyad display the same pattern of spacings and relative intensities in each dispersed fluorescence spectrum. Thus, polyads can be identified in the dispersed fluorescence spectra by a pattern recognition process that is quite similar to that described

in Example 1.[6,7] The individual polyads, illuminated by a given set of bright states, represent patterns that can be extracted from dispersed fluorescence spectra obtained from multiple $\tilde{A}$ state vibrational intermediates.

3. *Atmospheric emission simulation experiments with carbon monoxide.* Time-resolved infrared emission spectra of CO, following excitation with a pulsed electron beam, have been recorded by Lipson *et al.* at the Phillips Laboratory LABCEDE facility.[8] These spectra consist of overlapping progressions of vibrational fundamental ($\Delta v = 1$) emission bands ranging from $v' = 1$ up to at least $v' = 12$. Most of these bands can be analyzed by least-squares fitting with vibrational basis sets, which yields kinetic data of atmospheric importance, but the $v = 1 \rightarrow 0$ emission is highly self-absorbed (due to the abundance of ground-state CO), and cannot be modeled accurately. An alternative approach to analyzing this emission band is to consider it a pattern that is repeated in each of the various time-resolved spectra with a different amplitude. The isolation of this band pattern permits a determination of the time dependence of the $v = 1 \rightarrow 0$ emission.

Each of these types of spectroscopic data sets is an example of what we mean by ''identifying patterns that are repeated in multiple spectra'' Example 1 will be considered in detail in this paper. The application of the XCC to Examples 2[9] and 3[10] will be addressed in future publications.

This paper introduces the extended cross correlation, emphasizing simple, concrete examples. Section II describes the motivation for the XCC, and illustrates its use for identifying patterns that are repeated in two synthetic spectra. Section III illustrates the use of the XCC with real experimental data, namely, spectra of mixtures of deuterated ammonia isotopomers. This data set is used to illustrate how the XCC can be used to isolate spectra of individual species within spectra of mixtures of species, as described in Example 1 above. We conclude in Sec. IV with comments on the strengths of the XCC technique.

This paper is the first in a two-part series. The companion paper that follows illustrates the power and generality of the XCC technique, with an emphasis on a careful delineation of the applicability of the technique. Expressions are presented which permit the identification of an arbitrary number of patterns in an arbitrary number of spectra, and the example of mixtures of deuterated ammonia isotopomers, which is introduced in this paper, is given a more thorough treatment with an expanded arsenal of techniques.

II. AN INTRODUCTION TO THE EXTENDED CROSS-CORRELATION TECHNIQUE

We introduce the extended cross correlation function (XCC) by applying it to a synthetic data set that illustrates one of the simplest applications of the XCC: The partitioning of spectra of mixtures of chemical species into separate spectra of each species. In this application, the patterns to be identified are the spectra of the individual species.

In Fig. 1(a) we define two patterns (the spectra of two individual chemical species). In Fig. 1(b) we depict two synthetic spectra that are generated by taking two different linear superpositions of the patterns in Fig. 1(a). That is,

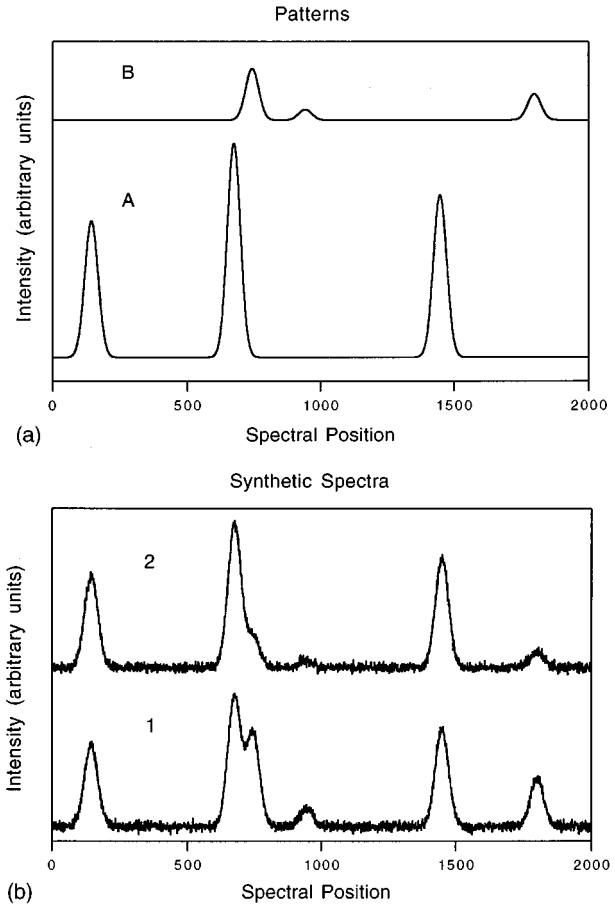

FIG. 1. Synthetic spectra used for illustration of the XCC technique. Linear combinations of two patterns plus noise generate two spectra. (a) shows patterns A and B, which contain features with Gaussian line shapes of half width at half maximum (HWHM) of 30. (b) shows the synthetic spectra 1 and 2 constructed by taking two different linear combinations of the patterns, and adding in Gaussian random noise.

$$I_1 = a_1 I_A + b_1 I_B, \qquad I_2 = a_2 I_A + b_2 I_B, \tag{1}$$

in which we adopt the convention of using numbers to label spectra, and letters to label patterns. The coefficients a_1, a_2, b_1, and b_2 describe the intensities of the patterns in each spectrum, and in this particular example take the values of 0.99, 1.1, 3.0, and 1.0, respectively. In addition, to make the spectra resemble real, experimental data sets, Gaussian random noise is superimposed upon each of the synthetic data sets.

We define in Fig. 2 a recursion map for the two synthetic spectra. This recursion map is the central conceptual underpinning of the XCC. The recursion map in this case is two dimensional, with the coordinates representing the intensity values in the two spectra. That is, the spectra are represented on the recursion map by plotting each spectral element of the entire data set as a point; the coordinates of the point are the intensities in the two spectra of the given resolution element.

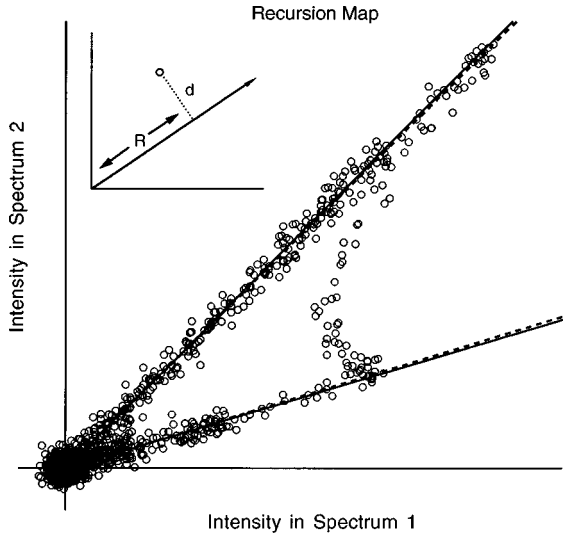

FIG. 2. Recursion map of the spectra in Fig. 1(b). The axes of the recursion map represent intensity values in each of the two spectra. The inset shows the (R,d) coordinates that are used in defining the XCC merit function. Ratio directions optimized from the merit function are shown as dashed lines, and the ratio directions used in creating the spectra as solid lines.

No information about spectral positions appears on the recursion map. The points on the recursion map can be categorized as follows:

1. *Points near the origin.* These points correspond to low intensities in both spectra. Although these points may have some signal content, this signal content is too weak relative to the noise to be useful in identifying patterns. The scatter of these points about the origin is due to the Gaussian random noise that is added to the synthetic spectra.
2. *Points that cluster about ''rays'' that pass through the origin.* The points that scatter about these rays have signal content that can be associated with one of the two patterns. That is, these points correspond to resolution elements in the spectra that lie on spectral features which are not overlapped. The scatter of the points about the rays is due to noise. The most distant points from the origin represent the strongest features in a pattern.
3. *Points that cross between, and possibly through, rays.* These points are generated where patterns overlap, by the spectral elements in the overlapped region.

For the goal of identifying patterns from the synthetic spectra, the points in category two are of the greatest interest. The presence of two rays of points in the recursion map clearly indicates that two patterns are present in the data set. The upper of these rays comprises points that are well-described by

$$I_1 \approx a_1 I_A , \qquad I_2 \approx a_2 I_A , \tag{2}$$

while the lower ray comprises points well-described by

$$I_1 \approx b_1 I_B , \qquad I_2 \approx b_2 I_B . \tag{3}$$

Prior to defining a numerically rigorous technique for identifying the patterns, it is clear that one could crudely identify which features in the spectra correspond to which patterns simply by partitioning the points into those that scatter about one or the other of the two rays.

The task of the XCC is to provide a numerically rigorous and optimal method for this process of identifying the patterns in the spectra. Following Eqs. (2) and (3) above, we consider each pattern to be defined by sets of resolution elements in which the ratio of intensities in the two spectra is nearly constant. This ratio of intensities we refer to as the ratio direction, and each pattern contained in the spectra can be uniquely labeled by a ratio direction. The simplest numerical definition of the ratio direction for a given pattern is the slope of the ray of points that define the pattern. In the case of the synthetic data, we know the ratio directions *a priori*, and we can express them in terms of the coefficients a_1, a_2, b_1, and b_2 defined in Eq. (1). Specifically, pattern A has a ratio direction of a_2/a_1 (1.11) and pattern B a ratio direction of b_2/b_1 (0.33).

In experimental data, however, the ratio directions are not known *a priori* and it is the task of the XCC to determine an unbiased estimate of the ratio direction for each pattern. At first glance, conventional least-squares fitting algorithms might appear to be appropriate, since finding an unbiased estimate of a ratio direction is equivalent to finding the slope of a best-fit line that is constrained to pass through the origin. However, linear least-squares fitting is a global optimization technique in the sense that it determines one set of model parameters which best describes all of the data. By contrast, we desire, for the synthetic data, unbiased estimates of two ratio directions. Linear least-squares fitting with the recursion map data results in a best-fit line with a slope of 0.91, in between the two correct ratio directions. Obviously, since least squares provides a single best-fit line, the ''best-fit slope'' does not provide a good estimate for either ratio direction.

From a different perspective, least-squares fitting is undesirable for the purpose of obtaining estimates of the ratio directions because of its well-known sensitivity to outliers. Least-squares fitting uses the chi-squared statistic as the figure-of-merit function; since chi-squared is defined as the sum of squares of deviations from the model, outliers strongly influence the best-fit parameters. Thus, when attempting to estimate the ratio direction for pattern A in the synthetic data, for instance, all of the points which are determined primarily by pattern B would be outliers in the least-squares fit, and *vice versa*.

Least-squares fitting has become firmly entrenched in spectroscopic practice. As a result, alternative merit functions often are not considered. However, other classes of merit functions, which minimize the effects of outliers but still provide an unbiased estimator of the desired parameters, have been used in optimization on entire data sets. Fitting

techniques that are based on merit functions that are influenced by outliers to a lesser degree than chi-squared are often referred to as robust fitting techniques; one common robust technique uses as a merit function the sum of the absolute deviations from the model.[11] A special class of robust estimators is referred to as redescending robust estimates.[12] These merit functions consist of point-by-point sums of weight functions which have small magnitudes for outliers and larger magnitudes for points that are well described by the model, which is the opposite of the chi-squared merit function used in least-squares fitting. A redescending robust estimator is desirable for the task of identifying the two model ratio directions in the recursion map precisely because extraction of more than one model estimate is desired.

The XCC is based on a redescending robust estimate, which we label G, which in the case of two data records takes the form

$$G(\alpha)=\sum_i g_i(\alpha)=\sum_i R_i * \exp(-d_i^2/2V_d). \tag{4}$$

Since the ''fit line'' is constrained to pass through the origin, the merit function is taken to be a function of just one parameter, α, which represents the ratio direction. In practice, α may represent either the slope of the fit line, or, equivalently, the angle between the fit line and one of the axes. The sum over i represents a sum over all resolution elements (all points on the recursion map). We refer to the g_i as weight functions; thus, the merit function takes the form of a sum of weight functions which are computed for each point on the recursion map.

The weight function in Eq. (4) consists of a product of two terms. We discuss the second term first, which takes the form of a Gaussian function of d, which represents the distance of any point in the recursion map from the fit line. Thus, points which are more distant from the fit line are weighted less than those near the fit line. This gives the merit function the property that it can estimate ratio directions for more than one pattern (i.e., this second term in the weight function makes the merit function a redescending robust estimate). V_d represents the expected variance of a point on the recursion map along the d direction. If the noise amplitude in the two spectra is identical and independent of intensity, then $V_d=\sigma_0^2$, where σ_0 is the baseline noise amplitude associated with the spectra. By incorporating V_d into the weight function, not only are points that are irrelevant to the fit automatically excluded, but the weights of each point are also determined in a statistically optimal manner, based on knowledge of the noise in the spectra, thus providing an accurate fit.

The first term in the weight function, R, is simply the projection of the point on the recursion map onto the fit line (see the inset in Fig. 2). The justification for the form of this term is less rigorous than that for the second, Gaussian term, and rests on the assumption that the resolution elements in the spectra with the strongest intensities are the least likely to be corrupted by overlap with other patterns, or by noise or other experimental artifacts. Some function of R could, in theory, be substituted for R; other options have been discussed previously.[1] However, we have found that simply using R in the weight function provides accurate estimates of pattern ratio directions in tests on synthetic data with many different characteristics. In addition, as we demonstrate below, the inclusion of R in the weight function guarantees that plotting the weight functions for the points with the ratio direction held at one of the optimum positions replicates, to a good approximation, the pattern associated with that ratio direction.

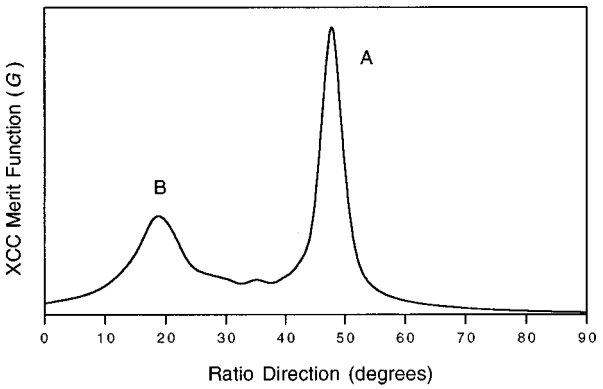

FIG. 3. XCC merit function [Eq. (4)] computed for the data set of Figs. 1 and 2 for ratio directions making angles between 0° and 90° with the axis for intensity in spectrum 1. Locating ratio directions in the recursion map of Fig. 2 is equivalent to finding peaks in the XCC merit function.

Note that R, as we have defined it, can have both positive and negative values. The only points which generate negative values of R are those which fall into category one (baseline noise points) and have negative values associated with one of the spectra. It can be shown that for any ratio direction, one-half of all baseline points will have positive values of R and one-half will have negative values of R, which results in a suppression of the effects of baseline noise points on the merit function.

Figure 3 shows the XCC merit function as a function of ratio direction for the simulated spectra, using the known variance of the added noise. Two maxima are observed in the merit function at 18.8° and 47.8°, and are marked as dashed lines on Fig. 2. They differ only slightly from the ratios used to construct the spectra (18.4° and 48.0°), which are marked as solid lines in the figure.

With the number of patterns and the ratio directions identified, it is now possible to assign spectral features to patterns. Several approaches to this task are feasible. Among these are the following two:

1. *XCC weights method.* In this method, the value of the weight function at one of the pattern ratio directions is plotted for each spectral element. Since the weight functions are largest for those points which are well described by intensity derived from only one pattern, it is straightforward to identify which features in the spectra are assignable unambiguously to one pattern. These weights are approximately linear in the intensity within a

single pattern and suppress contributions from other patterns. Smoothing the weights reduces noise-induced fluctuations in the weights.

2. *Inversion method.* Note that Eq. (1) is invertible; the patterns can be determined from the spectral intensities if the coefficients a_1, a_2, b_1, and b_2 are known. Having used the XCC to estimate the pattern ratio directions, these coefficients can be assigned. Although it may appear that we are attempting to use two pattern ratio directions to determine four coefficients, two of the coefficients, such as a_1 and b_1, can be assigned arbitrary values; this is equivalent to introducing arbitrary scaling factors for the patterns, I_a and I_b.

The results from technique 1 are shown in Fig. 4(a). The weights evaluated at the maxima in Fig. 3 clearly identify features in the original spectra as belonging to one or the other of the patterns. An overlapped feature is correctly separated into two components. Figure 4(b) shows the weights smoothed by convolution with a Gaussian line shape with a width equal to one-half of the width used in constructing the data set, resulting in ''reconstructed patterns'' which resemble quite closely the original patterns used to construct the synthetic data.

The results of linear inversion are shown in Fig. 4(c). Note that the signal-to-noise in the reconstructed patterns using the linear inversion technique is lower than in the synthetic spectra. This ''noise amplification'' effect is generic to the linear inversion technique, and can be understood by consideration of two extreme cases:

1. If the ratio directions for the patterns are identical, then the patterns are indistinguishable (in essence, the signal-to-noise of the patterns after linear inversion is zero, and the noise amplification effect is infinite).

2. If the ratio directions for the patterns are 0° and 90°, then the patterns are already separated in the spectra, and no linear inversion is necessary. The signal-to-noise of the patterns is identical to that of the spectra (the noise amplification effect is zero).

Thus, it is clear that this noise amplification effect will increase when the ratio directions for the two patterns are closer together, and decrease for ratio directions that are further separated. A mathematical treatment of the noise amplification effect can be found in paper II of this series.

Note, however, that the linear inversion method reconstructs the line shapes, line positions, and intensities of the original patterns to a much better approximation than the weights method, even with smoothing. Thus, because the ratio directions are determined by the least overlapped part of strong features, linear inversion allows determination of the line shape and correct intensity of features which are completely obscured by overlap.

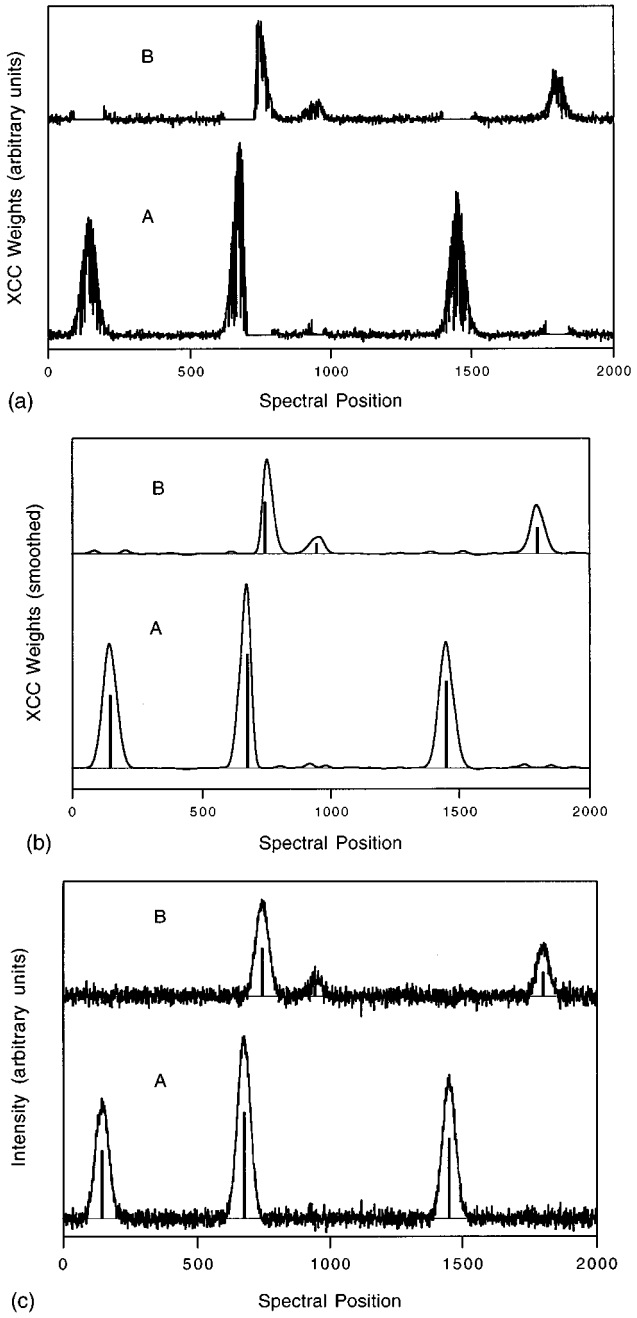

FIG. 4. Reconstruction of the patterns from the spectral data. (a) shows the results of the ''weights method'', and (b) depicts the data in (a) after convolution with a Gaussian to reduce the noise and replicate approximately the line shapes in the spectra in Fig. 1. (c) displays the results of the linear inversion method. Linear inversion results in a worse S/N ratio than the weights method (a), but provides a better line shape for overlapped features. The vertical bars in (b) and (c) represent the positions and intensities of the features in the patterns used to create the synthetic spectra.

III. APPLICATION TO SPECTRA OF NH_3/ND_3 MIXTURES

To illustrate the application of the XCC to real experimental data, we use the technique to extract the spectra of pure isotopomers from the infrared spectra[13] of mixtures containing ND_3, ND_2H, NDH_2, and NH_3. Because H and D exchange rapidly in these mixtures, it is not possible to obtain spectra of the pure species directly. Separation and analysis of these spectra would make a considerable contribution to the understanding of the potential-energy surface of

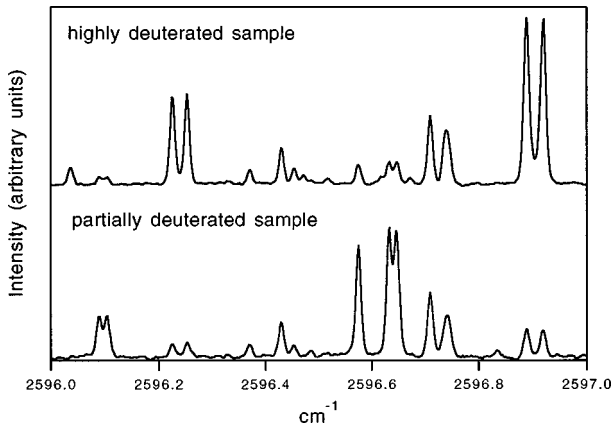

FIG. 5. Two infrared spectra of mixtures of deuterated ammonia isotopomers. The upper spectrum is of a sample of ND_3 in a sample cell pretreated with D_2O, so that ND_3 is expected to be the most abundant species. The lower spectrum is of a mixture of ND_3 and NH_3 in a 1:2 ratio in a cell pretreated with D_2O and H_2O in the same ratio.

ammonia, shedding light on the normal mode-local mode transition, stretch–bend interactions, and other vibrational couplings.[14]

Spectra of ammonia mixed isotopes are difficult to obtain with known isotopic ratios because of the strong absorption of ammonia and water on most cell surfaces. To obtain highly deuterated spectra usually requires preconditioning cell surfaces with ND_3 and/or D_2O to displace exchangeable hydrogen (H). In Fig. 5 we display a small section of two spectra obtained by Hernandez, Lehmann, and Lafferty[13] of mixtures of the ammonia isotopomers, which were prepared in two distinct fashions. The upper spectrum was obtained by introducing ND_3 into a cell that was preconditioned with D_2O. In the absence of contamination, this sample should contain only the ND_3 isotopomer; however, low levels of contamination by H often prove difficult to avoid. The lower spectrum was obtained with a sample that consisted of a mixture of ND_3 and NH_3 in a ratio of 1:2. The cell in this case was preconditioned with a mixture of D_2O and H_2O in the same ratio. This latter mixture is expected to contain all four ammonia isotopomers, the relative abundances of which can be estimated by the binomial formula.

The full spectra made available to us by Hernandez *et al.*[13] contain several thousand features and cover the entire ND and NH stretch fundamental regions. For ease of presentation, we have chosen a one cm^{-1} section between 2596 and 2597 cm^{-1} for analysis here. This region contains absorption due to the N–D stretch chromophore, and has not been analyzed in the literature, although Professor Martin Quack (ETH) has informed us that some work has been done.[15] Because the N–H stretch does not contribute in this region, we expect to find patterns due to ND_3, ND_2H, and NDH_2, only. Even in this small section of spectrum, and with only three of the four species contributing, the number of lines is large. Without a technique for labeling the lines according to which species produced them, it would be difficult to apply traditional assignment techniques like combination differences.

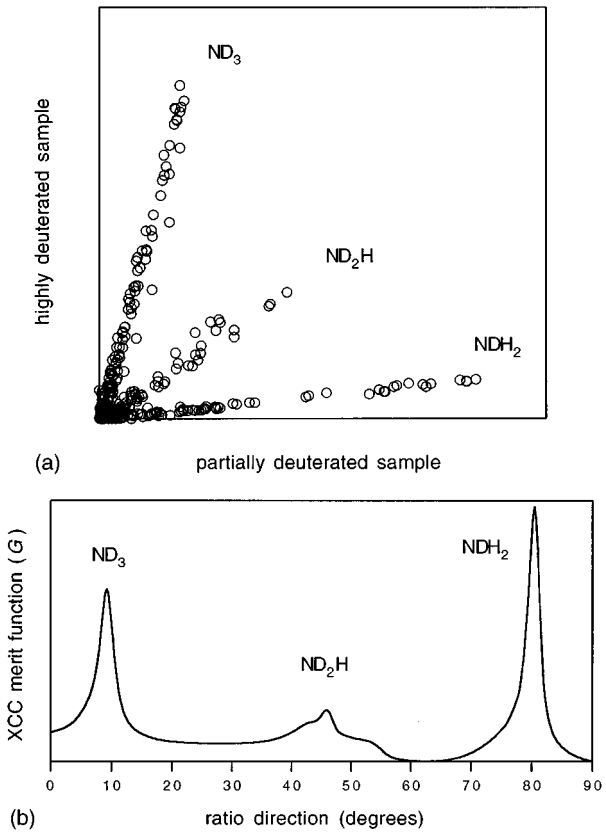

FIG. 6. Determination of the ratio direction corresponding to each contributing isotopomer. (a) displays the recursion map for the data in Fig. 5. (b) is the XCC merit function plotted as a function of the ratio direction (angle between the ''fit line'' and the x axis). Each ratio direction can be assigned to a specific isotopic species using knowledge of how the samples were prepared.

In the previous section we demonstrated how the XCC technique could be used to identify patterns that were repeated in two different spectra. In the present example, the patterns correspond to the spectra of the three isotopomers that absorb in this energy region. Figure 6(a) depicts the recursion map for the spectra in Fig. 5. Three rays of points are clearly observed, indicating three patterns, one corresponding to each contributing isotopomer. Since we expect ND_3 to contribute more strongly to the more highly deuterated sample, and NDH_2 to contribute more strongly to the less deuterated sample, we can immediately assign each of the patterns to one of the isotopomers (it is evident that the sample of ''pure ND_3'' must have been contaminated to some extent by H_2O or NH_3 due to the presence of ND_2H, and NDH_2 in this sample). The noise characteristics of the data can be estimated by inspecting the recursion map.

The XCC can be used to determine the ratio directions that correspond to each of the isotopomers, as shown in Fig. 6(b). The weights, as a function of energy for each of the three maxima in the XCC, are plotted in Fig. 7, after convolution with a Gaussian to replicate approximately the line shapes and linewidths observed in the spectra. This plot can be used in a simple fashion to identify which peaks in the spectra belong to which isotopomer. Virtually all of the

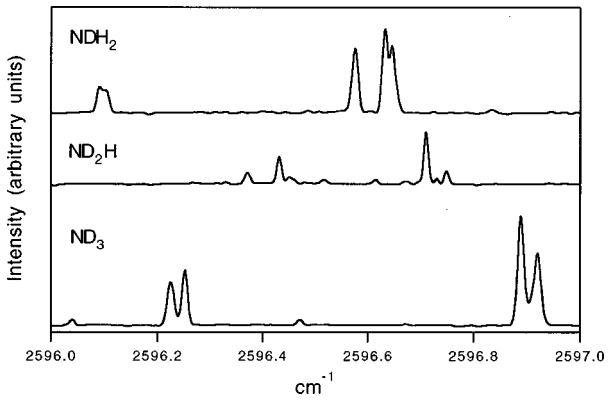

FIG. 7. Results of the XCC weights method for identifying the patterns present in the spectroscopic data. Each trace corresponds to one of the pattern ratio directions determined as a local maximum in the merit function in Fig. 6(b). This plot permits assignments of most of the lines in Fig. 5 to one of the isotopic species.

strong lines, as well as a few of the weaker lines, in the spectra can be assigned in this fashion. As we noted in the previous section, the weight functions will not necessarily accurately represent the intensities of the patterns (in this case, the intensities in the spectra of the individual isotopomers). For example, two ND_3 doublets are observed in the spectra with a splitting of $\sim 0.04\ cm^{-1}$ and nearly equal intensities for each member of the doublet. The splitting of the doublets is convincingly replicated in the weight functions for the ND_3 pattern, but the intensities are no longer equal. Because the discrepancy in intensities observed in the weights is fairly small, this effect can be accounted for by random statistical behavior.

In other cases, discrepancies in the intensities observed in the weights may contain diagnostic information. For instance, the first two moderately intense peaks just below $2596.8\ cm^{-1}$ are both assigned by the XCC weights method to be ND_2H peaks. The ''intensities'' of these two peaks in the weights for the ND_2H pattern, however, do not match the ratio of intensities observed in the spectra. In this case, the discrepancy is rather large, and is unlikely to be accounted for by statistical fluctuations. It is possible that the second peak (the nearest one to $2596.8\ cm^{-1}$) actually does not belong to ND_2H, but actually arises from some impurity in the sample, such as HOD, which might happen to have a similar, but not identical, ratio direction. Another possibility is that the peak does arise from ND_2H, but that the intensities in the spectra are ''corrupted'' by overlap with a small peak from another species, or by some experimental artifact. These various possibilities cannot be evaluated with the data available; in any case, the assignment of this peak should be viewed with suspicion. Thus, the XCC method identifies which features are securely apportioned between patterns and which features remain problematic.

The second technique that we described in the previous section for partitioning the spectra into patterns was entitled ''linear inversion.'' The strength of the linear inversion technique is its ability to determine accurate intensities and lineshapes for the patterns, even when features from more than one pattern overlap. Unfortunately, the inversion technique is not directly applicable to the analysis that we have just performed because we wish to recover three patterns from two spectra (the inversion is underdetermined). In fact, Hernandez *et al.* did record a third spectrum of mixed ammonia isotopomers with a deuterium fraction intermediate between the two presented in Fig. 5. Thus, it can be envisioned that a linear inversion from three spectra to three patterns would be possible if the XCC could be defined for three spectra. In fact, the XCC can be generalized in a straightforward fashion for any number of spectra, but this generalization requires the introduction of somewhat more elaborate notation and will be the focus of the second paper in the series.

IV. DISCUSSION

In this paper we have introduced a powerful pattern recognition technique, entitled extended cross correlation (XCC), that permits the identification of patterns that are repeated in multiple spectra. The XCC can be applied in a model-free way, meaning that the form and number of patterns to be identified can be completely unknown at the outset. The XCC permits the identification of multiple patterns within a set of spectra, including the possibility of identifying larger numbers of patterns than the number of spectra. Finally, the XCC takes into account knowledge about noise in the spectroscopic data in a natural fashion.

The XCC is similar in spirit to several other pattern recognition techniques that have been reported, particularly those that are based on principal component analysis (PCA), such as classification analysis[16] and iterative target transformation factor analysis (ITTFA).[17,18] Another pattern recognition technique with similar applications that has recently been brought to our attention is ''covariance mapping.''[19] These techniques start from the same fundamental assumption as the XCC: That a set of spectra can be considered to be linear superpositions of patterns. Another similarity is that each of these techniques can, in principle, be used to determine the number of patterns that are contained in a data set without any *a priori* knowledge. However, in the case of PCA-based techniques, the ''patterns'' that are obtained directly from principal component analysis generally do not have any physical meaning, although techniques have been reported that permit the transformation of the abstract principal components into physically meaningful patterns.[17,18] In addition, the successful use of both the PCA-based techniques and the covariance mapping technique generally requires large numbers of spectral inputs. In this respect, XCC provides an attractive alternative to these other statistical pattern recognition techniques. In cases in which spectra do not consist primarily of well-resolved features, however, PCA-based techniques may hold an advantage; the range of applicability of the XCC technique and its relationship with PCA-based techniques are described in greater detail in paper II of this series.

The examples of the application of the XCC technique that we have presented are simple ones, and it would be possible to identify by eye the patterns that are present in

both the synthetic data in Sec. II and the deuterated ammonia isotopomer data in Sec. III. However, the ammonia spectra that were presented in Sec. III represent only a small fraction of the total available spectra. The spectra extend over hundreds of cm^{-1}, and the numerical pattern recognition techniques that we have presented can easily be automated to provide isotopomer assignments for most of the lines in the entire data set. In addition to avoiding tedious analysis of large data sets, a numerical pattern recognition technique, such as the XCC, can also be particularly useful for analyzing complex data sets, which consist of many spectra and/or contain a high density of overlapping peaks. The use of the XCC to identify patterns in an arbitrary number of spectra will be described in paper II of this series.

ACKNOWLEDGMENTS

This research has been supported by AFOSR grants F49620-94-1-0068 and F49620-97-1-0040. M.P.J. thanks the Department of the Army for support under a National Defense Science and Engineering Graduate Fellowship. We are grateful to Professor Rigoberto Hernandez of the University of Pennsylvania, Professor Kevin Lehmann of Princeton, and Dr. Walter Lafferty of National Institute of Standards and Technology (NIST) for use of the ammonia mixed isotopes spectra. We thank Professors David Hercules and Joel Tellinghuisen for introducing us to the ITTFA technique.

[1] S. L. Coy, D. Chasman, and R. W. Field, in *Molecular Dynamics and Spectroscopy by Stimulated Emission Pumping*, edited by H.-L. Dai and R. W. Field (World Scientific, 1995).

[2] D. M. Jonas, S. A. B. Solina, B. Rajaram, R. J. Silbey, R. W. Field, K. Yamanouchi, and S. Tsuchiya, J. Chem. Phys. **99**, 7350 (1993).

[3] M. E. Kellman, J. Chem. Phys. **93**, 6630 (1990).

[4] M. E. Kellman and G. Chen, J. Chem. Phys. **95**, 8671 (1991).

[5] L. E. Fried and G. S. Ezra, J. Chem. Phys. **86**, 6270 (1987).

[6] S. A. B. Solina, J. P. O'Brien, R. W. Field, and W. F. Polik, Ber. Bunsen-Ges. Phys. Chem. **99**, 555 (1995).

[7] S. A. B. Solina, J. P. O'Brien, R. W. Field, and W. F. Polik, J. Phys. Chem. **100**, 7797 (1996).

[8] S. J. Lipson, R. B. Lockwood, P. S. Armstrong, D. L. Vititoe, and W. A. M. Blumberg (unpublished data).

[9] J. P. O'Brien, M. P. Jacobson, J. J. Sokol, S. L. Coy, and R. W. Field (in preparation).

[10] M. P. Jacobson, S. L. Coy, R. W. Field, S. J. Lipson, R. B. Lockwood, P. S. Armstrong, D. L. Vititoe, and W. A. M. Blumberg, (in preparation).

[11] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in FORTRAN: The Art of Scientific Computing*, 2nd ed. (Cambridge University Press, New York, 1994), Chap. 15.7.

[12] W. A. Stahel, A. F. Ruckstuhl, P. Senn, and K. Dressler, J. Am. Stat. Assoc. **89**, 788 (1994).

[13] The spectra of ammonia mixed isotopes were recorded by Rigoberto Hernandez and Kevin Lehmann of Princeton, and Walter Lafferty of NIST. We thank them for providing the spectra for use in this work.

[14] Vibrational analysis based on NH_3 spectra are discussed in S. L. Coy and K. K. Lehmann, Specterochim. Acta. **45**A, 47 (1989); K. K. Lehmann and S. L. Coy, J. Chem. Soc., Faraday Trans. 2, **84**, 1389 (1988).

[15] Professor Martin Quack (Eidgenossische Technische Hochschule, Zurich) (private communication), reports that some analysis has been done on the mixed ammonia isotope spectra in the stretch fundamental region.

[16] M. Meloun, J. Militky, and M. Forina, *Chemometrics for Analytical Chemistry, Vol. 1: PC-Aided Statistical Data Analysis* (Ellis Horwood, New York, 1992).

[17] D. M. Hercules, M. Houalla, A. Proctor, and J. N. Fiedor, Analytica Chemica Acta **283**, 42 (1993).

[18] J. N. Fiedor, A. Proctor, M. Houalla, and D. M. Hercules, Surf. Interface Anal. **20**, 1 (1993).

[19] L. J. Frasinski, K. Codling, and P. A. Hatherly, Science, **246**, 1029 (1989).