Anal. Chem. **2001,** *73,* 321−331

Comparison of the Performance of Different Discriminant Algorithms in Analyte Discrimination Tasks Using an Array of Carbon Black−Polymer Composite Vapor Detectors

Thomas P. Vaid, Michael C. Burl,[†] and Nathan S. Lewis*

Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125

An array of 20 compositionally different carbon black−polymer composite chemiresistor vapor detectors was challenged under laboratory conditions to discriminate between a pair of extremely similar pure analytes (H_2O and D_2O), compositionally similar mixtures of pairs of compounds, and low concentrations of vapors of similar chemicals. Several discriminant algorithms were utilized, including *k* nearest neighbors (*k*NN, with $k = 1$), linear discriminant analysis (LDA, or Fisher's linear discriminant), quadratic discriminant analysis (QDA), regularized discriminant analysis (RDA, a hybrid of LDA and QDA), partial least squares, and soft independent modeling of class analogy (SIMCA). H_2O and D_2O were perfectly classified by most of the discriminants when a separate training and test set was used. As expected, discrimination performance decreased as the analyte concentration decreased, and performance decreased as the composition of the analyte mixtures became more similar. RDA was the overall best-performing discriminant, and LDA was the best-performing discriminant that did not require several cross-validations for optimization.

A. Background and Goals. Arrays of broadly responsive detectors, in conjunction with pattern recognition algorithms, have attracted significant recent attention for use in vapor detection.[1] Such detector arrays have been shown to allow identification, classification, and in some cases quantification of various organic vapors.[1] Unlike traditional "lock and key" chemical sensing, in the array approach an individual sensor need not be highly selective toward the analyte of interest. Instead, variations in the pattern of responses produced by the detector array are used to differentiate between various analytes.

The ability of such detector arrays to discriminate between various analytes comprises one figure of merit for the sensing system as a whole. This figure of merit is analogous to the selectivity ratio of an individual, traditional chemical sensor for the target analyte relative to interferences, because when only one channel of data is available, the performance of a sensor system is identical to the performance of the sensor. However, one broadly responsive detector gives no information about an unknown analyte presented at an unknown concentration. In contrast, two differently responding, only partially correlated, detectors that each respond linearly to analyte concentration will yield a unique quantity, the ratio of their signals, for any given analyte. When an array of *n* detectors is exposed to an analyte, it generates *n* responses, which can be plotted as a single point in *n*-dimensional space. A set of exposures to a given analyte at a given concentration will yield a set of points in detector space, which are separated only by the variations in the detector responses. "Training" the array with many exposures to many known analytes will lead to several clusters, one for each analyte. Various discriminant algorithms can then be used to assign a single exposure of an unknown analyte to one of the clusters obtained from the training set, thus identifying the unknown with a specific probability of success.

Clearly, when an array approach to sensing is used, the system-level discrimination performance not only is a function of the detector performance but also is related to the performance of the accompanying data-processing algorithm. Expressions for the signal-to-noise ratio, sensitivity, and selectivity of a detector array system have been given by Lorber[2] and utilized by Kowalski and co-workers.[3,4] Previous studies have compared the performance of some of these algorithms on both real chemical data and simulated data.[5−7] The goal of this work is to evaluate the performance of various data-processing algorithms on a specific vapor detector array used in some several, relatively demanding discrimination tasks.

The detector arrays in the present study are formed from chemically sensitive resistors. Each detector material consists of regions of a conductor interspersed into regions of an insulating organic polymer. Typically the conductor is carbon black, and the dc electrical resistance of the composite is modulated by the swelling of the polymer that results from sorption of the analyte vapor.[8] Diversity in the response of the various detectors in the

[†] Jet Propulsion Laboratory, 4800 Oak Grove Blvd., Pasadena, CA 91125.

(1) Albert, K. J.; Lewis, N. S.; Schauer, C. L.; Sotzing, G. A.; Stitzel, S. E.; Vaid, T. P.; Walt, D. R. *Chem. Rev.* **2000**, *100*, 2595.
(2) Lorber, A. *Anal. Chem.* **1986**, *58*, 1167.
(3) Carey, W. P.; Kowalski, B. R. *Anal. Chem.* **1986**, *58*, 3077.
(4) Carey, W. P.; Beebe, K. R.; Kowalski, B. R. *Anal. Chem.* **1987**, *59*, 1529.
(5) Aeberhard, S.; Coomans, D.; de Vel, O. *J. Chemom.* **1993**, *7*, 99.
(6) Wu, W.; Mallet, Y.; Walczak, B.; Penninckx, W.; Massart, D. L.; Heuerding, S.; Erni, F. *Anal. Chim. Acta* **1996**, *329*, 257.
(7) Frank, I. E.; Friedman, J. H. *J. Chemom.* **1989**, *3*, 463.
(8) Lonergan, M. C.; Severin, E. J.; Doleman, B. J.; Beaber, S. A.; Grubbs, R. H.; Lewis, N. S. *Chem. Mater.* **1996**, *8*, 2298.

10.1021/ac000792f CCC: $20.00
Published on Web 12/15/2000

array is achieved by using different insulating organic polymers to form the composite films of the chemiresistors.

Detector arrays formed from carbon black composite chemiresistors have been shown to exhibit excellent pairwise discrimination between even closely related sets of analytes when a statistically based, linear discriminant algorithm is used to analyze the responses of 10−20 chemically diverse conducting polymer composites.[9] To compare the relative performance of various discriminant algorithms in conjunction with these detector array data, the array must be presented with pairs of analytes that will not be perfectly classified by at least some of the discriminant methods. This was not the case with pairs of simple organic vapors, all of which were essentially perfectly separated from each other, including structural isomers such as *o*- and *m*-xylene.[10] As part of this work, we have challenged a carbon black−polymer composite detector array with a pair of compounds that are very chemically similar, H_2O and D_2O.

In addition, it is of interest to evaluate the array performance on analyte mixtures. The steady-state relative differential resistance responses of the carbon black composite detectors, which serve as the descriptors that form an *n*-dimensional odor space from an *n*-member detector array, are linear with analyte concentration, and the response of a binary mixture of analytes is the response of the pure analytes weighted by the mole fraction of analytes in the mixture.[11] For each exposure, the responses of the *d* detectors can be mapped to *d* orthogonal axes. In this space, the Euclidean distance between a binary vapor mixture that is 0.5 mol fraction of each constituent and a binary mixture that is a 0.6:0.4 distribution of these same analytes should be one-tenth of the Euclidean distance between the array responses of the individual pure analytes. Several different binary mixtures of 1-propanol and 2-propanol, and of *n*-hexane and *n*-heptane, were therefore utilized as part of the present work.

Another method to decrease the discriminating ability of a detector array is to decrease the signal-to-noise ratio of the individual detectors. Delivery of low concentrations of analytes will decrease the detector signal and therefore reduce the signal-to-noise ratio, broadening the clusters relative to their separation. A number of low-concentration (≤1.0% of the vapor pressure) exposures to 1-propanol, 2-propanol, *n*-hexane, and *n*-heptane were therefore studied, and the performance of different discriminant algorithms was also assessed for these specific sensing tasks.

B. Description of Selected Discriminant Algorithms. Discriminant algorithms generally fall into two categories: parametric methods, which assume that the data have a certain distribution (usually a normal Gaussian distribution), and nonparametric methods, which make no assumptions about the underlying structure of the data. The classical parametric methods include linear discriminant analysis[12,13] (LDA, also known as Fisher's linear discriminant) and quadratic discriminant analysis (QDA). A hybrid of LDA and QDA, termed regularized discriminant analysis (RDA), has been more recently introduced.[5,14] The classic nonparametric discriminant is *k* nearest neighbors (*k*NN),[13] which has been applied to chemical data as well as to other types of data.[15] Many other classifiers have been developed, including artificial neural networks (ANN),[16] partial least-squares methods (PLS),[17] and soft independent modeling of class analogy (SIMCA).[18,19] In this work, the performance of the *k*NN, LDA, QDA, RDA, PLS, and SIMCA discriminant algorithms was compared for various analyte discrimination tasks using data from the carbon black composite detector array. Brief explanations of the various discriminant algorithms are provided below.

1. *k*-Nearest Neighbor Discriminant. The *k*NN algorithm involves calculation of the distance between the response of a test analyte and the responses of all of the examples in the training set.[13] The most commonly used distance metric is the Euclidean distance, which in two and thre dimensions is the familiar spatial distance. For an arbitrary number of dimensions, the Euclidean distance is simply

$$\text{distance}_{ij} = [\sum_{n=1}^{d}(X_{in} - X_{jn})^2]^{1/2} \qquad (1)$$

where X_{in} and X_{jn} are the coordinates of the *i*th and *j*th point in the *n*th dimension, respectively, and *d* is the number of dimensions. The test sample is then assigned to the class having the largest number of nearest neighbors to the test data. For example, if $k = 3$, the classes of the three nearest neighbors are compared, and the unknown is assigned to the class with the majority of nearest neighbors. When choosing from more than two classes, any $k > 1$ allows the possibility of a tie. For this reason, and because it has been shown that $k = 1$ is the best method for a wide variety of distributions,[20] $k = 1$ has been used in our study. It has also been shown that any classification rule, including those with information about the statistical distribution of the data, can perform at best twice as well as *k*NN ($k = 1$) in the asymptotic case in which the training set includes a very large number of examples from each class.[20] The straightforward *k*NN classifier is therefore a good benchmark against which to measure other, more sophisticated, discriminants.

***2.* Linear Discriminant Analysis.** LDA is typically taken to mean Fisher's linear discriminant.[12] The orthogonal projection of points in a *d*-dimensional space onto a line reduces the classification problem from *d* dimensions to one dimension. When the data are projected onto one dimension, it is desirable to maximize the distance between the means of the two classes being separated, while minimizing their within-class variation. Such a ratio can be expressed as a resolution factor, RF (eq 2), where δ is the distance

$$\text{RF} = \frac{\delta}{\sqrt{\sigma_1^2 + \sigma_2^2}} \qquad (2)$$

between the two class means and σ_1 and σ_2 are the standard

(9) Doleman, B. J.; Lonergan, M. C.; Severin, E. J.; Vaid, T. P.; Lewis, N. S. *Anal. Chem.* **1998**, *70*, 4177.
(10) Vaid, T. P.; Lewis, N. S., unpublished results.
(11) Severin, E. J.; Doleman, B. J.; Lewis, N. S. *Anal. Chem.* **2000**, *72*, 658.
(12) Fisher, R. A. *Ann. Eugneics* **1936**, *7* (Part II), 179.
(13) Duda, R. O.; Hart, P. E. *Pattern Classification and Scene Analysis*; John Wiley & Sons: New York, 1973.
(14) Friedman, J. H. *J. Am. Stat. Assoc.* **1989**, *84*, 165.
(15) Kowalski, B. R.; Bender, C. F. *Anal. Chem.* **1972**, *44*, 1405.
(16) Burns, J. A.; Whitesides, G. M. *Chem. Rev.* **1993**, *93*, 2583.
(17) Geladi, P.; Kowalski, B. R. *Anal. Chim. Acta* **1986**, *185*, 1.
(18) Wold, S. *Pattern Recognit.* **1976**, *8*, 127.
(19) Wold, S.; Sjöström, M. In *Chemometrics: Theory and Application*; Kowalski, B. R., Ed.; ACS Symposium Series 52; American Chemical Society: Washington, DC, 1977; p 243.
(20) Cover, T. M.; Hart, P. E. *IEEE Trans. Inf. Theory* **1967**, *IT-13*, 21.

deviations of the two classes, respectively. Fisher's discriminant finds the vector **w** onto which the data are projected that maximizes the RF. The Fisher method does not prescribe how the resulting one-dimensional data should be separated into classes. In our work, we have used a simple threshold that is derived using the assumption that the projected (one-dimensional) distributions for each class are Gaussian.

3. Quadratic Discriminant Analysis. QDA assumes a multivariate normal distribution of the data for each class.[13] A data point $\boldsymbol{x}$ is placed in the class ω_k that minimizes the value of $D_k(\boldsymbol{x})$, as given by

$$D_k(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) + \ln|\boldsymbol{\Sigma}_k| - 2\ln[P(\omega_k)] \quad (3)$$

In this equation, $\boldsymbol{\mu}_k$ is the mean vector of class ω_k, $\boldsymbol{\Sigma}_k$ is the covariance matrix of class ω_k, and $P(\omega_k)$ is the a priori probability of membership in class ω_k. The value of $P(\omega_k)$ was taken to be equal to the quantity 1/(number of classes) for all of the classes. QDA effectively measures the distance from the unknown point to the mean of a class, while normalizing for the variance in the individual measurements (dimensions). The unknown is assigned to the class with the minimum "normalized" distance, $D_k(\boldsymbol{x})$. In practice, the class-conditional mean vectors and covariance matrices are not known in advance, so these parameters are typically estimated from training data using the conventional maximum likelihood (ML) estimators.[13]

4. Regularized Discriminant Analysis. RDA minimizes the same $D_k(\boldsymbol{x})$ as is done in QDA (eq 3), but the ML estimates of the class-conditional covariance matrices are replaced with regularized estimates, $\boldsymbol{\Sigma}_k(\lambda,\gamma)$.[14] The first regularizing parameter, λ, converts the class covariance matrix to a linear combination of the class covariance matrix and the pooled covariance matrix (i.e., that of all training samples) (eqs 4−6). The second regularizing

$$\boldsymbol{Q}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} (\boldsymbol{x}_i^{(k)} - \boldsymbol{\mu}_k)(\boldsymbol{x}_i^{(k)} - \boldsymbol{\mu}_k)^T \quad (4)$$

$$\boldsymbol{Q}_p = \sum_{k=1}^{K} \frac{N_k}{N} \boldsymbol{Q}_k \quad (5)$$

$$\boldsymbol{\Sigma}_k(\lambda) = \frac{(1-\lambda)N_k\boldsymbol{Q}_k + \lambda N \boldsymbol{Q}_p}{(1-\lambda)N_k + \lambda N}, \quad 0 \le \lambda \le 1 \quad (6)$$

parameter, γ, shrinks the class covariance matrix toward a multiple of the identity matrix (eq 7). These regularizations correct for

$$\boldsymbol{\Sigma}_k(\lambda,\gamma) = (1-\lambda)\boldsymbol{\Sigma}_k(\lambda) + \frac{\gamma}{d}\mathrm{tr}[\boldsymbol{\Sigma}_k(\lambda)]\mathbf{I}, \quad 0 \le \gamma \le 1 \quad (7)$$

known discrepancies between the estimates of class distributions obtained from finite samples and the true population densities. The optimal values of λ and γ are determined by minimizing the misclassification in a leave-one-out cross-validation of all samples. The terms of eqs 4−7 are defined as follows: $\boldsymbol{Q}_k$ is the ML-estimated class-conditional covariance matrix of class ω_k, $\boldsymbol{Q}_p$ is the pooled covariance matrix, N_k is the number of objects in class ω_k, N is the total number of objects, K is the number of classes, $\boldsymbol{x}_i^{(k)}$ is the vector of the ith object in class ω_k, $\boldsymbol{\mu}_\kappa$ is the mean vector of class k, d is the number of variables (dimensions), $\mathrm{tr}[\boldsymbol{\Sigma}_k(\lambda)]$ is the trace of $\boldsymbol{\Sigma}_k(\lambda)$, and $\mathbf{I}$ is the identity matrix.

5. Partial Least Squares. A slightly different approach to classification is through the use of regression. Given a set of examples, we seek a weight vector **w** that will map each example to a desired target value. The target value is termed t_1 for class 1 and t_2 for class 2; t_1 is typically +1 and t_2 is typically −1. The parameter n_1 is defined as the number of examples in class 1, n_2 is defined as the number of examples in class 2, and n is defined as $n_1 + n_2$. If the examples from class 1 are arranged as rows in a matrix **X1** (each column is a detector) and the examples from class 2 are arranged as rows in a matrix **X2**, then **w** can be determined by solving the following multiple linear regression (MLR) problem:

$$\mathbf{t} = \mathbf{X}\mathbf{w} + \mathbf{e} \quad (8)$$

where **t** is a column vector containing n_1 rows of t_1 followed by n_2 rows of t_2. **X** is the vertical concatenation of the matrices **X1** and **X2**. The magnitude of the error vector, **e**, is minimized to solve the regression problem. Typically, the target vector and the measurements are mean-centered (and in some cases autoscaled as well). The minimum mean-squared error solution to the MLR problem is well known,[13] and is given by

$$\mathbf{w} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{t} \quad (9)$$

The effectiveness of **w** for classification can be determined by evaluating its predictive ability on new data (e.g., on a sequestered test set or on holdout examples in a leave-one-out cross-validation). If the target values are chosen as follows, $t_1 = n/n_1$ and $t_2 = -n/n_2$, then it can be shown that this approach reduces exactly to Fisher's linear discriminant.[13]

In some situations, such as when the measurements from different sensors are highly correlated or are noisy, obtaining a good weight vector through standard multiple linear regression is difficult due to the inverse appearing in eq 9. One method to resolve this problem is to perform a principal components analysis (PCA) on **X** to determine the directions that have the most variance. The data are projected onto this reduced dimensional subspace and directions with smaller variance are presumed to correspond to noise and discarded. The target values are then predicted from the projected subspace rather than from the original data. In the chemometrics literature, this approach is known as principal components regression (PCR).[21] The projected data are commonly referred to as the "score matrix".

PCR provides an alternative solution to the regression equation (eq 8) that may be better-behaved than the standard MLR solution. Partial least-squares regression is another method that provides an alternative solution to the regression equation.[17,21] The PLS method is similar to PCR, except that both the target vector and the measurements are used to determine a lower dimensional

(21) Livingstone, D. *Data Analysis for Chemists: Applications to QSAR and Chemical Product Design*; Oxford University Press: New York, 1995.

subspace from which the predictions will be made. Determination of the subspace is accomplished through an iterative procedure.[17]

6. SIMCA. The SIMCA algorithm, which was developed by Wold in the 1970s,[18,19] is based on representing each class with its own principal components model. If a class is viewed as a cloud of points in a d-dimensional space, PCA finds an orthonormal basis for the cloud. (Here we assume that the PCA is applied to mean-centered data.) The first principal component is the direction of maximum variance of the data. The second principal component is the direction of maximum variance in the subspace orthogonal to the first component, and so on. If the cloud is "thin" in some directions, the class can be accurately approximated as a linear combination of $k < d$ principal components.

In the original SIMCA formulation,[18] the distance of a point from a class was determined by the out-of-space distance, i.e., by the Euclidean distance of the point from the subspace spanned by the k principal components used to model the class. The underlying assumption was that the variances in the directions orthogonal to the PCA subspace were all equal (e.g., due to white noise). By considering the out-of-space distance relative to the average out-of-space distance observed for the training set (the training examples do not all lie exactly on the PCA subspace), the SIMCA algorithm determined whether an unknown point was well-modeled by a particular class.

In more recent formulations,[19] the SIMCA distance includes an in-space distance, as well as an out-of-space distance. The in-space distance is a measure of how well the projection of the point into the principal components subspace agrees with the projections of the known class data. The maximum and minimum values of the projected training data along each dimension of the subspace define a bounding box. SIMCA uses a slightly larger box (one standard deviation wider along each principal component direction) to represent the in-space distribution. If the projected point falls within the SIMCA box, i.e., within the "normal bounds", the in-space distance is 0; otherwise, the in-space distance is given by the weighted Euclidean distance of the point from the SIMCA box, where the weights correspond to the inverse variance along each dimension. The in-space and out-of-space distances are then combined and the unknown test point is assigned to the nearest class.

With a different definition of the in-space distance that is not based on a bounding box, but is based instead on a Gaussian model of the in-space distribution, it is readily shown that SIMCA is similar to a form of regularized QDA known in the chemometrics literature as DASCO (discriminant analysis with shrunken covariances).[7] The maximum likelihood estimates of the class covariance matrices used in standard QDA are replaced by a principal components estimate in which variances along the directions of highest variance are retained, while variances along directions of lowest variance are replaced with a constant value (related to the average out-of-space distance of the training set, which is used as a normalizing factor in SIMCA). Frank and Friedman discussed the connection between LDA, QDA, RDA, SIMCA, and DASCO in more detail.[7]

EXPERIMENTAL SECTION

A. Materials. Poly(ethylene-*co*-vinyl acetate) (70% vinyl acetate), polycaprolactone, cellulose acetate, hydroxypropylcellulose, poly(4-vinylpyridine), poly(vinyl acetate), ethyl cellulose, poly(ethylene-*co*-acrylic acid) (86% ethylene), 1,2-polybutadiene, poly(methyloctadecylsiloxane), and poly(styrene-*co*-acrylonitrile) were purchased from Scientific Polymer Products. Poly(4-vinylphenol), poly(vinyl butyral), and poly(ethylene glycol) were purchased from Polysciences. Poly(ethylene oxide), poly(ethylene-*co*-vinyl acetate) (18% vinyl acetate), poly(styrene-*co*-maleic anhydride) (50:50), poly(vinylpyrrolidone), polystyrene, and poly(methyl methacrylate) were purchased from Aldrich. The carbon black was Black Pearls 2000 from Cabot Corp. Bis(2-ethylhexyl) phthalate was purchased from Aldrich. *n*-Hexane was 99+% from Aldrich, heptane was supplied by Mallinckrodt, and 1-propanol and 2-propanol were obtained from EM Science. The H_2O was filtered through a Barnstead 18 MΩ·cm resistivity filter. D_2O was 99.9 atom % deuterium, purchased from Aldrich and used as received.

Table 1. Polymers in the 20-Detector Array

detector	polymer[a]
1	poly(ethylene-*co*-vinyl acetate) (70% vinyl acetate)
2	poly(ethylene oxide)
3	poly(vinylpyrrolidone) *P*
4	1,2-polybutadiene
5	polycaprolactone
6	poly(4-vinylphenol) *P*
7	poly(vinyl acetate) *P*
8	cellulose acetate
9	poly(4-vinylpyridine) *P*
10	poly(methyl methacrylate) *P*
11	poly(styrene-*co*-maleic anhydride) *P*
12	poly(vinyl butyral) *P*
13	hydroxypropylcellulose
14	ethyl cellulose
15	poly(ethylene-*co*-acrylic acid) (86% ethylene)
16	poly(methyloctadecylsiloxane)
17	poly(ethylene glycol)
18	poly(ethylene-*co*-vinyl acetate) (18% vinyl acetate)
19	polystyrene *P*
20	poly(styrene-*co*-acrylonitrile) *P*

[a] *P* indicates plasticization with 8% by mass bis(2-ethylhexyl) phthalate

B. Detectors and Instrumentation. Polymers were generally dissolved in tetrahydrofuran, except for poly(4-vinylpyridine) and poly(vinylpyrrolidone), which were dissolved in ethanol, and poly(ethylene-*co*-vinyl acetate) (18% vinyl acetate) and 1,2-poly(butadiene), which were dissolved in toluene. Each polymer (160 mg) was dissolved in 20 mL of its respective solvent either at room temperature or by heating to 35−40 °C for several hours. Carbon black (40 mg) was then added and the suspension was then sonicated for at least 20 min.

Corning microscope slides were cut into 10 mm × 25 mm pieces to provide substrates for the detectors. A 7−8-mm gap across the middle of each slide was masked and 300 nm of chromium and 500 nm of gold were then evaporated onto the ends of the slides to form the electrical contacts. Detectors were formed by spin-coating polymer−carbon black suspensions onto the prepared substrates. The resulting films were then allowed to dry overnight.

C. Measurements. The instrumentation and apparatus for resistance measurements and for the delivery of vapors have been described previously.[9] The array of 20 polymers listed in Table 1 was used for the measurements. All exposures were performed for a duration of 300 s and were separated by periods of 600 s of

flowing laboratory air. The first several exposures in a long series tended to give responses that were different from those of the remainder of the exposures, so the initial 40 exposures were excluded from analysis for every data set evaluated in this work. The background air contained 1.10± 0.15 parts per thousand of water vapor, but no active auxiliary control over the humidity of the solvents or over the ambient temperature of the bubblers or the detectors (generally 21.5± 1.5 °C) was performed during data collection.

1. H_2O vs D_2O. Two bubblers were filled with D_2O (labeled 1 and 3) and two with H_2O (labeled 2 and 4). For all exposures, vapors were diluted to P/P^o=0.050, where P is the partial pressure of the analyte and P^o is the vapor pressure of the analyte at room temperature. Forty exposures alternating between H_2O and D_2O were performed, and then 200 additional exposures were performed, cycling 50 times sequentially through bubblers 1−4.

2. Pairwise Resolution of Similar Analytes at Low Fractions of their Vapor Pressure. A series of 120 exposures to 1-propanol and 2-propanol were performed, with exposures alternating sequentially between each member of the pair of analytes. All exposures were initially performed at a partial pressure, P, such that $P/P^o = 0.01$ for the analyte in a background of laboratory air. Similar data were collected at partial pressures of $P/P^o = 7.5 \times 10^{-3}$, 5.0×10^{-3}, and 2.5×10^{-3}, with 120 alternating exposures to each member of the solvent pair performed at each analyte concentration. An identical exposure sequence and protocol was performed for collection of the detector response data for *n*-hexane vs *n*-heptane. The first 40 exposures in each sequence were not included in the data analysis.

3. Mixtures of Analytes. Vapor was delivered from two bubblers, one containing 2-propanol and the other containing 1-propanol. The 40 initial exposures (which were not used in the data analysis) consisted of a combination of 2-propanol at $P/P^o = 2.5 \times 10^{-2}$ and 1-propanol at $P/P^o = 2.5 \times 10^{-2}$. For data collection, exposure 1 consisted of a combination of 2-propanol at $P/P^o = 2.5 \times 10^{-2}$ and 1-propanol at $P/P^o = 2.5 \times 10^{-2}$. Exposure 2 consisted of 2-propanol $P/P^o = 2.7 \times 10^{-2}$ and 1-propanol $P/P^o = 2.3 \times 10^{-2}$; exposure 3, 2-propanol $P/P^o = 2.1 \times 10^{-2}$ and 1-propanol $P/P^o = 2.9 \times 10^{-2}$; exposure 4, 2-propanol $P/P^o = 3.5 \times 10^{-2}$ and 1-propanol $P/P^o = 1.5 \times 10^{-2}$. The series of exposures 1−4 was repeated 100 times, for a total of 400 exposures. An analogous data set was collected for *n*-hexane and *n*-heptane.

D. Data Reduction. The average of resistance readings for the 60 s immediately prior to the beginning of the exposure was used as the baseline resistance, R_b, and the average of the resistance readings for the last 60 s of the exposure was taken as the steady-state response, R_{ss}. The quantity used in data analysis was the steady-state relative differential resistance change, $\Delta R/R_b$, where $\Delta R = R_{ss} - R_b$. Data were converted to $\Delta R/R_b$ form in Microsoft Excel, while all subsequent manipulations were performed using Matlab. Original Matlab code was written to analyze the data, but the SIMCA routine was based upon one by Donald B. Dahlberg, available on the Internet at ftp://ftp.cdrom.com/pub/MacSciTech/chem/chemometrics/Dahlberg SIMCA.text.

The $\Delta R/R_b$ data were evaluated in three different forms—unnormalized and normalized by two different methods. In the first normalization (n_a), for each exposure the signal ($X_i = \Delta R/R_b$) of the ith detector was divided by the sum of the X_i signals of all 20 detectors in the array (eq 10). In the second normalization

$$\boldsymbol{X}^{(n_a)} = \boldsymbol{X} / \sum_{n=1}^{d} X_n \tag{10}$$

$$\boldsymbol{X}^{(n_g)} = \boldsymbol{X} / [\sum_{n=1}^{d} (X_n)^2]^{1/2} \tag{11}$$

(n_g), signals were divided by the square root of the sum of the squares of the signals across the array (eq 11). In three dimensions, the first normalization method maps the data onto a plane, whereas the second normalization method maps the data onto the unit sphere. Because the responses of the carbon black composite detectors to various analytes have been observed to vary linearly with concentration of the analyte in the vapor phase,[9] either normalization results in a unique, concentration-insensitive signature for an analyte of interest. The two normalizations had a very similar effect on the classification accuracy of the discriminants studied herein; therefore, only the results from n_a are presented.

Except where otherwise specified, all the discriminants were evaluated using a leave-one-out cross-validation methodology. In this procedure, one exposure (data vector) is left out of the data set and the remaining exposures are used as a training set to create the classification boundary. The left-out exposure is then classified by this rule and the classification is checked against the analyte's true class. The procedure is repeated for each member of the data set, and the rate of correct classification is a useful measure of a particular discriminant's performance.

RESULTS

A. Discrimination Between H_2O and D_2O. Figure 1 presents the average responses and standard deviations of the detectors in response to 100 exposures of H_2O and 100 exposures of D_2O. Despite the similarities in response that were expected, and observed, for these two compounds, it was possible to discriminate robustly between the light and heavy water exposures based on the small differences in response patterns that were produced on the carbon black−polymer composite chemiresistor array.

Table 2 presents the resolution factors between D_2O and H_2O obtained from Fisher's linear discriminant when each bubbler is treated as a separate class. Bubblers containing H_2O were well-differentiated from bubblers containing D_2O, with resolution factors between 8.1 and 10.1.

Interestingly, the analyte exposures from bubbler 1 were resolved from analyte exposures from bubbler 3 by a factor of 2.1, even though both contained D_2O. Similarly, analytes from bubblers 2 and 4 were both nominally H_2O, yet were resolved by a factor of 1.8. Resolution factors obtained using the LDA algorithm will never be zero with a finite sample size. Additionally, small amounts of contamination in the bubblers and lines could possibly contribute to the differences in patterns from nominally identical analytes placed in different bubblers. As a test for differences between bubblers, the exposures were divided into four sets, two each of H_2O and D_2O, but with each set containing data from a combination of two bubblers. As shown in Table 3, resolution factors between H_2O and H_2O and between D_2O and D_2O were then only 0.8 and 0.9, clearly indicating that some of

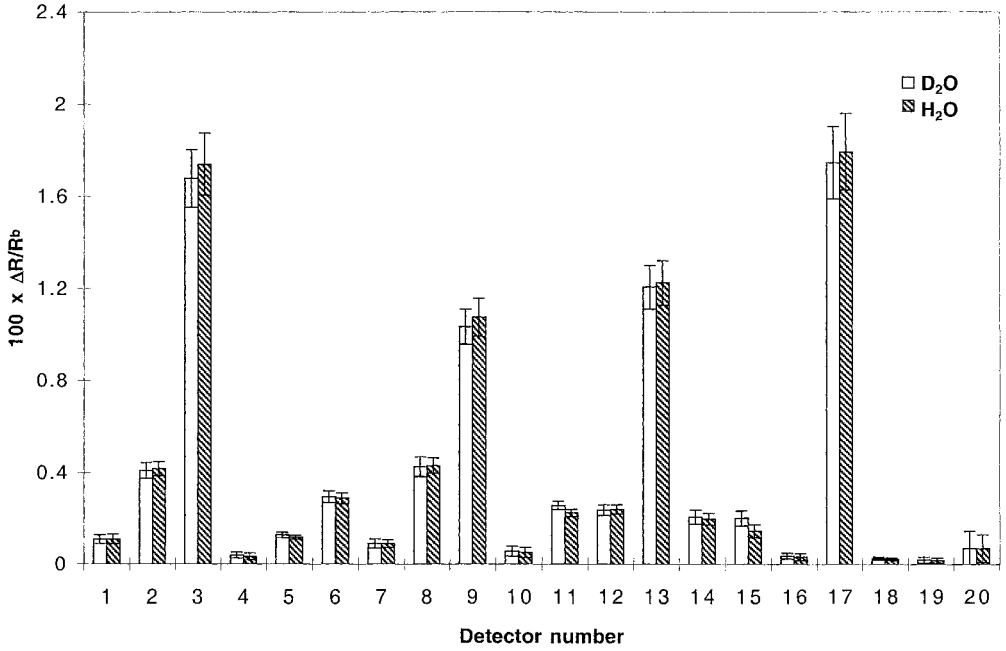

Figure 1. Steady-state relative differential resistance response, $\Delta R/R_b$, of carbon black−polymer composite vapor detectors to H_2O and D_2O (error bars are plus and minus one standard deviation). All exposures were at $P/P^o = 0.050$. Data represent means and standard deviations for 100 exposures to each analyte. The detector number indicates the polymer used to form the composite, with the detector numbering corresponding to the polymer composition listed in Table 1.

Table 2. Resolution Factors for H_2O versus D_2O Using LDA When Data from Each Bubbler Is Treated as a Separate Class

bubbler	1, D_2O	2, H_2O	3, D_2O	4, H_2O
1, D_2O	0.0			
2, H_2O	8.2	0.0		
3, D_2O	2.1	8.1	0.0	
4, H_2O	9.3	1.8	10.1	0.0

Table 3. Resolution Factors for H_2O versus D_2O Using LDA When Data Are Grouped into Four Classes, with the Two H_2O Classes Each a Random Combination of Half the H_2O Exposures and the Two D_2O Classes Each a Random Combination of Half the D_2O Exposures

analyte	D_2O	H_2O	D_2O	H_2O
D_2O	0.0			
H_2O	8.6	0.0		
D_2O	0.9	8.5	0.0	
H_2O	8.4	0.8	8.3	0.0

Table 4. Leave-One-Out Cross-Validation Error Rates for H_2O versus D_2O (Complete Data Set)

						SIMCA	
	*k*NN	LDA	QDA	RDA	PLS	12 PCs	best no. of PCs
unnormalized	0.125	0	0	0	0	0.015	0 (17)
n_a	0.37	0	0	0	0	0.005	0 (16)

the original discrmination was due to differences in what was delivered from the bubblers. The RF values for discrimination between these grouped exposures of H_2O and D_2O were still quite significant and fell in the range of RF = 8.3−8.6 (Table 3).

To further test that discrimination was occurring because of differences between H_2O and D_2O, and not because of various impurities in the bubblers or some other cause, the data were divided into two halves, one of which was used as a training set and the other of which was used a test set. The array was trained on the exposures from bubblers 1 (D_2O) and 2 (H_2O) and LDA was then used to classify the exposures from bubblers 3 (D_2O) and 4 (H_2O). All 100 of these exposures from bubbler 3 or 4 were correctly identified as either H_2O or D_2O using this procedure. Similarly, training on bubblers 3 and 4 and testing on 1 and 2 yielded 100 correct identifications. Training on 100 randomly selected exposures taken from all four bubblers and then testing on the other 100 exposures also produced perfect classification.

Table 4 presents the leave-one-out cross-validation error rates for all of the data obtained on this system. All the discriminants except for *k*NN and SIMCA (when a fixed number of principal components were used) were perfect in their classification. Normalization decreased the performance of the *k*NN algorithm, whereas it enhanced the performance of SIMCA. The degradation in performance of the *k*NN algorithm upon normalization of the response data occurred because the normalization produced less overall amplitude differences between the patterns, and the *k*NN algorithm utilized such differences in classifying the analytes.

B. Resolution of Analytes at Low Fractions of Their Vapor Pressure. 1. Form of the Data. Figure 2 shows the unnormalized response data for each detector in the array to hexane and to heptane, with each analyte at $P/P^o = 7.5 \times 10^{-3}$. Figure 3 displays similar data at an analyte partial pressure of $P/P^o = 2.5 \times 10^{-3}$. At a fixed fraction of the analyte's vapor pressure, the response patterns for hexane and heptane are quite similar, as would be expected from their similar chemical structure and properties. The mean magnitude of the response from detectors that showed significant signals when exposed to hexane (detectors 1, 2, 4, 5, 8, 12−19) decreased by a factor of 3.0 when the hexane partial pressure was decreased from $P/P^o = 7.5 \times 10^{-3}$ to $P/P^o =$

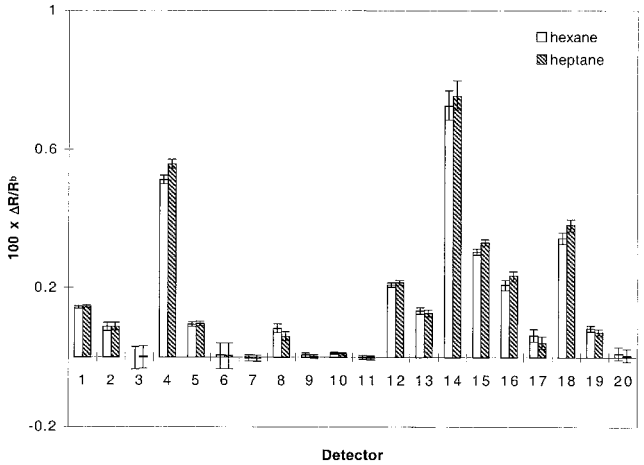

Figure 2. $\Delta R/R_b$ response of an array of carbon black−polymer composite vapor detectors to *n*-hexane or *n*-heptane at P/P^o = 0.0075. Means and standard deviations are for 100 exposures to each analyte, with exposures alternating sequentially between each member of the pair of analytes.

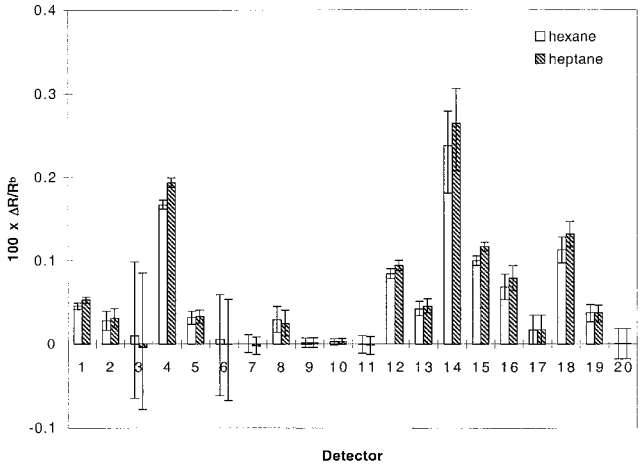

Figure 3. $\Delta R/R_b$ response of an array of carbon black−polymer composite vapor detectors to *n*-hexane or *n*-heptane at P/P^o = 0.0025. Means and standard deviations are for 100 exposures to each analyte, with exposures alternating sequentially between each member of the pair of analytes.

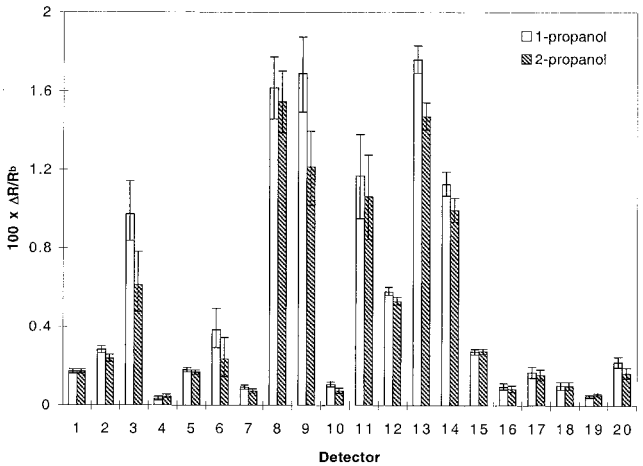

Figure 4. $\Delta R/R_b$ response of an array of carbon black−polymer composite vapor detectors to 1-propanol or 2-propanol at P/P^o = 0.010. Means and standard deviations are for 100 exposures to each analyte, with exposures alternating sequentially between each member of the pair of analytes.

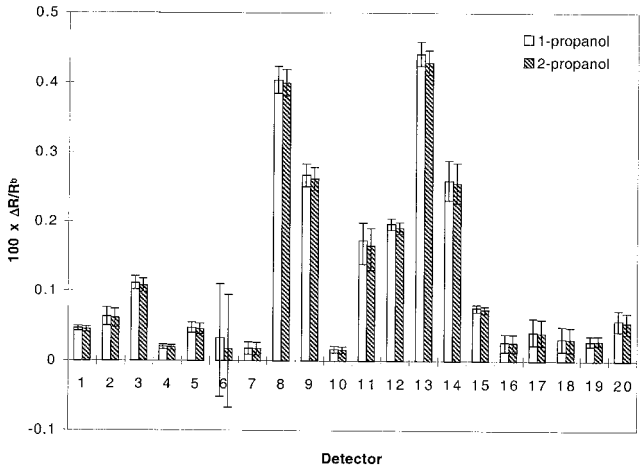

Figure 5. $\Delta R/R_b$ response of an array of carbon black−polymer composite vapor detectors to 1-propanol or 2-propanol at P/P^o = 0.0025. Means and standard deviations are for 100 exposures to each analyte, with exposures alternating sequentially between each member of the pair of analytes.

2.5×10^{-3}, whereas the same decrease in heptane partial pressure produced a mean signal decrease of 2.7 across the same set of detectors. These data are in accord with the linearity of response of carbon black composite detectors to analyte concentration that has been observed previously.[9]

In contrast, the absolute standard deviation of the responses across the set of 100 exposures was essentially constant as the analyte concentration was varied. For example, the ratio of the standard deviation of a detector's responses to hexane at P/P^o=7.5 $\times 10^{-3}$ to that at P/P^o=2.5 $\times 10^{-3}$ had an average of 1.15 across the set of detectors that responded well to hexane (1, 2, 4, 5, 8, 12−19), and this ratio had a value of 1.12 for heptane. Thus, the absolute signal strength decreased as the analyte partial pressure declined, but the absolute variance remained essentially constant, so the discrimination ability of the array is expected to become worse at lower analyte partial pressures.

A quite different situation was, however, observed for 1-propanol and 2-propanol. The absolute standard deviations decreased by an average of 3.91 for 1-propanol and by an average of 3.54 for 2-propanol when the partial pressure of these analytes was reduced from $P/P^o = 1.0 \times 10^{-2}$ to 2.5×10^{-3} (Figures 4 and 5). The main cause for the difference was not a change in random noise, but a steady drift in some of the detector responses over the course of this particular interval of data collection. The effect was more pronounced at $P/P^o = 1.0 \times 10^{-2}$ than at $P/P^o = 2.5 \times 10^{-3}$, accounting for the larger absolute standard deviation values observed at the higher analyte concentration. For illustration, Figure 6 shows the data for 100 responses of detector 8 to 1-propanol at $P/P^o = 1.0 \times 10^{-2}$ and $P/P^o = 2.5 \times 10^{-3}$, respectively. At the higher concentration, the signal drifted by 32% over 50 h, while at the lower concentration it drifted by only 10%. When a simple linear correction was applied to the data (Figure 6), the standard deviation of the higher concentration data decreased by a factor of 3.3, while that of the lower concentration data decreased by a factor of 1.3.

2. Performance of Various Discriminant Algorithms. Table 5 presents the leave-one-out cross-validation error rates for the different discriminant algorithms for the 1-propanol/2-propanol

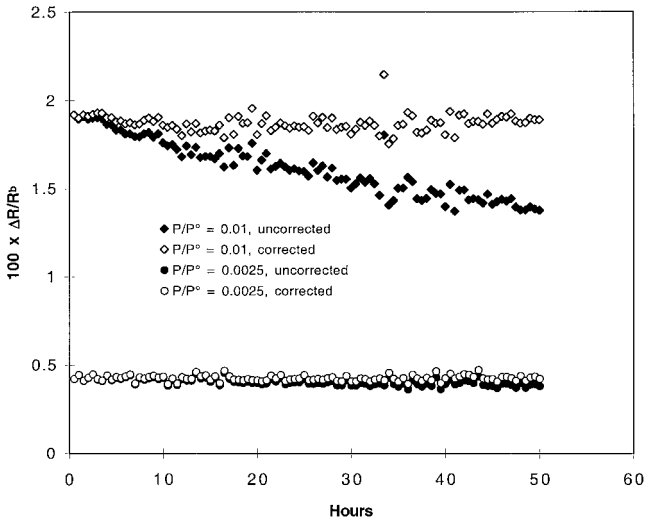

Figure 6. $\Delta R/R_b$ response of detector 8 to 1-propanol and 2-propanol at $P/P^\circ = 0.010$ (diamonds) and at $P/P^\circ = 0.0025$ (circles). Uncorrected, raw data are indicated by filled symbols, and data produced through the use of a linear correction to yield a regression line with slope of zero are indicated by unfilled symbols.

and hexane/heptane data sets. For both the 1-propanol/2-propanol and *n*-hexane/*n*-heptane classifications, the error rate increased for all discriminants at lower partial pressures of analyte. For the unnormalized data, LDA and RDA were the best discriminants (average error rates of 0.079 for hexane vs heptane) with RDA offering only a very slight improvement upon LDA. The PLS algorithm had an average error rate of 0.089, followed by QDA and optimized SIMCA at ~0.10. The *k*NN discriminant had an average error rate of 0.117, and the worst-performing discriminant was SIMCA, with an average error rate of 0.13.

The discriminants were more uniform in their leave-one-out cross-validated performance on normalized data. Once again, SIMCA and *k*NN were the worst classifiers. LDA and QDA were similar overall in their classification accuracy, but their classification performance differed somewhat in different tasks. Because RDA can vary between LDA and QDA, and necessarily chooses the best of these two limiting algorithms based upon cross-validation, RDA was the best discriminant for these normalized data.

C. Discrimination between Compositionally Similar Binary Analyte Mixtures. 1. Structure of Data. Figure 7 displays the average responses of each detector to the four different hexane/heptane binary mixtures. The detector responses exhibited a monotonic trend as the mole fraction of hexane was increased, as expected. In contrast, the response of some detectors was not monotonic for the 1-propanol/2-propanol vapor mixtures (Figure 8). Standard deviations of the detector responses for the 1-propanol/2-propanol vapor mixtures were also generally larger than those for the hexane/heptane mixtures. The larger standard deviations can be attributed to a steady change (usually a decrease) in the response of a detector observed over the course of that particular data collection interval, and the error introduced by the drift may account for the fact that the change across a series is not always monotonic, especially when comparing the very similar 50/50 and 54/46 binary mixtures of 2-propanol and 1-propanol.

2. Performance of Discriminant Algorithms. The leave-one-out cross-validation error rates for this data set are given in Table 6. For both the 1-propanol/2-propanol and *n*-hexane/*n*-heptane classifications, the error rate decreased for all discriminants as the separation in mole fraction between the analytes increased. Normalization did not have a large effect on discriminant performance. The LDA and RDA algorithms were the best-performing discriminants, with average error rates near 0.024. The RDA algorithm was nearly identical in performance to LDA and usually converged to the grid point $(\lambda,\gamma) = (1,0)$, equivalent to LDA. The PLS discriminant was almost as proficient as LDA and RDA, with average error rates of ~0.025. The other discriminants followed in the order, best to worst: QDA, optimized SIMCA, SIMCA, and kNN.

DISCUSSION

A. Discrimination between H_2O and D_2O. Although H_2O and D_2O have very similar physical properties, there are many quite measurable differences, including, for example, boiling point (100 vs 101.4 °C) and melting point (0 vs 3.8 °C).[22] Note that in Figure 1 the detectors with the largest responses (those that are most polar and hydrogen-bonding) tended to respond more strongly to H_2O than D_2O, while the converse is true of the relative responses of the less-polar polymers.

An examination of Figure 1 (and specifically the indicated standard deviations) reveals that most detectors would individually perform very poorly in distinguishing H_2O from D_2O. Detector 11 is the most discriminating individual detector, as reflected by the fact that the **w** vectors found between H_2O and D_2O always had their largest coefficients for 11. Even so, when 11 was removed from the data set, RFs of 8−10 were still obtained, and identification tests were perfect.

B. Performance of LDA and QDA. The H_2O and D_2O data do not provide an appropriate challenge for evaluating the performance of discriminant algorithms, because perfect classification was achieved for most of the algorithms investigated. Such comparisons could be made, however, for both of the experiments involving analytes at low fractions of their vapor pressure and for experiments involving compositionally similar binary analyte mixtures. In these tasks, LDA performed better than QDA. In RDA, where the floating parameter λ allows hybridization between LDA and RDA, a λ value near 1, corresponding to LDA, was generally found to be optimal. These results may at first seem surprising, because QDA is a more general classifier and because QDA reduces to LDA in the specific case when the class covariance matrices are equal. LDA simply uses the pooled covariance matrix, effectively assuming that all the class covariance matrices are equal.

If the true class covariance matrices are the same, then the two classifiers should perform identically in the asymptotic situation in which an infinite number of training examples are available and the class statistics are known exactly. However, in the present situation, the statistics must be estimated from a finite number of training examples. The QDA algorithm estimates a ($d \times d$) covariance matrix for each class, whereas LDA estimates a ($d \times d$) covariance matrix for the pooled data. The covariance estimates produced by QDA will be based on half as much data as in the LDA case and therefore are less likely to reflect the "true" covariance matrix. Also, as shown below, QDA emphasizes the differences in covariance structure between the two classes. From

(22) *CRC Handbook of Chemistry and Physics*, 67th ed.; Weast, R. C., Ed.; CRC Press: Boca Raton, FL, 1986.

Table 5. Leave-One-Out Cross-Validation Error Rates[a] for 1-Propanol versus 2-Propanol and *n*-Hexane versus *n*-Heptane at Low Concentration

						SIMCA	
$100 \times P/P°$	*k*NN	LDA	QDA	RDA	PLS	12 PCs	best no. of PCs
			1-Propanol vs 2-Propanol				
0.01	0 (0.01)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0.025)	0 (0.005)
0.0075	0.01 (0.015)	0.005 (0)	0.005 (0.01)	0.005 (0)	0.015 (0)	0.02 (0.005)	0.01 (0)
0.005	0.41 (0.495)	0.26 (0.48)	0.335 (0.41)	0.255 (0.39)	0.3 (0.495)	0.47 (0.445)	0.36 (0.395)
0.0025	0.385 (0.465)	0.35 (0.495)	0.435 (0.47)	0.35 (0.415)	0.38 (0.55)	0.44 (0.515)	0.4 (0.42)
			n-Hexane vs *n*-Heptane				
0.01	0.03 (0.01)	0.005 (0.005)	0.005 (0.005)	0.005 (0.005)	0.005 (0.005)	0.01 (0.005)	0 (0.005)
0.0075	0.035 (0.065)	0.005 (0.01)	0.01 (0.045)	0.005 (0.01)	0.005 (0.01)	0.065 (0.11)	0.03 (0.075)
0.005	0.02 (0.285)	0.005 (0.18)	0.005 (0.245)	0.005 (0.165)	0.005 (0.16)	0.01 (0.29)	0.005 (0.285)
0.0025	0.045 (0.41)	0.005 (0.35)	0.01 (0.32)	0.005 (0.29)	0.005 (0.305)	0.025 (0.36)	0.01 (0.345)
averages	0.134 (0.219)	0.079 (0.190)	0.101 (0.188)	0.079 (0.159)	0.089 (0.191)	0.130 (0.219)	0.102 (0.191)

[a] Error rate for unnormalized data; error rates for normalized data given in parentheses

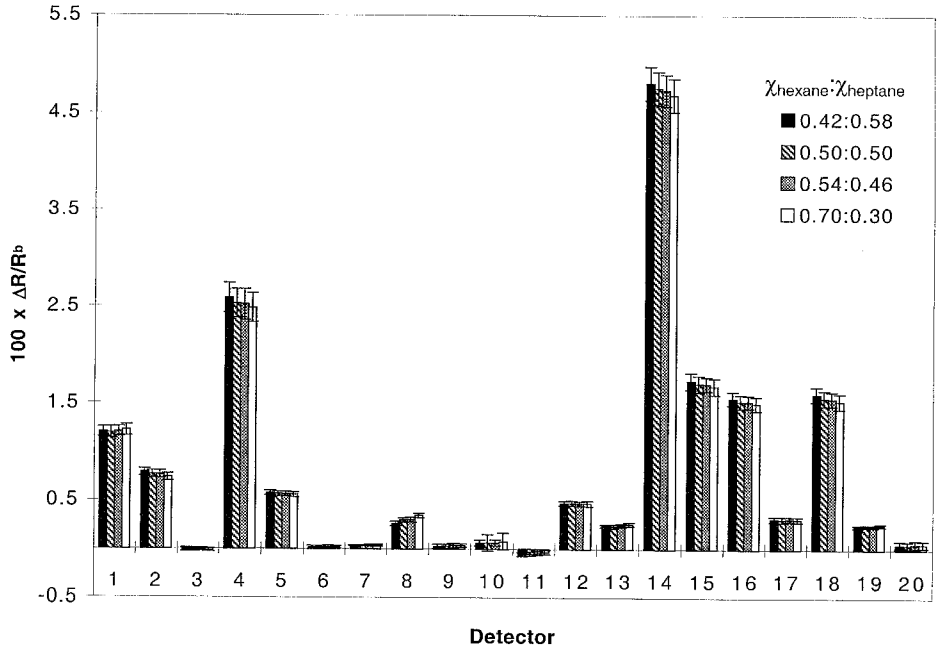

Figure 7. $\Delta R/R_b$ response of an array of carbon black−polymer composite vapor detectors to mixtures of hexane and heptane. The partial pressure of heptane is $[0.050-P/P°_{hexane}]P°_{heptane}$, where P is the partial pressure of hexane for a given exposure. The value of 20 $P/P°$ for each analyte in the mixture is indicated in the legend.

eq 3, we have

$$D_1(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x} - \mu_1) + \ln|\boldsymbol{\Sigma}_1| - 2\ln[P(\omega_1)] \quad (12)$$

$$D_2(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2) + \ln|\boldsymbol{\Sigma}_2| - 2\ln[P(\omega_2)] \quad (13)$$

After some manipulation one obtains

$$D_2(\boldsymbol{x}) - D_1(\boldsymbol{x}) = \boldsymbol{x}^T(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})x + 2(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1)^T \boldsymbol{x} + c \quad (14)$$

When $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are identical, the first term drops out and the LDA classifier is obtained. However, when $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are replaced with their estimated versions, which are not likely to be exactly equal, the first term remains, leading to suboptimal classification.

C. Performance of PLS and SIMCA. The performance of PLS tracked very closely with that of LDA. The PLS discriminant is fundamentally a form of multiple linear regression, and, as explained above, linear regression is equivalent to LDA. It is therefore not surprising that, through different algorithms for optimization, PLS and LDA give similar results. The LDA algorithm might be the preferred method because it is somewhat simpler to implement.

When compared to the other discriminants evaluated, SIMCA performed rather poorly on the discrimination tasks investigated in this work. When the model with the optimal number of principal components was chosen, 16 or 17 principal components were often found to give near-optimal (or optimal) classification accuracy. At these higher limits, SIMCA becomes somewhat similar to QDA, because it is using almost the full dimensionality of the data. Both SIMCA and QDA create a separate model for each class. In situations where the covariance matrices (size, shape, and orienta-

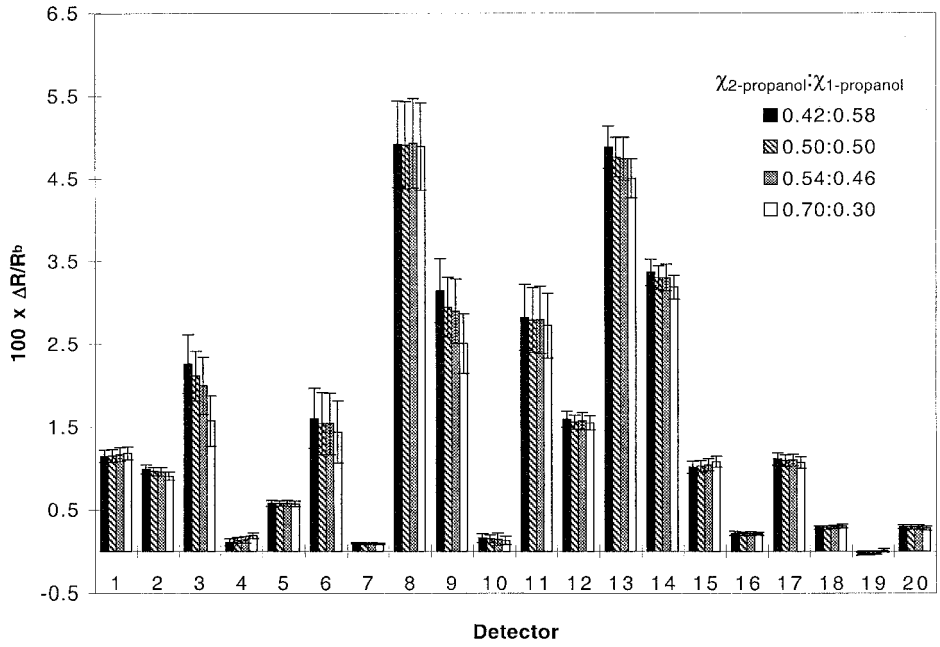

Figure 8. $\Delta R/R_b$ response of an array of carbon black−polymer composite vapor detectors to mixtures of 1-propanol and 2-propanol. The partial pressure of 2-propanol is $[0.050-P/P^\circ_{\text{1-propanol}}]P^\circ_{\text{2-propanol}}$, where P is the partial pressure of 1-propanol for a given exposure. The value of 20 P/P° for each analyte in the mixture is indicated in the legend.

Table 6. Leave-One-Out Cross-Validation Error Rates[a] for Compositionally Similar Analyte Mixtures of 1-Propanol/2-Propanol and *n*-Hexane/*n*-Heptane

						SIMCA	
Δ mixture[b]	*k*NN	LDA	QDA	RDA	PLS	12 PCs	best no. of PCs
			1-Propanol and 2-Propanol				
4	0.325 (0.165)	0.03 (0.025)	0.075 (0.08)	0.03 (0.025)	0.03 (0.025)	0.205 (0.145)	0.13 (0.12)
8	0.12 (0.105)	0.005 (0.01)	0.015 (0.02)	0.005 (0.01)	0.005 (0.01)	0.19 (0.21)	0.045 (0.065)
12	0.065 (0.045)	0 (0)	0 (0)	0 (0)	0 (0)	0.1 (0.05)	0.005 (0.005)
16	0.01 (0.01)	0 (0)	0 (0)	0 (0)	0 (0)	0.01 (0)	0 (0)
20	0.005 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.005 (0.005)	0 (0)
28	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
			n-Hexane and *n*-Heptane				
4	0.42 (0.45)	0.225 (0.21)	0.31 (0.285)	0.225 (0.21)	0.23 (0.21)	0.34 (0.365)	0.28 (0.305)
8	0.365 (0.295)	0.025 (0.04)	0.04 (0.06)	0.025 (0.01)	0.025 (0.025)	0.085 (0.13)	0.06 (0.105)
12	0.3 (0.27)	0.005 (0)	0.01 (0.05)	0.005 (0)	0.005 (0.005)	0.025 (0.045)	0.015 (0.025)
16	0.31 (0.265)	0.005 (0.005)	0.005 (0)	0.005 (0)	0.005 (0.005)	0.035 (0.025)	0.015 (0.025)
20	0.215 (0.135)	0 (0)	0 (0)	0 (0)	0 (0)	0.005 (0)	0 (0)
28	0.065 (0.01)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
averages:	0.183 (0.146)	0.025 (0.024)	0.038 (0.038)	0.025 (0.024)	0.025 (0.025)	0.083 (0.081)	0.046 (0.054)

[a] Error rate for unnormalized data; error rates for normalized data given in parentheses. [b] Δ mixture indicates the difference in mole fraction between the pairs of mixtures being discriminated, as follows: Δ = 4 refers to 50:50 (1-propanol/2-propanol) vs 46:54 (1-propanol/2-propanol); Δ = 8, 50:50 vs 58:42; Δ = 12, 46:54 vs 58:42; Δ = 16, 50:50 vs 30:70; Δ = 20, 50:50 vs 30:70; Δ = 28, 58:42 vs 30:70, and the total of P/P° was 0.050 for all mixtures. Analogous formulas apply to discrimination between mixtures of *n*-hexane and *n*-heptane, with the composition of all mixtures given in the Experimental Section.

tion of the data "cloud") of the two classes under study are very different, it is advantageous to have these separate models. However, as observed in the comparison of LDA with QDA, the data in our tasks generally consist of pairs of classes that have similar covariance matrices. There is therefore little advantage in forming two separate models.

When comparing SIMCA to the other discriminants, it is important to keep in mind the manner in which the models were formed. For LDA, QDA, and PLS, the model is created using the training data, and then unknown "test" data are classified according to the model. The situation is similar for SIMCA when 12 principal components was chosen as an approximately optimal number and used for both classes in all the tasks. In contrast, the optimized SIMCA model was customized for each classification by performing a leave-one-out cross-validation for models that used from 6 to 18 principal components. It is therefore most appropriate to compare the optimized SIMCA to RDA, which also built many models that were tested by cross-validation, and from which the best-performing model was chosen for each classification task.

Overall, QDA and RDA both outperformed SIMCA, whether it was optimized or not. Frank and Friedman discuss some shortcomings of SIMCA that may explain its relatively poor performance.[7]

D. Effects of Normalization. 1. Analytes at Low Fractions of Their Vapor Pressure. Because all pairs of vapors were delivered at the same fraction of their vapor pressure, to a first

approximation, the total response across the array should be similar for different analytes.[23] This is the case in our experiments, especially because the pairs of analytes investigated are so chemically similar. There are differences, however, with heptane giving a slightly larger total response than hexane, and 1-propanol producing a larger total response than 2-propanol. Normalization using eq 10 forces the total response across the array to be the same for every single exposure. If the response patterns of two analytes are similar but differ in magnitude, normalization will make their discrimination more difficult, and this was indeed found to be the case for both analyte pairs across all the discriminants (Table 5). However, normalization is necessary when one has no auxiliary information about the concentration of the analyte and is attempting to perform a classification/identification task for members of these analyte pairs.

2. Compositionally Similar Binary Analyte Mixtures. In contrast to the situation for pure analyte discrimination described above, for the binary mixture data, both normalization procedures led to an improvement in the performance of *k*NN, while the performance of the other discriminants was essentially unaffected by data normalization. This behavior occurs because each exposure is normalized individually, so the effects of variations in external parameters that influence all the detectors in the same way is eliminated though the normalization process. For example, if variations are present in the amount of analyte that is delivered to the array among nominally identical exposures, normalization will ideally correct all the response patterns to the same normalized pattern. Variance in detector response due to other external parameters (perhaps the temperature or the humidity of the background air) that affect the detector signals in the same direction, albeit by different relative magnitudes, will also be canceled to some extent by normalization. One large effect of this type is the drift of the detector signals over the course of the experiment. If the drift is in the same direction for all the detectors, it will be partially ameliorated by normalization. The standard deviations for individual detectors across a set of responses will decrease, but it is not clear how the classification accuracy of the discriminants will be affected.

The drift was much larger for the propanols than for the alkanes and decreased significantly for the propanols between $P/P^\circ = 0.01$ and 0.0025. The largest baseline resistance drifts of any of the sensors over the course of data collection was ~10%, and this appeared to have no correlation with the largest drifts in $\Delta R/R$. The largest downward drifts in $\Delta R/R_b$ (for propanols) were observed for hydrogen-bonding polymers, including poly(vinylpyrrolidone), poly(4-vinylphenol), cellulose acetate, poly(4-vinylpyridine), and poly(styrene-*co*-maleic anhydride), whereas the one polymer in which a significant upward drift in $\Delta R/R_b$ (for propanols) was observed was 1,2-polybutadiene.

E. Extension to Other Vapor Sensor Array Data Sets. Our experiments were carried out under controlled laboratory conditions using carbon black composite chemiresistors; thus, the conclusions regarding which discriminant performed optimally will not necessarily apply to other situations in which variations in detector responses can be produced by a variety of additional factors. For example, a hand-held detector array system that is utilized outdoors may encounter a variety of ambient temperatures, humidities, and background vapors. The resulting class covariance matrices may have a different form and relation to each other than those encountered in our experiments. We point out, however, that a 20-member array of polymer−carbon black detectors has little difficulty in distinguishing two analytes at significant fractions of their vapor pressure unless they are extremely similar (i.e., more similar than H_2O and D_2O). Therefore, the cases in which a choice of discriminant is important will occur only in classification of very similar vapors or at relatively low analyte concentrations. Training of such an array under the variety of conditions under which it will be expected to perform classifications of unknowns will presumably result in similar variances (and relationships between variances on different detectors, i.e., covariances) because the analytes themselves are so similar. The LDA algorithm, which assumes identical covariance matrices for both classes, will therefore likely perform well relative to the other discriminant algorithms evaluated in this work most situations in which the discrimination ability of such an array is challenged.

The conclusions described herein may well also apply to other polymer-based sensor arrays. Polymer-coated quartz resonators of either quartz crystal microbalance (QCM, also called thickness-shear mode resonators) or surface acoustic wave (SAW) devices[1] also utilize sorption of a vapor by the polymer film to detect an analyte. Because these methodologies also rely upon vapor sorption by a polymer film to produce a signal, the conclusions obtained above may apply to the data from such systems as well.

SUMMARY AND CONCLUSIONS

In summary, an array of 20 compositionally different carbon black−polymer composite chemiresistor vapor detectors was challenged under laboratory conditions to discriminate between a pair of extremely similar pure analytes (H_2O and D_2O), compositionally similar mixtures of pairs of compounds, and low concentrations of vapors of similar chemicals. H_2O and D_2O were perfectly separated from each other, and all 100 examples in a test set were correctly classified based on 100 examples in a training set. Discrimination performance decreased as the analyte concentration decreased, and for *n*-hexane and *n*-heptane, classification error rates on normalized data using a leave-one-out cross-validation method exceeded 18% when the analyte concentration was less than 0.005 P/P^o. Mixtures of chemically similar analytes were also robustly discriminated (error of 1% or less) when the analyte compositions differed by more than 0.006 P/P^o (and the total analyte concentration was 0.05 P/P^o), with classification error rates using the leave-one-out cross-validation method exceeding 20% only when the mole fractions of the hexane and heptane differed by less than 0.002 P/P^o in composition (and the total analyte concentration was 0.05 P/P^o). Excluding regularized discriminant analysis, which required the building and cross-validation of many models and which tended to become linear discriminant analysis under optimization, Fisher's classic linear discriminant was the best-performing method under the conditions evaluated in this work.

ACKNOWLEDGMENT

We acknowledge DARPA, the Army Research Office through a MURI grant, the Department of Energy, and NASA for support of this work.

Received for review July 10, 2000. Accepted October 5, 2000.

AC000792F

(23) Doleman, B. J.; Severin, E. J.; Lewis, N. S. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5442.