Computer Methods and Programs in Biomedicine (2005) **79**, 259–271

ELSEVIER

Computer Methods and Programs in Biomedicine

www.intl.elsevierhealth.com/journals/cmpb

Data reduction in headspace analysis of blood and urine samples for robust bacterial identification

J.W.T. Yates[a,*], M.J. Chappell[a], J.W. Gardner[a], C.S. Dow[b], C. Dowson[b], A. Hamood[b], F. Bolt[b] L. Beeby[b]

[a] *School of Engineering, University of Warwick, Coventry CV4 7AL, UK*
[b] *Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, UK*

Received 15 January 2004; received in revised form 10 April 2005; accepted 14 April 2005

KEYWORDS
Model reduction;
Bacterial identification;
Headspace analysis;
Electronic nose;
Black box models;
Intrinsic dimension

Summary This paper demonstrates the application of chemical headspace analysis to the problem of classifying the presence of bacteria in biomedical samples by using computational tools. Blood and urine samples of disparate forms were analysed using a Cyrano Sciences C320 electronic nose together with an Agilent 4440 Chemosensor. The high dimensional data sets resulting from these devices present computational problems for parameter estimation of discriminant models. A variety of data reduction and pattern recognition techniques were employed in an attempt to optimise the classification process. A 100% successful classification rate for the blood data from the Agilent 4440 was achieved by combining a Sammon mapping with a radial basis function neural network. In comparison a successful classification rate of 80% was achieved for the urine data from the C320 which were analysed using a novel nonlinear time series model.

1. Introduction

Headspace analysis involves the detection of volatile molecules produced by a liquid or solid sample. This form of chemical analysis, using gas sensors, is proving to be promising as a diagnostic tool in the medical field. Several authors (see for example [1,2]) have reported that, using such systems, it is possible to detect the presence of bacteria and provide sufficient information to discern both the species and metabolic state of these bacteria. These investigations have used commercial instruments, such as the Agilent 4440 or Cyrano Sciences C320 as a headspace analyser [3—5]; these are often referred to as electronic noses. Alternatively, gas chromatography [6] has been used for the chemical analysis of medical samples.

The aim of applying this form of analysis is to screen biological samples rapidly. Using current microbiological techniques, processing of a sample takes 24–48 h in order to characterise and identify a pathogenic organism [7]. This is not ideal, especially when rapid antibiotic therapies are crucial in many infections; these treatments are

* Corresponding author. Present address: AstraZeneca, 8AF2, Alderley Park, Macclesfield, Cheshire SK10 4TG, UK

0169-2607/$ — see front matter.
doi:10.1016/j.cmpb.2005.04.003

pathogen specific and worst still there is a growing resistance to antibiotics in the microorganism population [8,9]. It is therefore desirable to reduce the time of analysis of a sample to, ideally, minutes using a technique that is easy to apply. In this way the infection may be treated rapidly with the most suitable medication.

The task of diagnosis entails some method of discriminant analysis in order to identify the form of bacterial infection. Techniques for discriminant analysis include principal components analysis (PCA) for data visualisation, linear discriminant analysis, and neural networks of various geni and topologies [10]. In all cases these have been used in an attempt to discriminate between samples of different types; be it different species of bacteria or different metabolic states.

These models are generally identified and simulated upon a microcomputer such as a standard PC or Unix workstation. This means that, in a medical context where speed and accuracy are necessary, the computational intensity of the data analysis task is important. Robust parameter estimation is examined with respect to data from biological experiments.

In this paper data from two different headspace analysis experiments are considered. The first were from a blood screening survey, where blood has been used as the growth medium for bacteria. The blood samples were inoculated with various species of bacteria—including methicillin susceptible (MSSA) and methicillin resistant (MRSA) *S. aureus*. This organism has the propensity to develop resistance to multiple antibiotics and is thus a growing public health concern [11]. The aim was to see if this dangerous antibiotic resistant pathogen might be detectable amongst other species.

The second set of data were the result of routine medical urine sample screening. These were supplied by Walsgrave Hospital, Coventry, U.K. after they had been processed in their microbiological laboratory. These originated from patients who were asked to provide a routine urine sample at the hospital. The hospital processes a few hundred such samples every week and kindly provided an arbitrary selection for our use. The aim was to detect the presence of urinary tract infections (UTI).

In both cases electronic noses were used to analyse the samples. Electronic nose data have three important characteristics which makes preprocessing essential. The first is that the data have a predominance of intensity information [12] causing correlation between sensor responses; the second is that the system is time dependent, which becomes evident in the results of this paper; the third characteristic is that which is characterised as 'chemical noise' [13,14] or interference, which needs to be accounted for.

Biological samples contain a rich mixture of chemicals relating to the function of the human body. There will certainly be substances present that relate to processes other than those that are of interest; for example, results of metabolisation of nutrients and medication [14]. There will be a whole host of information relating to whether the patient has been active or sedentary, and other pathogenic infections elsewhere in the body. These will all be detected by the gas sensor array as background noise and so will potentially affect the data and the results of any analysis.

The aim of this investigation therefore was to apply different data reduction techniques to two very disparate data sets from a C320 electronic nose and an Agilent 4440 mass spectrometer. This was in order to analyse blood and urine samples for specific bacteria and optimise system model identification. The interaction between the reduction techniques and black box models for discrimination are investigated by comparing the observed successful classification rates of the resulting identified models. This was done in order to see which models perform 'best' with respect to discrimination and which data reduction techniques interact well with each of the models.

2. Agilent 4440 and Cyrano Sciences C320

The C320 (Cyrano Sciences, Inc., USA) unit is a self contained, hand-held electronic nose. It possesses 32 carbon black polymer sensors [15] which alter their electrical conductivity when they come into contact with certain odourous molecules in the air. The time dependent response is logged and from this the resistance change is computed. This feature is used to characterise each sensor's response to a particular sample: this is the input pattern in the model building stage.

The Agilent 4440 (Supplier: Gerstel, Berlin, Germany) comprises two units: an automatic headspace sampler that handles a hopper of 40 vials which are heated, individually, in an 'oven' for 4 min; vapour drawn from the atmosphere within the vial is passed to a quadrapole mass spectrometer. The mass spectrometer outputs a mass profile in the range 46–550 Da which is recorded by a desktop computer. Fig. 1 shows this device in a laboratory setting with the C320.

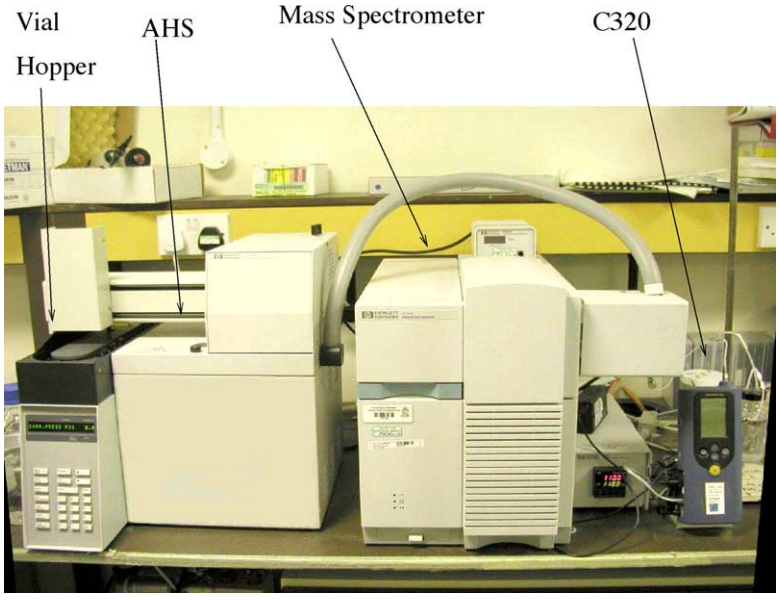

Fig. 1 An Agilent 4440 Chemosensor (PC not shown).

Importantly the AHS is a separate, autonomous system which synchronises the mass spectrometer's operation. This means that a C320 unit may be connected alongside and can analyse samples in parallel with the Agilent 4440. A schematic of this is shown in Fig. 2.

3. Dimension reduction

The Agilent 4440 measures a mass range between 46 and 550 Da, and the C320 has 32 different solid-state gas sensors which respond to various airborne molecules. This results in very high dimensional data sets and so the models required to analyse these data would have many free parameters to identify.

Supervised pattern recognition has two inherent stages and these are often referred to as training and testing. The training stage is the calibration part of the process which attempts to find empirical models to distinguish between data classes. Validation of the model is attempted during the test stage. Within the training stage there is an interaction between gathered data and a model of some form to be 'fitted' to the data set; this can be viewed as a calibration of the model. Hence the focus of this paper is not whether the model is, or the data are, good in their own right, but whether a model with sufficient 'discriminant power' may be estimated robustly from the data available; the model is tuned, or identified, using some form of optimisation algorithm. Optimisation of experimental design is not considered here.

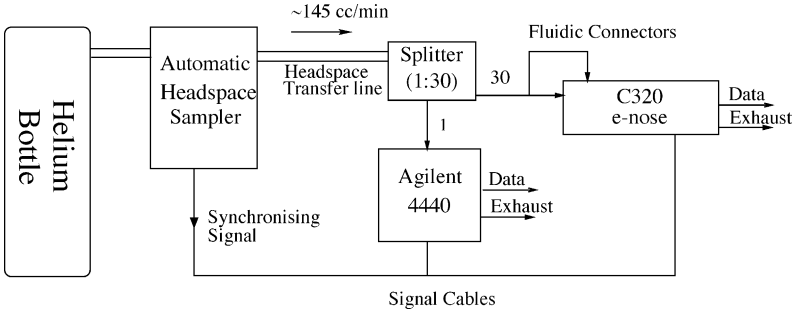

Fig. 2 Schematic of integration of the C320™ with the Agilent 4440. The AHS is fed a supply of helium, and this is used to flush the sampled headspace through the system. The 1:30 splitter passes the sampled vapour to the Agilent 4440 and the C320. The sample supply is connected to both the purge and sample inlets of the C320, this is because the helium supply from the AHS is used to 'wash' the sensors after each sample analysis. The sample vapour is vented to the atmosphere after it has been analysed.

It is this consideration that is of interest in the data analysis stage of this study. Parameter estimation needs to be a well-posed problem, that is there are sufficient data to identify the parameters. In this sense the problem is neither under- nor over-determined and there is a 'unique' set of parameters that fit the data. It is apparent then that the size of the data set dictates the complexity of the models that can possibly be identified, and vice versa. There is a data dimension issue here as this study deals with an electronic nose with an output of 32 channels and a mass spectrometer with an output of 505 channels. This means that, due to the number of degrees of freedom in the data, that a large amount of data would be required to identify a discriminant model. To satisfy this form of data requirement is difficult as a great deal of laboratory time is required; it takes about 180 min for the Agilent 4440 to process 40 samples at 4.5 min per sample.

It is therefore important to ascertain a method of reducing the dimension of the input to a pattern recognition algorithm in such a way that 'essential' information [16] is not lost. In order to do this a knowledge of the data set structure is necessary, or some a priori assumptions need to be made about this structure, i.e. there is some level of interdependency of the channels.

3.1. Current data set reduction methods

The problems of high dimensional data, and the need to reduce the dimension of a data set, have been investigated for some time. Much of the work has come from the 'intrinsic dimension' [17] point of view. This arises from the assumed existence of some co-dependence between, or constraints upon, the observed outputs of a system. This means the hypothesised probability distribution, and hence the sampled data, lie in some hypersurface within the system output space. A usual definition is that the intrinsic dimension of a data set is the number of free parameters in the minimal model that describes the set [18]. However, this model is not necessarily known a priori.

There is some confusion over this definition; for example in [18] this is taken to mean the dimension of the subspace of the output space that the data lie in. Considering the case of data sitting upon the surface of a sphere shows this to be ambiguous; the dimension of the surface, that is the number of parameters necessary to describe it, will be one less than the dimension of the vector space in which it sits, but the vector subspace in which it lies is the entire space. Thus for generality the data must be considered to lie in a manifold.

It has therefore been the approach to use tools that attempt to approximate this minimal model by making certain assumptions. In this way an attempt to identify a mapping that lowers the dimension of the data set, in line with an assumed form of the manifold, may be made. For example, principal components analysis (PCA) [19], the Sammon map [20] or some hierarchical cluster analyses [21] have been used. PCA looks for vectors along which the sampled data are varying the most. This commonly used algorithm outputs any specified number of these vectors, referred to as principal components, up to the number of dimensions, or the number of samples, whichever is lower. It is quite common to find that in electronic nose data the first three or four principal components describe something in the region of 95% of the variance in the data [17]. So if, for example, linear co-dependence between channels is assumed, these new coordinates should encompass the majority of the behaviour of the system being modelled.

The Sammon map is a metric preserving map. In fact the Sammon map algorithm searches for coordinates in a specified lower dimensional space that preserve interpoint distances. It can be seen that this technique will be most effective in reducing the dimension if the 'intrinsic dimension' of the data is much lower than the space in which they sit. For example, data sampled from a system with a four dimensional output may all lie in a three dimensional subspace; hence the Sammon map is designed to find three dimensional coordinates that retain the inherent metric information.

In this paper, variations of the Sammon map and a technique based upon correlation will be considered. Dimension reduction will be performed before applying four different black box models; two linear, the other two nonlinear. These black box models are applied as discriminant techniques. It is apparent in this context that dimension reduction will be beneficial as any interdependence will mean there is not a unique solution for the set of free parameters. PCA, though it is used as a data visualisation tool, is not considered for data reduction as it too requires a large amount of data for good estimation. Secondly it should be noted that data classes may not separate in the directions of largest variation, which is an assumption when using PCA.

Besides normalisation, two approaches are taken, although they both result in a linear projection onto a lower dimension space. However, the distinction is made in the following way. One approach is to analyse the information provided by

each individual channel, relative to the rest. Due to this leading to a selection of the most informative channels this is referred to as *sensor selection*. The second approach is a more general embodiment of the assumption of channel interdependence. A mapping which produces new coordinates, in a lower dimensional space, which is a function of the original channels is sought. This is referred to as *feature extraction*.

3.2. New set reduction methods

It is assumed that not all sensors supply data containing information relevant to the discriminant task. There may be noise from the environment such as temperature effects or there may also be information about other aspects of the sample, i.e. interference. Sensor drift is a key issue for chemical sensors and it is necessary to have a way of selecting some sensors, or functions of sensors, that reduce the dimension of the input of the discriminant map such that the data are sufficient to estimate the free parameters within the model. Here two different approaches are taken. The first is based upon the observation that any channel, whose output is dependent upon, or correlated with, that of another is redundant (to some extent). Indeed, if there is a mutual dependence between two channels then the uncertainty of the output of one will reduce when there is knowledge of the output of the other. Hence it will have lower information content than if it stood alone; the greater the dependence, the less information supplied. It can then be seen from this that there will not necessarily be a unique parameterisation for a discriminant model. Thus, removing a channel whose output is dependent upon others' will result in a better posed parameter estimation problem where the majority of the information content of the full output has been retained.

A linear cross-correlation technique is considered. It is an ad hoc approach based upon the observation that highly correlating sensor responses imply redundancy in the data set. The full correlation coefficient matrix is calculated for the data set containing the sensor responses/mass spectra. A corresponding matrix containing the sign of each of the correlations coefficient entries is produced. The sign is encoded as a '1' for a positive value, a '−1' for a negative value and a '0' for zero; the result is a matrix which is the same size as the correlation coefficient matrix, but has only entries from $\{-1, 0, 1\}$. The sums of the columns of this matrix are then calculated—giving a correlation score for each sensors. The sensors can then be ranked either in descending or ascending order, so that positive or negative correlation may be examined.

The second approach adopts the Sammon map. It is primarily used as a data visualisation tool, where the algorithm searches for coordinates that preserve the sample space interpoint distances in a much lower dimension; for this reason it is a non-linear mapping. However, here it is extended to estimate a linear projection to apply to a previously unseen data set. This projection may then be used as a feature selection mapping to process data prior to a black box model.

A projection is produced from the Sammon map by first applying the algorithm to a subset of the data. This results in a set of coordinates in a lower dimension which preserve, as much as possible, the interpoint distances of the original data. A linear mapping is then sought that maps the original data set a closely as possible to its Sammon map image. This is the projection that is required.

Formally, if D is defined to be the original data set and D_p to be the corresponding coordinates produced by the Sammon algorithm (the convention that samples are in rows and the channel outputs are in columns is used), then a projection P is sought such that

$$PD^{\top} = D_p^{\top} \tag{1}$$

This will be an under determined problem. However, by applying the Penrose–Moore pseudoinverse [22] of $D^{\top}$, $D^{\top +}$, the best possible approximation, P_0 is obtained thus:

$$P_0 = D_p^{\top} D^{\top +} \tag{2}$$

Thus a linear projection which respects the interpoint distances of the original data subset is obtained, and can be used as a form of feature selection. It should be noted that the image of the original data set under P_0 was very similar to the output of the Sammon map.

4. Black box models

A number of types of black box model for pattern recognition are considered in this study. A full review of each is not carried out here but the reader is referred to the references provided for more details. These models assume some relationship between the input and the output of the system. Here the system considered is the the inverse system; the input is the output measurements for a sample and the output is the classification of the sample. The classifications considered in this article are binary in nature; a value of -1 will denote a 'negative'

sample and 1 will denote a 'positive' sample. Negative will be taken to mean the absence of a urinary tract infection, or the absence of MRSA, positive will denote the converse.

Of the four models used in the study, two are linear in structure and are akin to a linear regression on the observed input–output behaviour. The remaining two are nonlinear and use Gaussian bell shaped curves to yield a more probabilistic structure.

4.1. Multilayer perceptron (MLP)

The architecture for the multilayer perceptrons (MLPs) [10] considered in this study has an input of the same dimension as the experimental data, 8 in the hidden layer and a single output node for binary classification; this results in a network with 272 free parameters for an input of 32 dimensions. This model generates 8 linear combinations of the inputs and then calculates a weighted sum of these. The output is the classification of the input. A similar topology was used for the blood data. This particular architecture was chosen as the resulting number of free parameters may be estimated with the data available; it is the most flexible MLP model that can be realistically identified.

4.2. Autoregressive exogenous model

A standard linear transfer function type black box model was also considered; specifically the autoregressive exogenous type (ARX) [23]. This type was used as it takes into account the past history of the sampling equipment. This analysis indicates whether the system being modelled is changing with time.

This important class of black box predictive models are linear systems of the general form

$$y(t) = \Phi^{\top}(t)\theta \tag{3}$$

for its input–output characteristics. Here

$$\Phi(t) = \begin{bmatrix} \phi_1(t) \\ \vdots \\ \phi_n(t) \end{bmatrix} \tag{4}$$

is the input vector, $y(t)$ the output and θ the parameter vector representing the parameterisation of the system.

In particular, ARX is of the form:

$$\mathrm{ARX}(q,r) : Y[n] = \sum_{i=1}^{q} \alpha_i Y[n-i] + \sum_{k=0}^{r} \gamma_k U[n-k] + e[n] \tag{5}$$

or in matrix notation

$$A(z)Y[n] = B(z)U[n] + e[n] \tag{6}$$

where $A(z)$ and $B(z)$ are polynomial matrices,

$$A(z) = 1 - \alpha_1 z^{-1} - \cdots - \alpha_q z^{-q} \tag{7}$$

$$B(z) = 1 + \gamma_1 z^{-1} + \cdots + \gamma_r z^{-r} \tag{8}$$

the orders of which are defined by q and r in Eq. (5) above and the dimensions are dependent on the number of inputs and outputs; e is the error. Here the autoregressive part is defined by the matrix **A**, and the exogenous (input) part is defined by matrix **B**. These are polynomials in z^{-1} which represents a shift one time step in time; z^{-2} is two steps back and so on. Thus this model represents a weighted sum of the inputs to the system up to the current one, and all previous outputs. The output of the model is the predicted current output.

The model assumes a number of things. Firstly that the system is linear in behaviour. Secondly, that the system is deterministic and is only determined by a short history of input–output responses; the system is finite and causal. Finally that the noise is zero biased and white in nature. This last assumption is necessary for least squares fitting to be an unbiased estimate of the maximum likelihood error.

4.3. Radial basis function neural network

Radial basis functions [10] (RBFs) are spherically symmetrical functions. The type considered here are Gaussian because of their nonlinearity; because importance is stressed upon the distance between points in the data set and the fact that they have been shown to be universal approximators [24]. These are of the form:

$$G_{\sigma}(\mathbf{x}, \mathbf{c}) = \exp -\frac{\|\mathbf{x} - \mathbf{c}\|}{2\sigma^2} \tag{9}$$

Here **x** is the model input. The **c**'s are *centres* (means) and σ is the width parameter for this bell shaped curve. These functions form the hidden layer of a nonlinear type of MLP. The output node is a weighted sum. The weights are identified by use of the support vector machine (SVM) algorithm [25]. This nonlinear optimisation algorithm selects centres from amongst the training points. The optimisation performed minimises complexity as well as the error to yield a well identified model.

The resulting model may be written in the form

$$F(\mathbf{x}) = \omega_i \sum_i G_\sigma(\mathbf{x}, \mathbf{c}_i) + b \tag{10}$$

where the weights, ω_i, and the bias, b, are estimated using the experimental data. $F(\mathbf{X})$ is the classification of the input $\mathbf{x}$.

4.4. Nonlinear time series model

A hybrid, nonlinear, time series analysis model is considered as well. It was suggested in [26] that RBFs might prove useful for time series analysis. The theory of nonlinear black box models may also be found in articles such as [27] and [28]. Here the acronym NARMAX was coined for use with a generic nonlinear form of ARMAX models. Here a novel way of combining the concept of the ARX type model with RBFs is demonstrated which results in a model whose parameters may be identified using the support vector method: the NARX. The advantage being, as will be demonstrated later, that a model is generated that is endowed with robust parameter estimation.

If the form of a standard ARX model is considered then, in the many-in-single-out (MISO) case (Eq. (5)), it may be expanded into the form

$$\begin{aligned}\alpha_1 y(n) + \alpha_2 y(n-1) + \cdots + \alpha_{na+1} y(n-na) \\ = \beta_1 \cdot u(n) + \beta_2 \cdot u(n-1) + \cdots \\ + \beta_{nb} \cdot u(n-nb) + e(n)\end{aligned} \tag{11}$$

Here y represents the output of the system (in this case the classification of the sample) and u is the input (the sensor responses) and e is the error; n indexes the nth element of the input and output sequences. The α's and β's are coefficients of polynomials in the shift operator and na and nb are the orders of these polynomials. The input to the model may be represented as a vector of the form

$$u(\cdot) = \begin{bmatrix} u_1(\cdot) \\ u_2(\cdot) \\ \vdots \\ u_d(\cdot) \end{bmatrix} \tag{12}$$

By rearranging, the following is obtained

$$\begin{aligned}\alpha_1 y(n) = \beta_1 \cdot u(n) + \beta_2 \cdot u(n-1) + \cdots \\ + \beta_{nb} \cdot u(n-nb) - \alpha_2 y(n-1) - \cdots \\ - \alpha_{na+1} y(n-na) + e(n)\end{aligned} \tag{13}$$

or

$$\begin{aligned}\alpha_1 y(n) = [\beta_1, \ldots, \beta_{nb}, -\alpha_2, \ldots, -\alpha_{na+1}] \\ \times \begin{bmatrix} u(n) \\ u(n-1) \\ \vdots \\ u(n-nb) \\ y(n-1) \\ \vdots \\ y(n-na) \end{bmatrix} + e(t)\end{aligned} \tag{14}$$

This allows us to place the model in the context of a support vector machine using Gaussian kernels. The dot product in Eq. (14) above can be replaced with kernel products of the form

$$\langle \mathbf{x}, \mathbf{y} \rangle = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2} \right) \tag{15}$$

where $\mathbf{x}$ and $\mathbf{y}$ are two vectors of the same dimension and σ the width parameter as before.

Hence the result is an RBF network whose input is the column vector on the right-hand side of (14), and the target output is $y(t)$. The parameters α_i and β_i are found using the SVM algorithm. All model identification and simulation was performed using MATLAB (Mathworks) Version 5 on a Sun Microsystems Workstation with the system identification and optimisation toolboxes.

5. Experimental method

The following sample preparation was carried out in the Department of Biological Sciences at the University of Warwick, UK.

5.1. Preparation of frozen samples

In order to obtain a culture of a known value of colony forming units (cfu) per milliliter, frozen aliquots of the various species considered were created. For each strain three cultures were prepared using 10 ml of brain heart infusion (BHI) and one colony inoculated into each. These inoculated bottles were incubated overnight at 37 °C and shaken at 75 rpm. Following incubation one bottle containing each culture was removed, vortexed and pipetted in 1 ml aliquots into 1.5 ml eppendorfs. The aliquots were then centrifuged into a pellet (13,000 rpm, 2 min) and the supernatant removed. The second culture bottle was then aliquoted in a similar manner into the eppendorf containing the

pellet and the above steps repeated. The third bottle was processed similarly. Finally, 1 ml BHI + 15% glycerol was added to the eppendorf containing the pellet and vortexed to create a suspension which was quick frozen to −80 °C.

5.2. Inoculation of blood culture bottles

All experiments were conducted in BacT/Alert SA (aerobic) (Biomerieux UK Ltd.) sterile culture bottles. The bottles contained 40 ml media plus an internal sensor that detects carbon dioxide dissolved in the culture medium. The medium formulation consists of pancreatic digest of casein (1.7%, w/v), papaic digest of soybean meal (0.3%, w/v), sodium polyaneatholesulfonate (0.035%, w/v), polyiodoxine HCl (0.001%, w/v) and other complex amino acid and carbohydrate substrates in purified water.

Microorganism presence and the consequent production of carbon dioxide results in a colour change in the gas permeable sensor at the bottom of the tube.

The culture bottles were injected with 10 ml of blood immediately prior to use to recreate a clinical situation of adding a patient blood sample. A sample of 10 ml of blood is recommended, although lower blood volumes can be used, but recovery may not be as great. During the investigation experiments were conducted using defibrinated horse blood (no preservative) (Oxoid Ltd.).

The bottles were inoculated using the above prepared microorganism pellets and incubated at 37 °C and shaken at 75 rpm to recreate a standard incubator used in a hospital.

5.3. Preparation of samples for the automatic headspace sampler

The samples to be measured by the Agilent 4440 were transferred as 1.5 ml aliquots in 10 ml sterile flat bottom headspace vials (Agilent Technologies, Inc.). Blanks of BacT/Alert SA and blood were run alongside inoculated samples in addition to BHI broth as standards to confirm the accuracy of the equipment. It should be noted that this broth is extremely odourous with the headspace being very variable.

5.4. Urine preparation

The samples arrived from the Walsgrave hospital in standard urine sample vials. Each sample had a boric acid (62 Da) additive to suspend biological activity. The exact concentration of the additive in solution is unknown because it is placed in the vial by the manufacturer and is also dependent upon the volume of the sample provided by the patient. This will introduce a significant variability into the measurement procedure.

Samples were again transferred as 1.5 ml aliquots to the standard 10 ml vials used by the Agilent 4440. In addition to this other microbiological techniques were used to 'classify' each sample with respect to evidence of bacterial infection. Infection levels were measured via:

- *Classical microbiology*. That is, inoculation of growth media with prepared samples to demonstrate the presence bacteria. This was also to identify species using staining growth media.
- *Cellfacts*. This instrument (Cellfacts Ltd.) measures the distribution of particle sizes in a fluid sample. The observed diameter indicates the level of bacterial presence, white and red blood cells. The results of this were used to classify the urine samples into being either 'positive' or 'negative' with regards to being indicative of a UTI.

Fig. 3 shows a typical sensor response for six sensors when the C320 is challenged with the headspace of a urine sample. The response is a output from Cyrano Sciences own data logging software and is proportional to the resistance of the sensor. Each sensor response is coded onto as real numbers using the fractional change measure:

$$\text{Measure (Response)} = \frac{\text{Maximum} - \text{Baseline}}{\text{Baseline}} \quad (16)$$

and this is used as the input to the black box models. Thus the time series of each experimental response is not considered in the models, only the time series of exposures as measured by the fractional change.

6. Results

The results for urine and blood are considered separately. In both cases, model identification and cross validations are achieved using the 'leave-one-out' [10] algorithm. Cross validation is when the discriminant model is tested using previously unseen data. In the case of the leave-one-out algorithm, a calibration set comprising of all but one of the data is used and the remaining datum is used to test the model. This is carried out for each data point in turn and the accuracy is the average of all these tests. Accuracy is defined as

$$\text{Accuracy} = \frac{\text{Number of samples correctly identified}}{\text{Total number of test samples}} \quad (17)$$

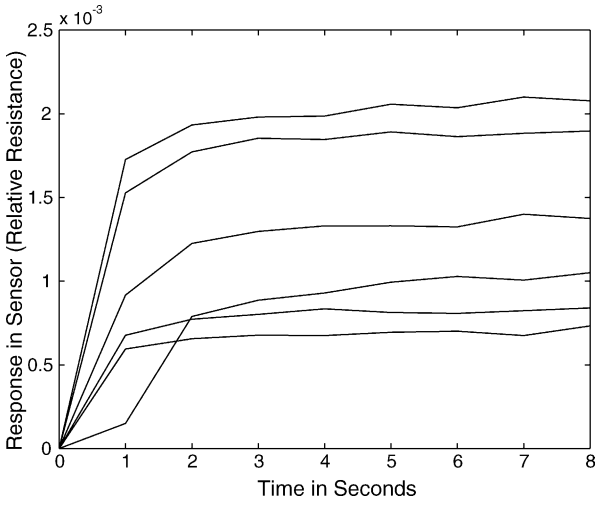

Fig. 3 Typical time series responses for six of the C320 sensors when exposed to a urine sample. The measured response is proportional to the resistance of each sensor, and was provided by the manufacturer's software.

In both experiments the samples were screened using the Agilent 4440 and the C320 unit. However it was found that the mass spectrometer performed poorly for the urine screening experiment, and vice versa for the blood screening experiments. No explanation for this observation is proposed at this time.

6.1. Urine

For these data the aim was to classify accurately whether each sample was 'negative' or 'positive'. This is a difficult problem due to the 'chemical noise' of the samples and the range of different indicators used to positively identify a UTI. As discussed above, the attempt is to identify the part of the signal from the system which relates to the presence of bacteria.

The urine screening experiment resulted in a set of 189 sensor responses from the C320. By using data from all of the 32 sensors an accuracy of 65% was obtained using an ARX model. The order of the model was optimised to $na = 3$, $nb = 4$ to gain the greatest success rate; in all the models below the order was optimised by testing a range of values. The number of sensors was reduced to 19 by using the most negatively correlating sensors, giving 67%. Selecting 19 out of 32 sensors is still a high proportion of the sensors, but this still yielded a more robust model. Negative correlation was chosen as it resulted in a much better success rate when compared to positive correlation, and the number of sensors chosen was optimal with respect to the success rate of the model. The plots in Fig. 4(a) and (b) show the relationship between order and accuracy. The former is for the full array of 32 sensors and the latter is for the reduced 19 sensor set. Note that once sensor reduction has been performed then lower order models are preferred. This suggests that the sensors that have been discarded contained no information with regard to the discriminant task and so only provided some false 'trends' to which the model attempted to fit. Normalisation gave 60% but when coupled with the most negatively correlating sensor technique 71% was achieved.

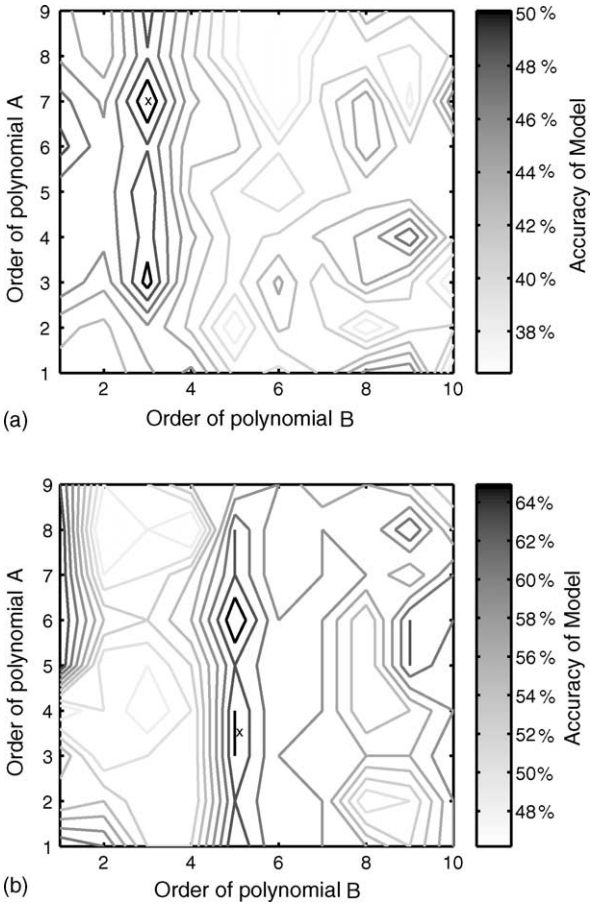

Fig. 4 Model order against accuracy plots for the ARX type model: (a) the 32 sensor data set and (b) the reduced 19 sensor data set.

RBF success on these data was approximately 50%. Here the width parameter was optimised by applying the SVM algorithm to a range of values for this parameter. Applying the correlation results to RBF networks it was found that a maximum accuracy of 65% was possible.

The hybrid NARX type model was considered and, for $na = 3$, $nb = 4$, 80% was the maximum accuracy achieved and the best overall for the urine data.

Table 1 Key to data plots Fig. 5(a) and (b)

Strain	Symbol
MRSA	◯
NCTC	+
S. Epidermis	*
S. Warneri	◇
S. Simularis	□
S. Haemalyticus	⋆
S. Lugdenensis	✡
Growth medium	.

Reducing the number of sensors lowered the accuracy to 73%. This was perhaps because the hybrid model relies on the distance between points, so extra dimensions are an advantage. Further, it should be noted that radial basis functions have a regularising, smoothing property which allows them to avoid over-fitting.

6.2. Blood

The blood data were recorded using the mass spectrometer. A useful data preparation technique was performed by removing the first two measurements of each data file, due to instability in the measurement of the first two samples.

Applying an ARX model to the full data set was not suitable due to this being a very poorly determined problem; each input to the inverse model has 505 dimensions and so the number of free parameters in even a low order model would be an order of magnitude higher than the amount of data available. By using the top 200 positively correlating masses a cross-validated accuracy of 94% was achieved. The result of the correlation technique may be observed by comparing Fig. 5(a) and (b); note that separation has not been affected but the data may have much simpler models fitted to them. The PCA analysis here is just as a means of visualisation, not data analysis.

Applying the hybrid model gave a successful classification rate of 90%. However the best results were achieved by using the Sammon map and training a standard RBF network. This resulted in an excellent accuracy of 100% successful classifications.

The results are summarised in Table 2 and are compared with a multilayer perceptron (dimension of input $\times$ 8 $\times$ 1). The results demonstrate, in certain cases, combining certain data reduction techniques with black box models yields an unsatisfactory classification rate. However these results are included in order to demonstrate the effect that each technique has on the data sets.

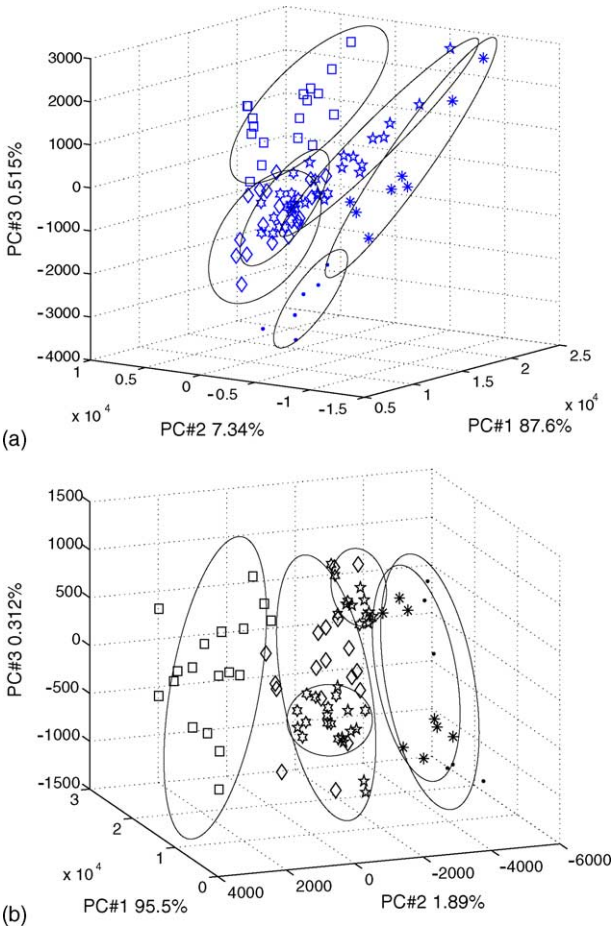

Fig. 5 PCA plots of blood data: the ellipses illustrate the spread and overlap of the classes of data: (a) the full mass range and (b) a reduced data set produced using the correlation method. Refer to Table 1 for key.

7. Discussion and conclusion

Two different types of data sets have been analysed using dimension reduction techniques and black box models. A marked increase (in the region of 20%) has been observed in the success of the models after judicious pruning of the dimension of the data set analysed.

The high order of ARX models necessary for the urine data suggest that the history of experiments needs to be taken into account; this is evidence of a time dependent system. Although the sensors respond over a 10-s period, the ARX model results suggest that there is a memory effect lasting for a considerably longer time. This could mean anything from sensor drift to environmental factors affecting the sampling equipment, and also the samples themselves changing over time. Why this is so for the urine data and not the blood could be due to the different sampling equipment used

Table 2 Summary of results as percentage accuracy

Model	Optimisation	Urine	Blood
ARX	Polynomial order	60	Not feasible
−Corr ARX	No. of remaining sensors	67	50
+Corr ARX	No. of remaining sensors	45	94
RBF	Width parameter	65	71
Sammon RBF	No. of features to extract	50	100
Norm RBF	Width parameter	50	75
Nonlinear model	Kernel width parameter	80	72
MLP	Fixed number of nodes	38	50
MLP +Corr	No. of remaining sensors	40	72
MLP Sammon	No. of extracted features	10	65
MLP Norm	Fixed number of nodes	63	50

Note that an ARX model cannot be applied to blood data as there are too few data points to estimate the required number of weights.

for the two sets of measurements. The urine sample headspace was analysed using a C320 which uses chemical sensors which react with volatile molecules; the urine is acidic (and also contains ammonia) and so may be 'poisoning' the sensors. The blood data were produced using an Agilent 4440, which measures the distribution of molecular masses using a quadrapole mass spectrometer. Thus it maybe that C320's response is dependent upon those samples that it has interacted with in previous experiments.

It appears that the correlation techniques work most successfully with linear models, such as ARX, whereas the Sammon map works best with the nonlinear models. This may be due to the nonlinear models using radial basis neural networks. These functions are spherically symmetric, hence distance information is important. With a linear model, using linear combinations of sensor outputs, correlation between sensors implies redundancy. It is unclear, however, why negative correlation is most effective in one case and positive correlation is effective in the other case. This may again be due to the sampling equipment used. The blood samples were analysed using a mass spectrometer and thus the counts of each molecular weight are always positive. Hence any intensity information that is manifested as correlation will tend to be positive. In the case of the urine samples, the device used was the C320 electronic nose. The sensor response may exhibit a positive or negative resistance change. It may be that components of the urine with no bearing on the presence or not of an infection correlate negatively with compounds denoting infection; this has yet to be verified.

It is possible that the Sammon map works best as it is biased to conserve the distance between mutually remote points rather than neighbouring points, which would be hypothesised to belong to the same data class. Correlation is effective as it 'prunes away' redundant, correlating sensors.

The results for the various models suggest that separation between classes is inherently nonlinear; RBF-based techniques faired much better than the linear MLP. In addition the SVM method seems to effective at avoiding over-fitting.

These techniques still beg some questions. The first is that of choosing the optimal reduced dimension of the input for the black box model. As discussed above, lower dimensions result in a much more robust estimates of the parameters. However, Cover's theorem [10] dictates that a high dimensional feature space is desirable[1] for maximal discriminant power due to the greater degree of model freedom this affords. Hence an optimal dimension is a trade-off between experimental ability and the complexity necessary of a model to explain the input–output behaviour of the physical system 'sufficiently well'.

Using the urine data, the least squares error of the Sammon map algorithm may be used to estimate how well the data fit into a given number of dimensions. In this way it is possible to estimate the 'intrinsic dimension' of the data by observing how the topological structure is preserved. We can compare this with how much variance is described by each principal component resulting from PCA. Fig. 6 shows this comparison for the urine data; the corresponding analysis on the blood data are shown in Fig. 7. The eigenvalues of each principal component are used to denote the variance. It may be

[1] The basis of the proof of this is the greater degree of freedom afforded models on such spaces.

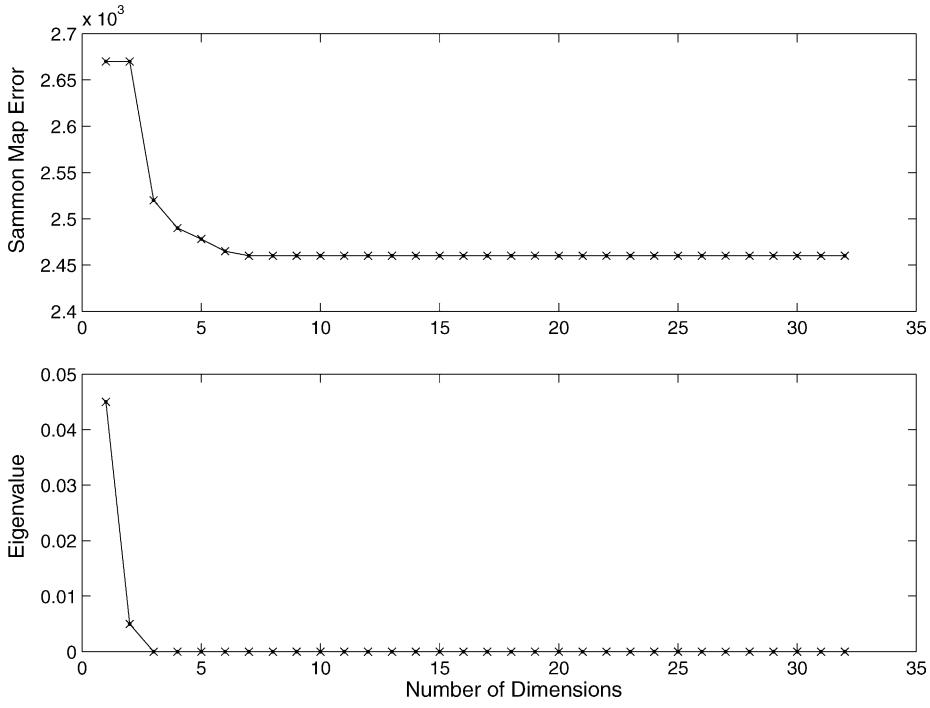

Fig. 6 Two graphs demonstrating different ways of measuring the intrinsic dimension of the urine data.

seen that both suggest that the data may be reduced to three or four dimensional sets. This could be a method of optimising the input dimension of a model computationally.

It should be noted that the definition of intrinsic dimension given above requires that the form of co-dependence of channels is known. At the very least some assumptions about the form of these

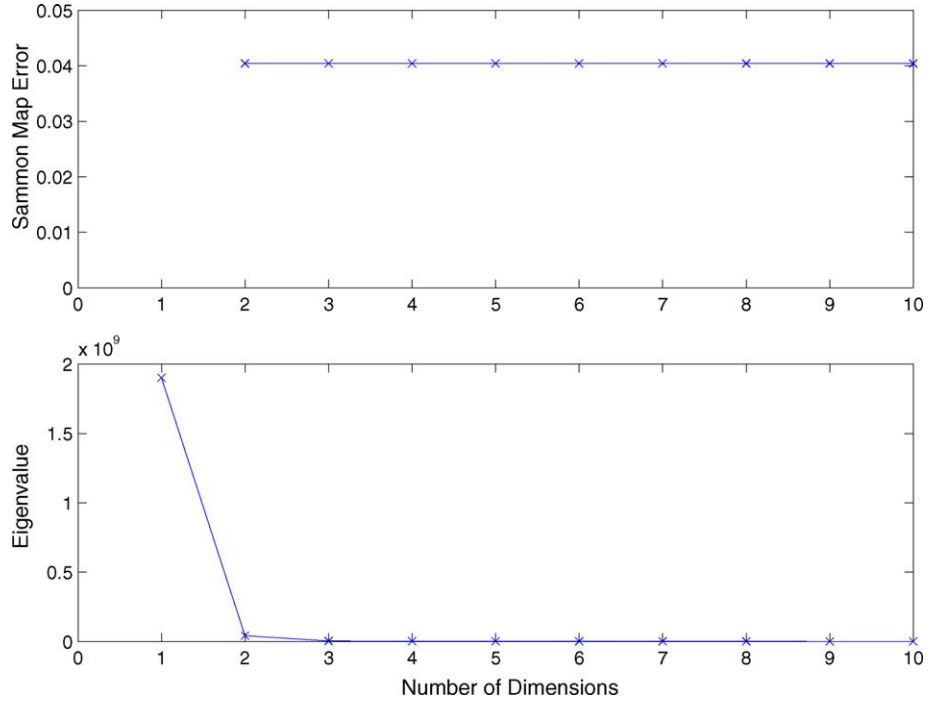

Fig. 7 Two graphs demonstrating different ways of measuring the intrinsic dimension of the blood data.

relationships have to be made in order to estimate the intrinsic dimension. It seems then that this is another aspect of black box modelling, and requires the same in-depth analysis and investigation that has been afforded to input–output models.

This work suggests that data set dimensional reduction can improve the robustness of a model and that the specific method used is dependent upon both the data type and the black box model being considered. When applied to electronic nose data the improvements in success rates are very encouraging, and this may help to improve the technology, via computational methods, in the area of rapid medical diagnosis.

References

[1] R. Esteves de Matos, D.J. Mason, C.S. Dow, J.W. Gardner, Investigation of the growth characteristics of *E. coli* using headspace analysis, in: J.W. Gardner, K.C. Persaud (Eds.), Olfaction & Electronic Noses 2000, IOP Publishing, Bristol, UK, 2000.

[2] J.W. Gardner, H.W. Shin, E.L. Hines, An electronic nose system to diagnose illness, Sens. Actuators 70 (2000) 19—24.

[3] F. Pena, S. Cardenas, M. Gallego, M. Valcarcel, Characterisation of olive oil classes using a chemsensor and pattern recognition techniques, J. Am. Oil Chem. Soc. 79 (11) (2002) 1103—1108.

[4] M.B. Wise, M.R. Guerin, Direct sampling MS for environmental screening, Anal. Chem. 69 (1) (1997) A26—A32.

[5] S.Y. Lai, O.F. Deffenderfer, W. Hanson, M.P. Phillips, E.R. Thaler, Identification of upper respiratory bacterial pathogens with the electronic nose, Laryngoscope 112 (6) (2002) 975—979.

[6] T. Chen, M. Rimpiläinen, R. Luukkainen, T. Möttönen, T. Yli-Jama, J. Jalava, O. Vainio, P. Toivanen, Bacterial components in the synovial tissue of patients with advanced rheumatoid arthritis or osteoarthritis: analysis with gas chromatography—mass spectrometry and pan-bacterial polymerase chain reaction, Arthrit. Rheumat. (Arthrit. Care Res.) 49 (2003) 328—334.

[7] A.K. Pavlou, N. Magan, D. Sharp, J. Brown, H. Barr, A.P.F. Turner, An intelligent rapid odour recognition model in discrimination of helicobacter pylori and other gastroesophageal isolates in vitro, Biosensors Bioelectron. 15 (2000) 333—342.

[8] R. Canton, T.M. Coque, F. Baquero, Multi-resistant Gram-negative bacilli: from epidemics to endemics, Curr. Opin. Infect. Dis. 16 (4) (2003) 315—325.

[9] A.E. Aiello, E. Larson, Antibacterial cleaning and hygiene products as an emerging risk factor for antibiotic resistance in the community, Lancet Infect. Dis. 3 (8) (2003) 501—506.

[10] S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice-Hall, New Jersey, 1999.

[11] P.D. Fey, B. Säid-Salim, M.E. Rupp, S.H. Hinrichs, D.J. Boxrud, C.C. Davis, B.N. Kreiswirth, P.M. Schlievert, Comparative analysis analysis of community- or hospital acquired methicillin-resistant *Staphylococcus aureus*, Antimicrob. Agents Chemother. 47 (1) (2003) 196.

[12] D.M. Wilson, K. Dunman, T. Roppel, R. Kalim, Rank extraction in tin-oxide sensor arrays, Sens. Actuators B 62 (2000) 199—210.

[13] A.N. Krutchinsky, B.T. Chait, On the nature of chemical noise in MALDI mass spectra, J. Am. Soc. Mass Spectrom. 13 (2) (2002) 129—134.

[14] A. Polettini, A. Groppi, C. Vignali, M. Mantagna, Fully-automated systematic toxicological analysis of drugs, poisons, and metabolites in whole blood, urine, and plasma by gas chromatography full scan mass spectrometry, J. Chromatogr. 713 (1) (1998) 265—279.

[15] M.C. Lonergan, E.J. Severin, B.J. Doleman, S.A. Beaber, R.H. Grubbs, N.S. Lewis, Array-based vapor sensing using chemically sensitive, carbon black-polymer resistors, Chem. Mater. 8 (1996) 2298—2312.

[16] A. Potapov, M.K. Ali, Neural networks for estimating intrinsic dimension, Phys. Rev. E 65 (4) (2002) (Article Number 046212).

[17] Michael Hörnquist, John Hertz, Mattias Wahde, Effective dimensionality of large-scale expression data using principal components analysis, Biosystems 65 (2002) 147—156.

[18] F. Camastra, A. Vinciarelli, Estimating the intrinsic dimension of data with a fractal-based method, IEEE Trans. Pattern Anal. Mach. Intell. 24 (10) (2002) 1404—1407.

[19] Ker-Chau Li, Kerby Shedden, Identification of shared components in large ensembles of time series using dimension reduction, J. Am. Stat. Assoc. 97 (2002) 759—765.

[20] W. Dzwinel, How to make Sammon mapping useful for multidimensional data structure analysis, Pattern Recogn. 27 (7) (1994) 949—959.

[21] C.A. Nicolaou, S.Y. Tamura, B.P. Kelley, S.I. Bassett, R.F. Nutt, Analysis of large screening data sets adaptively grown phylogenetic-like trees, J. Chem. Inform. Comput. Sci. 42 (2002) 1069—1079.

[22] A. Albert, Regression and the Moore—Penrose Pseudoinverse, Academic Press, London, 1972.

[23] L. Ljung, System Identification: Theory for the User, Series: Information and System Sciences Series, Prentice-Hall, New Jersey, USA, 1987.

[24] M. Powell, Radial basis functions for multivariable interpolation, in: Proceedings of the Conference on Algorithms for the Approximation of Functions and Data, 1985, pp. 143—167.

[25] V. Vapnik, Statistical Learning Theory, Wiley/InterScience, Danvers, MA, USA, 1998.

[26] M. Casdagli, Nonlinear prediction of chaotic time series, Physica D 35 (1989) 335—356.

[27] I.J. Leontaritis, S.A. Billings, Input—output parametric models for non-linear systems. Part I. Deterministic non-linear systems, Int. J. Contr. 41 (2) 1985 303—328.

[28] S. Chen, S.A. Billings, Representations of non-linear systems: the NARMAX model, Int. J. Contr. 49 (3) (1989) 1013—1032.