1999

A comparison of techniques for calibration of non-linear electro-analytical data

Joel C. Swanson
San Jose State University

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_theses

Recommended Citation

Swanson, Joel C., "A comparison of techniques for calibration of non-linear electro-analytical data" (1999). *Master's Theses*. 1847.
DOI: <https://doi.org/10.31979/etd.j7fa-9dh2>
https://scholarworks.sjsu.edu/etd_theses/1847

This Thesis is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Theses by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

A COMPARISON OF TECHNIQUES FOR
CALIBRATION OF NON-LINEAR ELECTRO-
ANALYTICAL DATA

A Thesis

Presented to

the Faculty of the Department of Chemistry

San Jose State University

In Partial Fulfillment of the Requirements for the

Degree of

Master of Science

by

Joel C. Swanson

May, 1999

UMI Number: 1394559

Copyright 1999 by
Swanson, Joel Clyde

All rights reserved.

UMI Microform 1394559
Copyright 1999, by UMI Company. All rights reserved.

This microform edition is protected against unauthorized
copying under Title 17, United States Code.

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

© 1999

Joel Swanson

ALL RIGHTS RESERVED

APPROVED FOR THE DEPARTMENT OF CHEMISTRY

Dr. Sam P. Perone

Dr. Joseph J. Pesek

Dr. Herbert Silber

APPROVED FOR THE UNIVERSITY

ABSTRACT

A COMPARISON OF TECHNIQUES FOR CALIBRATION OF NON-LINEAR ELECTRO- ANALYTICAL DATA

by Joel Swanson

Some analytical methods suffer from matrix effects or sensor variance, which cause non-linear analyte responses. Linear Sweep Voltammetry (LSV) is such a method.

The research presented is a comparison of three linear and one non-linear techniques for constructing a calibration function for LSV data. The techniques are Multiple Linear Regression (MLR), Principal Component Regression (PCR), Partial Least Squares (PLS) Regression, and Artificial Neural Networks (ANN). These four techniques were applied to three different sets of calibration data for two-component mixtures with increasingly severe non-linear analyte response factors. The four calibration techniques were compared regarding accuracy, precision, reliability, and ease of application.

ACKNOWLEDGMENTS

I would like to thank my thesis advisor, Dr. Sam Perone, for his guidance and patience.

Thanks also to the members of my committee, Dr. Joseph J. Pesek, and Dr. Herbert Silber.

My particular appreciation and gratitude is given to my family.

TABLE OF CONTENTS

1	Introduction.....	1
1.1	Previous work.....	1
1.2	Objective.....	2
2	Theory.....	2
2.1	Non-Linear Data.....	2
2.2	Linear Sweep Voltammetry.....	3
2.3	Multivariate Data Analysis Techniques	5
2.3.1	Multiple Linear Regression.....	6
2.3.2	Principal Component Regression	7
2.3.3	Partial Least Squares Regression.....	9
2.3.4	Artificial Neural Networks	10
3	Experimental Approach.....	14
3.1	Instrumentation.....	14
3.2	Digital Simulation.....	14
3.3	Generation of the Sample Space	14
3.4	Selection of Cross-Validation Sets	17
3.5	MATLAB™ Analysis	21
3.5.1	Data Pretreatment	21
3.5.2	Implementation of Data Analysis Techniques	23
3.5.3	Figures of Merit.....	25
3.5.4	Optimization Using Predictor Set Error.....	26
4	Results	28
4.1	Comparison of Cross-Validated Prediction Results.....	31
4.2	Bias of Predicted Concentrations.....	34
4.3	Slope of Regression Line for Predicted Concentrations	38
4.4	Correlation of Bias and Slope.....	39
4.5	Comparison of Sets.....	41

4.6 Best Analysis Methods	45
4.6.1 Artificial Neural Network Comparison	45
4.7 Observations.....	46
5 Conclusions	47
6 Future Work.....	48
7 References.....	49

LIST OF FIGURES

Number.....	Page
Figure 1. Voltage and Current during Linear Sweep Voltammetry	3
Figure 2. Block Diagram of a Feed-Forward, Non-Linear Neural Network.....	12
Figure 3. Comparison of Variability of Voltammogram Shapes.....	17
Figure 4. Variation in Shapes of Voltammograms in the Leave-One-Out Groups in Data Set 2: [A] = [C] = 5 mM.....	19
Figure 5. Changes in a Data Set during Mean-Centering and Normalization	21
Figure 6. The Sigmoid Logistic Function	22
Figure 7. Use of Sum of Squared Errors to Optimize Data Analysis.....	28
Figure 8. Output from Four Data Analysis Techniques for Data Set 2, Subset L1.....	29
Figure 9. Output from Four Data Analysis Techniques for Data Set 1, Subset L9.....	30
Figure 10. PRMSE Values for Cross Validation from Data Set 1	31
Figure 11. PRMSE Values for Cross Validation from Data Set 2	33
Figure 12. PRMSE Values for Cross-Validation from Data Set 3	34
Figure 13. Biases for All Points in Data Set 1, Plotted versus Concentration	35
Figure 14. Biases for All Points in Data Set 2, Plotted versus Concentration	36
Figure 15. Biases for All Points in Data Set 3, Plotted versus Concentration	37
Figure 16. Prediction Biases for Cross-Validation Subsets in Data Set 1.....	38
Figure 17. Comparison of Regression Lines.....	39

LIST OF TABLES

Number.....	Page
Table 1. Designators and Corresponding Kinetics Values for Cross-Validation Subsets	20
Table 2. Correspondence between Subset Designators and Figure 4.	20
Table 3. Figures of Merit for Cross-Validated Predictions	42
Table 4. Best Techniques for Data Analysis, Based on Figures of Merit.....	45
Table 5. Comparison of ANN Analyses, Based on Figures of Merit.....	46

1 Introduction

1.1 Previous work

Linear sweep voltammetry (LSV) is a technique that has been used for chemical analysis, as well as to study oxidation-reduction reactions, electrode processes, and other electrochemical phenomena. It has been shown that voltammetric data may be used to classify compounds (1,2). The information content of voltammograms has been studied (3). The magnitude of the current measured, in response to a change in voltage at an electrode, may be used to determine the concentration of an electroactive compound. This is because of the linear relationship, at a particular voltage, between current and concentration. This linear relationship between current and concentration occurs when all other parameters affecting the current are constant. These parameters include diffusion rates of oxidized and reduced species, temperature of the system, rate of reaction for the exchange of electrons, rate of change of the voltage on the electrode, relative ease of oxidation reaction as compared to the reduction reaction, effective area of the electrode, and so forth.

In an application of linear sweep voltammetry in which electrode response factors change, the dependence of current on voltage will be non-linear with respect to electroactive compound concentration at a given voltage. This study is a comparison of data analysis techniques, applied to non-linear voltammetric data, and used to determine analyte concentrations. Other comparisons of data analysis techniques have been published recently (4, 5), which compared data analysis techniques for complex linear spectroscopic data. Another recent study (6) dealt with voltammetric data that are non-linear in character, due to intermetallic compound formation.

1.2 Objective

The purpose of this study was to compare different techniques for data analysis, as applied to a non-linear voltammetric system. The comparison was intended to identify the best technique for use in quantification of multiple component mixtures of electroactive compounds, as analyzed by linear sweep voltammetry (LSV) under non-linear conditions.

2 Theory

2.1 Non-Linear Data

Most analytical techniques rely on linearity of instrumental responses. This means that the system response is proportional to concentrations. This concept is represented in the following formula:

$$d_i = \sum_{j=1}^n R_{ij} C_j \quad (1)$$

Where d_i is the response (e.g., absorbance) in the i th channel, C_j is the concentration of the j th analyte, and R_{ij} is the proportionality constant (for example, the absorptivity) for the j th compound in the i th channel. An example of a data channel would be a particular wavelength in a spectrum, or a particular voltage in a voltammogram.

In a non-linear system, the above formula does not apply. One possible cause for this is that the proportionality constant, R_{ij} , is a function of variables which are not constant from one sample to another. Another possible cause could be interaction between analytes. In most analytical systems, the experiment is designed so that variables determining the value of R_{ij} are

held constant. Interactions between analytes are often addressed by separation methods or empirical interference correction factors.

In an analytical laboratory situation, the parameters affecting proportionality between concentration and instrument response may be held constant, and interferences quantified and corrected. On the other hand, for an unattended field experiment involving chemical sensors for analytes in environmental or wastewaters, response factors ($R_{i,j}$) could vary, requiring more robust data analysis methods to cope with non-linear performance.

2.2 Linear Sweep Voltammetry

In linear sweep voltammetry, the voltage of the working electrode is varied linearly with respect to time. If there is an electroactive species in the solution in which the electrode is immersed, and its oxidation-reduction potential is within the voltage range swept, a current will be developed. This is shown schematically in Figure 1.

Figure 1. Voltage and Current during Linear Sweep Voltammetry

The current is a function of the voltage and other parameters. The contributions of many of the parameters of interest to this study are shown by the Butler-Volmer formulation for the current at the electrode (7):

$$i = nFAk^0 \left[C_O(0,t) e^{-\alpha nF/RT(E-E^0)} - C_R(0,t) e^{(1-\alpha)nF/RT(E-E^0)} \right] \quad (2)$$

Where i , the current at the electrode, is the combination of anodic and cathodic currents. Other factors are: n , the number of electrons exchanged during oxidation or reduction; F , the Faraday constant; A , the area of the electrode; R , the gas constant; T , the temperature in degrees Kelvin; E , the voltage on the electrode; and E^0 , the formal potential for the oxidation-reduction couple. The concentrations, $C_O(0,t)$ and $C_R(0,t)$, are the concentrations of the oxidized and reduced species, respectively, at the surface of the electrode (distance = 0), which will also vary with time, t . The standard heterogeneous rate constant, k^0 , is the rate constant for the electrode reactions at the standard potential, and typically depends on the electrode material, and solution conditions. The transfer coefficient, α , has a value typically around 0.5, and can be thought of as a measure of how energy is partitioned as electrode potential changes. It describes how much of an increase in potential goes into enhancing the oxidation half reaction, and how much goes into deterring the reduction reaction. If α is equal to 0.5, the energy from a change in voltage will be equally divided between promoting the oxidation reaction and retarding the reduction reaction. If α is greater than 0.5, then for an increase in voltage, the oxidation reaction is promoted more than the reduction reaction is retarded. If α

is less than 0.5, then for an increase in voltage, the oxidation reaction is promoted less than the reduction reaction is retarded.

If certain parameters remain constant the relationship for a particular analyte, between concentration and current at a given voltage (R_{ij}) will be a fixed ratio. These parameters include temperature, voltage sweep rate, solution conditions, and electrode condition (for example, electrode condition can change due to a change in the effective area of the electrode, due to fouling or oxide formation). If these factors change, there will be variations in R_{ij} . The non-linear nature of LSV appears under these conditions. Under these circumstances, the usual approach for quantitative mixture analysis of using a single channel per analyte will fail, because the concentration will not be proportional to the response in a single data channel. Multivariate calibration methods should be used, which can use data from many channels to determine concentration.

2.3 Multivariate Data Analysis Techniques

To obtain quantitative concentration data from overlapping voltammograms, several multivariate data analysis techniques may be used. Multi-Linear Regression (MLR), Principal Component Regression (PCR), and Partial Least Squares Regression (PLS) are often used in spectroscopic data analysis. Excellent references and tutorials are available on these techniques (8-14). Artificial Neural Networks (15) have also been applied to quantitative analysis of voltammetric data (6).

2.3.1 Multiple Linear Regression

Multiple linear regression (also called multivariate regression or multilinear regression) is a commonly used technique, and is based on simple matrix algebra. The sensor response over many channels (e.g., absorption at various wavelengths in a spectrum or currents at various voltages in a voltammogram) is related to concentration, by some regression coefficients B. In terms of matrix algebra, this is shown by the following formula.

$$A * B = C \quad (3)$$

In this formula, A is a matrix of chemical sensor responses, such as a collection of spectra or voltammograms (each row of A is a spectrum or voltammogram). B is a matrix of regression coefficients, one for each data channel (spectrum wavelength or voltammogram voltage) and analyte. C is a matrix of concentrations of the analytes producing the sensor responses. Applying matrix algebra, the solution for the regression coefficients is found by the following formula.

$$B = (A' * A)^{-1} * A' * C \quad (4)$$

In this formula, A' is the transpose of the matrix A, and the superscript “-1” denotes the inverse of a matrix. For samples of unknown concentration, the product of a matrix of responses (spectra or voltammograms) is multiplied by the matrix of regression factors to yield the estimate of concentrations in the samples (8).

The MLR solution is based on all information provided, which includes the analyte response, contributions from interfering species, and noise. All data are weighted equally. Because of

these characteristics of MLR, this technique is best used for well-defined systems with little or no extraneous contributions to the signal. Its use assumes linear data.

2.3.2 Principal Component Regression

The power of principal component regression (PCR) is its ability to transform the data to lower dimensionality. This means that it is capable of screening out noise and redundancy. It reduces the dimensionality of the data by determining eigenvalues of the data matrix. The greater the magnitude of the eigenvalue, the greater the influence of the corresponding factor on the final data set.

For PCR regression, eigenvalues are first determined from the matrix of chemical response data. For all but the smallest data sets, eigenvalue calculation is accomplished by computer programs. The magnitude of the eigenvalues is an indication of how many components contribute to the chemical response. For example, a matrix of spectra with a thousand channels in each spectrum, which spectra are due to two chemical components at various concentrations, can generate one thousand eigenvalues. Only two eigenvalues will be large in comparison to the other nine hundred and ninety-eight.

To calculate unknown concentrations with PCR, one possible approach is as follows. First, the chemical response matrix eigenvalues are determined. Then, the matrix of eigenvalues is truncated to eliminate non-essential dimensions of the data. For example, only the two largest eigenvalues would be used for the situation discussed in the previous paragraph. The remaining eigenvalues are used along with the matrix of chemical response data to generate a calibration factor matrix, as follows.

$$P = VC' * A \quad (5)$$

$$F = C * P' * (P * P')^{-1} \quad (6)$$

$$F_{cal} = F * VC' \quad (7)$$

In this series of equations, VC is the eigenvalue matrix, and VC' is its transpose. P is the product of the transpose of VC and A. As before, A is the matrix containing chemical responses (spectra or voltammograms). F is a matrix containing the calibration in factor space, which, for this study, is not used directly, but is used to generate F_{cal} , the calibration matrix.

After F_{cal} is determined, new chemical data (in the matrix A_{new}) can be analyzed to determine unknown concentrations, as follows.

$$C_{est} = F_{cal} * A_{new} \quad (8)$$

For this final formula, C_{est} is the matrix of estimated concentrations, and A_{new} is the new matrix of chemical data (11). Again, spectra or voltammograms are examples of the type of chemical data in A.

As with MLR, PCR assumes the data are linear, but, unlike MLR, PCR can discriminate against contributions from noise.

2.3.3 Partial Least Squares Regression

Partial least squares (PLS) regression, like principal component regression, uses factor-reduced data. During the calibration phase, it also uses concentration data to optimize the regression for the components of interest, while minimizing the contributions from interfering species and noise.

There are different approaches to PLS calibration (12,13). One technique uses singular value decomposition, or SVD, to obtain the necessary matrices for PLS regression (13). The singular value decomposition process factors a matrix into three matrices, typically referred to as U, S, and V. S is a diagonal matrix, of the same dimension as X and with nonnegative diagonal elements in decreasing order. The non-zero elements of S are the squares of the eigenvalues for matrix X. Matrices U and V are unitary (that is, $U^*U = V^*V = I$, the identity matrix). If a matrix X is factored into U, S, and V by singular value decomposition, then the following formula is true.

$$X = U * S * V^T \quad (9)$$

In partial least squares regression by singular value decomposition, the matrix that is factored is the product of A' and C. A' is the transposed matrix of spectra or voltammograms, and C is the matrix of analyte concentrations associated with each spectrum or voltammogram in matrix A. Because the singular value decomposition is being performed on a matrix that contains both chemical response and concentration data, the elements of S reflect only those factors in the data matrix, A, that are correlated to concentrations in matrix C. The effects due to contributions to the data in A, which are not correlated with concentrations in C, are minimized. Examples of non-correlated effects include interfering chemical species and noise.

For development of a PLS calibration, the following formulae are used.

$$A' * C = U * S * V' \quad (\text{singular value decomposition process}) \quad (9)$$

$$B = V * S^{-1} * U' * C \quad (10)$$

Once the regression matrix, B, is calculated, estimated concentrations can be determined for a new matrix of chemical data (A_{new}), as follows.

$$C_{\text{est}} = A_{\text{new}} * B \quad (11)$$

As before, C_{est} is the matrix of estimated concentrations for the spectra or voltammograms in A_{new} (13)

PLS regression, like PCR and MLR, also assumes that data respond linearly to concentration. Of the three linear data analysis techniques mentioned here, PLS regression is typically the most robust, due to its ability to discriminate against both noise and contributions from other interferences.

2.3.4 Artificial Neural Networks

Unlike the three previously discussed analysis techniques, artificial neural networks (ANN) do not assume that the response of data to concentrations of analytes is a linear relationship. The implementation of a non-linear feedforward backpropagation neural network, such as is used in this study, can be explained with the aid of a diagram, shown here as Figure 2.

In Figure 2, the inputs to the ANN have weights applied to them; that is, there are three sets of factors for the eight inputs shown in Figure 2. Eight inputs into three hidden neurons require twenty-four weighting factors for the non-linear neurons. Each non-linear neuron also has a bias added to the sum of the weighted inputs to it: three neurons in the hidden layer require three biasing values. The sum of the weighted inputs, plus the bias, is fed in to the non-linear function. The following sigmoidal function is often used in non-linear neural network applications (18).

$$y_j = \frac{1}{1 + e^{-x_j}} \quad (12)$$

In this equation, y_j is the output of the j th neuron, and x_j is the input to that same neuron, as described by the next equation:

$$x_j = b_j + \sum_{i=1}^8 d_i w_{i,j} \quad (13)$$

Here, b_j is the bias added to the summation of inputs to the neural net, d_i multiplied by the eight weighting factors for the j th neurons input values.

Figure 2. Block Diagram of a Feed-Forward, Non-Linear Neural Network,
with Eight Input Nodes, Three Non-Linear Neurons in the Hidden layer,
and Two Linear Output Neurons

Outputs from the non-linear neurons go to the output neurons, with the next set of weighting factors and biases applied. The simplest linear neuron may be used, for which the output equals the input. For Figure 2, the inputs to the linear output neurons is found by the following equation:

$$z_k = b'_k + \sum_{j=1}^3 y_j w'_{j,k} \quad (14)$$

Here, b'_k is the bias added to the summation of outputs from the hidden neurons, y_j , multiplied by the three weighting factors for the k th neuron's input values. The output from the final layer is the same as the input for the linear transfer function, and is the estimate of the concentration.

For the type of neural network used in this study, the method used to determine the best weights and biases for estimating concentrations is called backpropagation. The mean squared error for a training set is minimized by varying the weights and biases. The backpropagation refers to the process by which the output is compared to the target values in the training set, and the effect on the error due to each weight and bias is determined. A training set consists of selected observations (in this case, voltammetric data) and their corresponding concentrations. The errors being minimized are the differences between the outputs of the neural network and the training set concentrations. The neural network's weights and biases are changed in a manner that lowers the sum of the squared errors. The new weights and biases are used in the next iteration of training. Each cycle of training (from one change in weights and biases to the next) is called an epoch. Hundreds of epochs may be required to reach the minimum error (17).

Information on the use of non-linear ANN's is found in many references and tutorials (15-18). Artificial neural networks are best employed when the analyst knows the data are non-

linear in nature, and has no reliable model with which to relate measurements to concentrations.

3 Experimental Approach

3.1 Instrumentation

All programs used ran on an IBM clone personal computer, with a 100 MHz Pentium™ CPU. Programs were run under the Microsoft Windows95™ operating system.

3.2 Digital Simulation

In order to generate a large number of linear sweep voltammograms of controllable variation in response to concentrations, simulated voltammograms were generated by DigiSim™ 2.0 (Bioanalytical Systems, W. Lafayette, In.). This program allows the setting of all of the variables shown in equation 2 above (19). The voltammograms generated from a simulated experiment may be saved as tables of voltages and currents for export to other programs. Each voltammogram reflects a combined response from two single-electron transfer reactions.

3.3 Generation of the Sample Space

Each simulated voltammogram was due to the reductions of two electroactive species. It was decided that values for concentration, formal potential, transfer coefficient, and standard heterogeneous rate constant would be varied.

The concentration of the electroactive species ranged from 0.0001 to 0.001 moles per liter. Possible values for concentration were in increments of 0.0001 moles per liter. Values for the transfer coefficient, α , were set at 0.3, 0.5, or 0.7. Values for the standard heterogeneous rate constant, k^0 , were 0.01, 0.001, or 0.0001 cm/s for data set 1, and 0.018, 0.005, or 0.002 cm/s

for data sets 2 and 3. In data set 1, the values of k^0 were set arbitrarily, but in data sets 2 and 3, the values were chosen in order to limit the current peak variability to 100 mV. The values of k^0 were chosen to allow the current peaks to shift from their nominal locations (where they would be if k^0 was very large) to the maximum desired displacement. Values of α were chosen to vary from 0.3 to 0.7 due to a statement that “In most systems α turns out to lie between 0.3 and 0.7...” from Bard and Faulkner (7). The difference in formal potentials for the two analytes was set at 100 mV for data sets 1 and 2, and 150 mV for data set 3. Specifically, the formal potential for each analyte did not vary for a set of synthetic mixture voltammograms, but the observed peak potentials for each component varied due to changes in k^0 , α , and C^* .

In generating the data sets, it was important to prevent any correlation between the parameters that were being varied. For k^0 and α , orthogonality of the data sets was ensured by experimental design (20). For the concentrations for the two analytes, the values were randomized, and absence of correlation verified using an Excel™ spreadsheet to calculate the correlation between the varied parameters.

Each voltammogram also had a noise component. The noise added was generated by the DigiSim™ 2.0 program, at a level of 1×10^{-8} amperes. This level was chosen to approximate levels typically experienced during the author's previous LSV experiments with metallic analytes at millimolar concentrations, using a hanging mercury drop electrode and silver-silver chloride reference electrode.

It is also important to note that some things that could affect data non-linearity were not changed. These were held constant to facilitate experimental design, or because plausible changes in that factor did not cause significant changes in current peak shape or height, or because that factor could be easily controlled. Temperature was always set at 298.3 K (25 °C). Diffusion coefficients for both analytes A and C were set at $1 \times 10^{-5} \text{ cm}^2/\text{s}$. The sweep rate (rate at which electrode voltage was changed) was always 100 mV per second. The starting voltage was 0.0 volts, and ending voltage was -1.024 volts. Voltage and current were recorded every four millivolts.

In summary, three data sets were generated. Each data set contained 220 voltammograms. Data set 3 had the two analytes' peak potentials nominally 150 mV apart, with variability of each peak being about 100 mV. Data set 2 had the nominal analyte peak potentials about 100 mV apart, with the same variability in location of the peak. Data set 1 had the peaks nominally 100mV apart, but, due to the selection of the wide range of rate constant values, the variability of the peaks' locations was greater than 200 mV. The intention was for data analysis to be relatively easy for data set three, more difficult for data set two, and most difficult for data set one. Examples of voltammograms for a wide range of conditions are shown in Figure 2, for each of the three data sets. These voltammograms point out the non-linear nature of the simulated systems. All bulk concentrations in Figure 2 are the same: $[A] = [C] = 5 \text{ mM}$.

Figure 3. Comparison of Variability of Voltammogram Shapes in Data Set 3 (upper left), Data Set 2 (upper right), and Data Set 1 (lower center)

3.4 Selection of Cross-Validation Sets

Each data set was divided into nine cross-validation sets. For a particular data set, eight of the nine sets were used by the MLR, PCR, PLSR, and ANN routines to calibrate the technique. The left-out set was then analyzed as a predictor set, in order to determine how well the concentrations were estimated by the various techniques. This cross-validation process was repeated until all of the leave-one-out sets had been used as a predictor set. Overall performance is based on the standard error of cross-validated prediction over all nine subsets.

The criterion for a leave-one-out set was based on the degree of peak overlap. Within a data set, peak overlap was determined by each analyte's heterogeneous rate constant (k_A^0 and k_C^0) and transfer coefficient (α_C and α_A). There are twenty-seven combinations of heterogeneous rate constants and transfer coefficients within each data set. To get nine cross-validation sets, each of the nine possible combinations of k_A^0 and k_C^0 were chosen. All voltammograms with a particular k_A^0 and k_C^0 were saved out, not used for calibration, and then analyzed with the calibration models generated from the remaining voltammograms.

Specifically, each analyte had a possibility of being assigned a standard heterogeneous rate constant, k^0 , which was slow, medium, or fast. For data sets 2 and 3, a slow value for k^0 was 0.002 cm/s, a medium value was 0.005 cm/s, and a fast value was 0.018 cm/s. For data set one, a slow value for k^0 was 0.0001 cm/s, a medium value was 0.001 cm/s, and a fast value was 0.01 cm/s. The relative shapes can be seen in Figure 4. Variations in the peak shapes appear because, within a given set of kinetic parameters, the transfer coefficient (α) for analyte A varies. The smaller the standard heterogeneous rate constant, k^0 , the greater the effect caused by the transfer coefficient (α).

A consequence of using the standard heterogeneous rate constants to select cross-validation sets was that, within a prediction set for cross validation, only one value for analyte C's transfer coefficient (α) occurred. Although these parameters were not correlated for a given data set's entire 220 voltammograms, they were correlated for the cross-validation subsets. For example, in Data Set 2, for $k_A^0 = 0.005$ cm/s and $k_C^0 = 0.002$ cm/s, $\alpha_C = 0.3$ for all prediction set voltammograms. This can be seen in Figure 4, by noting that the current peak on the right

changes within a cross-validation set, whereas the current peak on the left does not. This may not be apparent where the peaks are not resolved. For the voltammograms pictured in Figure 4, it is helpful to note that the current peak due to analyte A is on the right, and the one due to analyte C is on the left.

Figure 4. Variation in Shapes of Voltammograms in the Leave-One-Out groups in Data Set 2: $[A] = [C] = 5 \text{ mM}$

Each of the above voltammograms in Figure 4 corresponds to a cross-validation subset. These subsets are given designations in Table 1. In Table 2, the subset designators are in the same relative locations as their corresponding voltammograms in Figure 4.

Table 1. Designators and Corresponding Kinetics Values for Cross-Validation Subsets

Subset Designator	Data Set 3		Data Set 2		Data Set 1	
	k_A^0 (cm/s)	k_C^0 (cm/s)	k_A^0 (cm/s)	k_C^0 (cm/s)	k_A^0 (cm/s)	k_C^0 (cm/s)
L1	0.002	0.002	0.002	0.002	0.0001	0.0001
L2	0.002	0.005	0.002	0.005	0.0001	0.001
L3	0.002	0.018	0.002	0.018	0.0001	0.01
L4	0.005	0.002	0.005	0.002	0.001	0.0001
L5	0.005	0.005	0.005	0.005	0.001	0.001
L6	0.005	0.018	0.005	0.018	0.001	0.01
L7	0.018	0.002	0.018	0.002	0.01	0.0001
L8	0.018	0.005	0.018	0.005	0.01	0.001
L9	0.018	0.018	0.018	0.018	0.01	0.01

Table 2. Correspondence between Subset Designators and Figure 4.

C Slow - more peak overlap	L7	L4	L1
C Medium	L8	L5	L2
C Fast - less peak overlap	L9	L6	L3
	A Fast - more peak overlap	A Medium	A Slow - less peak overlap

3.5 MATLAB™ Analysis

MATLAB™ (the MathWorks, Inc., Natick, MA.) is a high-level programming language, specifically designed to handle matrix mathematics. It has toolboxes (add-in programming) that enable it to perform efficiently in specialized applications. This study included the use of the Chemometrics and Neural Networks toolboxes, with MATLAB™ version 5.1.

3.5.1 Data Pretreatment

Data pretreatment was an essential step in the process of data analysis. Pretreatment consisted of mean-centering and normalizing. In the mean-centering step of pretreatment, the average of the currents at a given voltage was calculated, then a mean-centered set of data was determined by subtracting the mean current at each voltage from the current in each individual voltammogram. The normalized set of data was produced by dividing the mean-centered currents at each voltage by the standard deviation at each voltage. The final set of data analyzed had, at each voltage, a mean of zero and a standard deviation of one.

Figure 5. Changes in a Data Set during Mean-Centering and Normalization

Figure 5 shows the data pretreatment process graphically, from original voltammograms on the left, to mean-centered, normalized data on the right.

The hidden layer neurons of the artificial neural networks used here had a sigmoid transfer function, as shown in Figure 6. Failure to mean-center and normalize data at the input layer of the ANN would result in most of the data being truncated.

Because both voltammogram responses and concentrations were used to train the neural network, both sets of inputs (currents and concentrations) were mean-centered and normalized. This meant that the predictor set current values also had to be adjusted for the calibration mean and standard deviation values.

Figure 6. The Sigmoid Logistic Function

The estimated concentrations were therefore consistent with the mean-centered and normalized calibration set. Real predicted values were obtained by first multiplying the output values by the standard deviation of the calibration set concentrations, then adding the calibration set average concentration.

3.5.2 Implementation of Data Analysis Techniques

All four data analysis techniques (MLR, PCR, PLS, and ANN) were performed by MATLAB™ routines. Multilinear regression coefficients were determined by simple matrix algebra as described by Workman and Mark (8-10). Principal component(s) regression and partial least squares regression were performed through the methods provided by the MATLAB™ Chemometrics Toolbox.

The artificial neural network was implemented using the MATLAB™ Neural Networks Toolbox. The artificial neural network was much as the one described earlier (in Figure 2), but larger. There were forty-three inputs, and two output neurons, one for analyte A, and one for analyte C. The first network contained seven hidden (log-sigmoidal) neurons. Seven hidden neurons were chosen as follows: the number of hidden layer neurons was the integer closest to the square root of the sum of the number of input and output neurons:

$$\# \text{ of Hidden Neurons} = \sqrt{43 \text{ inputs} + 2 \text{ outputs}} = \sqrt{45} \approx 7 \quad (15)$$

This technique was suggested by Robb (18), and generated a value of seven, as seen above. This agrees with Zupan and Gasteiger's text (15), in which the recommendation is to use a number of hidden layer neurons that reflects the inherent parameters in the data. For the data

generated in this study, six parameters were varied. When plots of SSE versus number of factors for either PCR or PLS are examined, the apparent number of inherent parameters can be determined. The number of inherent parameters can be measured by the number of factors at which the SSE reaches a minimum or ceases to change significantly. This value typically ranged from five to ten factors for this study.

During MATLAB™ training of a feedforward back-propagation neural network, if the target error is not reached after a training cycle, a standard recommendation is displayed to try more hidden-layer neurons. Consequently, a second neural network analysis, with an ANN containing ten hidden layer neurons, was performed. Care was taken to monitor the predictor set SSE, and the training process stopped when the predictor set SSE is at a minimum. Using an excessive number of hidden layer neurons can lead to overfitting unless such a technique is used. Overfit is the condition whereby an artificial neural network is optimized too well for a specific training set, and cannot provide good prediction outside of that training set (it has a good specific solution, but not a good general solution). Data from the ANN using seven hidden layer neurons is labeled “ANN7,” and data from the neural network with ten hidden layer neurons is labeled “ANN10.”

Training of the neural network was performed with parameters selected in an effort to provide a slow-changing prediction set SSE. This was done to produce a U-shaped SSE versus training cycle curve, which provided confidence in selection of the weights and biases for the best solution. The training cycles were 110 epochs in length, with a learning rate of 0.0007, and other training parameters set to MATLAB™ default values.

Data input to the analysis routines was reduced so that only every fourth data point from the voltammograms was used, making the current data points 16 mV apart. The current values between 0.0 volts and -0.34 volts were not used, due to their being mostly noise. Only the voltage window between -0.034 and -1.024 V was used for data input to the ANN.

Prior to analysis of non-linear data, a test set of linear data was generated for testing the MATLAB™ routines. The linear data for the pre-test were the sums of overlapped gaussian peaks, with twenty data channels per “spectrum,” and twenty-four “spectra.” Analysis of nonlinear voltammetric data began after successful completion of this test.

3.5.3 Figures of Merit

The root-mean-squared error (RMSE) was chosen as one of the indicators of adequacy of a solution. The formula for the RMSE is shown below in equation (16).

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (c_i - \hat{c}_i)^2}{n}} \quad (16)$$

Where n is the number of data points in the predictor set, c_i is the actual value of the concentration of the predictor set, and $\hat{c}_i$ is the predicted concentration. When expressed as a percent relative to the average of the data, the percent root-mean-squared error, or PRMSE, is generated.

$$\text{PRMSE} = \frac{\sqrt{\frac{\sum_{i=1}^n (c_i - \hat{c}_i)^2}{n}}}{\frac{1}{n} \sum_{i=1}^n c_i} * 100\% \quad (17)$$

To further examine the output of a data analysis technique, other figures of merit were employed. The slope of the graph of predicted concentrations versus actual concentrations gave a second indicator of accuracy: a perfect solution would yield a slope of unity. Related indicators are the coefficient of determination (R^2) and standard error of the regression (s_{yx}). The bias of the data sets was also examined. All values were calculated in MATLAB™.

3.5.4 Optimization Using Predictor Set Error

The sum of squares of residuals (“sum-squared error”) is a useful metric for evaluating the calibration model. For a fixed number of samples, the sum-squared error (SSE) is proportional to the square of the RMSE. The dependence of the SSE on the number of factors used for PCR and PLS regression was determined. The number of factors that generated the minimum SSE value was chosen as the number used for the PCR or PLS calibration model. The SSE values calculated utilized data for both analytes A and C.

To determine the appropriate calibration model from ANN training, the training SSE can be examined as a function of training epochs. Unfortunately, with non-linear ANN’s, the training SSE can continue to decline as the data are overfit by the network. Thus, to avoid overfitting, the predictor set SSE was used for optimizing the training of the neural network. After each ANN training cycle (of 110 epochs), the prediction SSE was determined. The values of the

weighting factors and biases for the neurons in the net are saved for the training cycle that generates the lowest predictor set SSE. This optimum set is used to evaluate the prediction performance of the ANN.

A plot of the sum of squares errors versus the number of factors in PCR and PLS regression, and versus the training epochs in the ANN training, is shown as Figure 7. PCR and PLS regression generate a unique set of calibration coefficients for each set of factors. ANN's have non-unique solutions, depending on the (random) set of initial weights and biases. Therefore, the analysis for several replicate training events are produced in Figure 7 for the ANN. The error bars represent the one-sigma confidence interval for the SSE of the predictor set. From Figure 7, the optimum number of PLS factors was identified as seven. For PCR, the optimum number of factors was eight. The ANN best solution was reached at training cycle 12 (after 1320 epochs).

These optimum values varied depending on the specific calibration set being used. Typically, the greater the overlap of current peaks in the prediction set, the higher the number of factors required for PCR and PLS. For data sets 2 and 3, PCR and PLS typically needed between 5 and 10 factors. For data set 1, due to its greater degree of non-linearity, more factors were required to generate the lowest SSE. The optimum number of training cycles for the artificial neural networks averaged around 12 (1320 epochs) for data sets 2 and 3. For data Set 1, with the greatest non-linearity, SSE results were higher than for data sets 2 and 3, and the number of training cycles needed to reach the minimum SSE was lower, with an average of 3 (330 epochs).

Figure 7. Use of Sum of Squared Errors to Optimize Data Analysis

4 Results

Estimated concentrations for the predictor set, generated by cross-validation, were compared to reference concentrations. Cross-validation results were based on nine leave-one-out cases in each of the three data sets. Plots, for a single prediction subset's results from data set two, are shown as Figure 8. Plots for one of the data set 1 predictor sets are shown as Figure 9. A representative plot for data set three is not shown, due to data set three's predictor set plots looking quite similar to ones from data set two. The differences due to greater non-linearity of data are evident in the poor prediction values seen in the results from data set 1. The results for the entire cross-validated data sets are tabulated later.

Figure 8. Output from Four Data Analysis Techniques for Data Set 2, Subset L1. The solid line is the ideal calibration line: slope = 1, and intercept = 0

Analyte A, Data Set 1, Cross-Validation Subset L9

Analyte C, Data Set 1, Cross-Validation Subset L9

Figure 9. Output from Four Data Analysis Techniques for Data Set 1, Subset L9.

The solid line is the ideal calibration line: slope = 1, and intercept = 0

4.1 Comparison of Cross-Validated Prediction Results

For each data set, the cross-validated percent RMSE (PRMSE), slope of the cross-validated prediction versus true concentration line, bias, correlation coefficient, R^2 , and standard error of the regression line fit, $s_{y_{\text{pred}}}$ were calculated for both analytes A and C.

Below, in Figures 10, 11, and 12, one can see the variation in the results from the cross-validated predictor sets. Note that data set 1, having the greatest nonlinearity, has PRMSE values much higher than the other two data sets. Within each data set, one can also see cases where the data analysis technique had more difficulty than with other cross-validation sets. Only the results for analyte A are shown, as the results for analyte C are similar to those of A.

Figure 10. PRMSE Values for Cross Validation from Data Set 1

For data set one, one can see that the cross validation sets with $k_C^0 = 0.0001$ (L1, L4, and L7) had higher PRMSE values than the other leave-one-out sets. The reason is that the small rate constant resulted in a shift of the current peak of analyte C to create more overlap with that of analyte A.

In data sets 2 and 3, a slightly different variation in PRMSE with kinetics parameters is seen. In data set 2 (Figure 11), the cross-validation sets with the highest PRMSE are the ones for which analyte A's kinetics are fast relative to those of analyte C. In the case of subset L4, where $k_A^0 = 0.005$ and $k_C^0 = 0.002$, the transfer coefficient for analyte C, α_C , is 0.3 for all voltammograms. This creates greater overlap of analyte C's current peak on to analyte A's peak. For subsets L7 and L8, in which $k_A^0 = 0.018$ and $k_C^0 = 0.002$, and $k_A^0 = 0.018$ and $k_C^0 = 0.005$, the overlap is also affected by analyte C's transfer coefficient. One might expect the overlap to be worse for L7, but for this cross-validation set, the transfer coefficient for analyte C was 0.7, which tends to resolve the two current peaks. The overlap is worsened, and PRMSE consequently higher, for cross-validation set L8, due to analyte C's transfer coefficient being 0.5 (rather than 0.7 as it was for subset L7). Of note, these three subsets with the worst PRMSE values are the three in the upper-left corner of Figure 4. In other words, the most difficult prediction subsets are those where the greatest peak overlap occurs.

Figure 11. PRMSE Values for Cross Validation from Data Set 2

Figure 12. PRMSE Values for Cross-Validation from Data Set 3

In data set 3 the same effects of overlap are seen (Figure 12). However, subset L7 shows lower PRMSE than in data set two, due mainly to the greater separation of the formal potentials of the two analytes.

4.2 Bias of Predicted Concentrations

Bias was calculated for each data point, each cross-validation subset (L1 through L9), and for each data set. Bias was seen to be larger on data set one results than on the results for sets two or three. This difference in bias is shown in Figures 13, 14, and 15, which follow. One can also see increasing bias as concentration increases for analyte C in all data sets.

Figure 13. Biases for All Points in Data Set 1, Plotted versus Concentration

Figure 14. Biases for All Points in Data Set 2, Plotted versus Concentration

Figure 15. Biases for All Points in Data Set 3, Plotted versus Concentration

Also of note is that, for a given data analysis technique, the bias of analyte A is opposite in sign to that of analyte C, and the biases are roughly equal in magnitude. This is generally true for all data sets and cross-validation subsets. This is shown below, for data set one, in Figure 16.

Figure 16. Prediction Biases for Cross-Validation Subsets in Data Set 1 -

Note the Symmetry about Zero for Analytes A and C

4.3 Slope of Regression Line for Predicted Concentrations versus Actual Concentrations

Another figure of merit calculated for all data analysis techniques and all data sets was the slope of the best-fit line for predicted concentrations (dependent variable) versus actual concentrations (independent variable). With one exception, that of multilinear regression of analyte A on data set three, which had a slope of 1.006 ± 0.024 (2σ error), all slopes were less

than 1. For data sets 2 and 3, the slope for analyte A was higher than that of analyte C for all data analysis techniques. For data set one, the slope was higher for analyte A than it was for analyte C in MLR and ANN analyses, but lower for analyte A than it was for analyte C in PCR and PLS analyses.

4.4 Correlation of Bias and Slope

It is interesting to note that in any data set, within any analysis technique tested, that the analyte for which the regression line slope was higher had a positive bias, and the analyte with the lower slope had a negative bias. It is easiest to see this in data set one, where the biases are of sufficient magnitude to separate the best-fit lines from the zero-bias line (i.e., the ideal calibration line with a slope of 1 and an intercept of zero). This can be seen in Figure 17.

Figure 17. Comparison of Regression Lines

Although Figure 17 shows the worst case for bias and slope, the basic features of the regression lines are the same for other data analysis techniques results in all three data sets. These features include the following items. First, the slope of the upper line is greater than the slope of the lower line. Second, the intercept of the regression line is greater than zero. Lastly, the slopes of regression lines are less than one (with the exception, as mentioned before, of the PCR result for data set three).

An explanation for the features listed above may be found in the way that the non-linear data vary, coupled with the way the regressions are optimized by minimizing the combined SSE for both analytes. Changes in current peak location and shape (due to changes in α and k^0) result in the data channels' responses to a given concentration changing. The variability of the current response to analyte concentration is proportional to the concentration of the analyte. Therefore, for data resulting from high concentrations, the SSE will be greater. The data analysis techniques are all designed to minimize the SSE. The increase in SSE at high concentrations can be mitigated by reducing the gain on the data channels. This causes the average predicted data to be in error with a negative bias. The negative bias is also manifested as an increased SSE, which can be decreased by adding a constant value to all concentration estimates, which manifests itself on the calibration line as a positive regression intercept. The combination of lowering the gain and raising the value of the added constant results in the lowest SSE.

The divergence of the two calibration lines, with that for analyte A typically being higher, can be explained by the way peak shapes vary. When the current peak shifts, due to lower k^0 or α

values, it shifts away from its nominal location, and towards lower voltages, as simulated in this study. Referring to Figures 1, 3, or 4, this shift is to the right (voltage is more negative at the right of the abscissa, because these figures are in “classical” format for voltammograms, which reverses the values on the x-axis). In these voltammograms, the current peak for analyte A is on the right, and C is on the left. Therefore, when the current peak for analyte A shifts, it shifts away from the current peak for analyte C. When the current peak for analyte C shifts, it shifts towards the current peak for analyte A. This results in analyte A’s receiving a positive contribution from shifts in the peak location for analyte C.

4.5 Comparison of Sets

The values for PRMSE, slope, correlation coefficient, the standard error of the regression line, and the data set bias were based on cross-validated prediction values for all samples in each data set. The 95% confidence interval for the correlation coefficients was calculated using the method of Hoel (21).

The figures of merit for each data set are given in Table 3. Data Set 1 had a significantly greater degree of non-linearity of the data than did either set two or set three. Values for the best result (lowest PRMSE, slope closest to 1, R^2 closest to 1, lowest s_{yx} , and bias closest to zero) are in bold type.

Table 3. Figures of Merit for Cross-Validated Predictions

(95% Confidence Intervals, where given)

Figure of Merit		Data Set 1	
		Analyte A	Analyte C
PRMSE	MLR	68.1%	67.7%
	PCR	30.2%	30.4%
Best value is	PLS	29.4%	29.4%
nearest to zero	ANN7	28.5%	30.0%
	ANN10	29.8%	30.7%
Slope	MLR	0.807 ± 0.170	0.658 ± 0.165
± 2σ	PCR	0.575 ± 0.051	0.661 ± 0.063
Best value is	PLS	0.629 ± 0.057	0.684 ± 0.063
nearest to 1	ANN7	0.725 ± 0.063	0.707 ± 0.066
	ANN10	0.739 ± 0.068	0.715 ± 0.068
R ²	MLR	0.293 (0.164 - 0.411)	0.226 (0.094 - 0.350)
with 95% C.I.	PCR	0.703 (0.627 - 0.765)	0.671 (0.589 - 0.739)
Best value is	PLS	0.692 (0.614 - 0.756)	0.684 (0.604 - 0.749)
nearest to 1	ANN7	0.707 (0.632 - 0.768)	0.677 (0.597 - 0.744)
	ANN10	0.686 (0.607 - 0.751)	0.667 (0.584 - 0.735)
S _{yx}	MLR	36.20	35.14
values x 10 ⁵	PCR	10.80	13.36
Best value is	PLS	12.12	13.42
nearest to zero	ANN7	13.48	14.09
	ANN10	14.44	14.58
Data Set Bias	MLR	0.0190	-0.0179
	PCR	-0.0075	0.0062
Best value is	PLS	-0.0034	0.0025
nearest to zero	ANN7	0.0044	-0.0048
	ANN10	0.0052	-0.0061

Table 3, Continued

Figure of Merit		Data Set 2	
		Analyte A	Analyte C
PRMSE	MLR	9.7%	9.8%
	PCR	7.8%	7.9%
Best value is	PLS	6.7%	6.7%
nearest to zero	ANN7	8.4%	8.4%
	ANN10	8.2%	8.4%
Slope	MLR	0.995 $\pm$ 0.025	0.968 $\pm$ 0.025
$\pm 2\sigma$	PCR	0.996 $\pm$ 0.020	0.978 $\pm$ 0.020
Best value is	PLS	0.990 $\pm$ 0.017	0.985 $\pm$ 0.017
nearest to 1	ANN7	0.982 $\pm$ 0.021	0.972 $\pm$ 0.021
	ANN10	0.983 $\pm$ 0.021	0.976 $\pm$ 0.021
R ²	MLR	0.967 (0.957 – 0.975)	0.966 (0.956 - 0.974)
with 95% C.I.	PCR	0.979 (0.972 – 0.984)	0.978 (0.971 - 0.983)
Best value is	PLS	0.984 (0.979 – 0.988)	0.984 (0.979 - 0.988)
nearest to 1	ANN7	0.975 (0.967 – 0.981)	0.975 (0.968 - 0.981)
	ANN10	0.976 (0.969 – 0.982)	0.976 (0.968 - 0.981)
S _{yx}	MLR	5.28	5.22
values $\times 10^5$	PCR	4.26	4.22
Best value is	PLS	3.66	3.61
nearest to zero	ANN7	4.53	4.45
	ANN10	4.44	4.46
Data Set Bias	MLR	0.0021	-0.0022
	PCR	0.0016	-0.0016
Best value is	PLS	0.0012	-0.0010
nearest to zero	ANN7	0.0017	-0.0018
	ANN10	0.0017	-0.0017

Table 3, Continued

Figure of Merit		Data Set 3	
		Analyte A	Analyte C
PRMSE	MLR	9.5%	10.4%
	PCR	6.5%	6.7%
Best value is	PLS	6.0%	7.0%
nearest to zero	ANN7	6.4%	7.5%
	ANN10	6.3%	7.6%
Slope	MLR	1.006 ± 0.024	0.979 ± 0.027
$\pm 2\sigma$	PCR	0.989 ± 0.016	0.975 ± 0.017
Best value is	PLS	0.995 ± 0.015	0.987 ± 0.018
nearest to 1	ANN7	0.987 ± 0.016	0.973 ± 0.019
	ANN10	0.987 ± 0.016	0.979 ± 0.019
R^2	MLR	0.969 (0.960 - 0.976)	0.961 (0.949 - 0.970)
with 95% C.I.	PCR	0.985 (0.980 - 0.989)	0.984 (0.979 - 0.988)
Best value is	PLS	0.987 (0.983 - 0.990)	0.982 (0.977 - 0.987)
nearest to 1	ANN7	0.985 (0.981 - 0.989)	0.979 (0.972 - 0.984)
	ANN10	0.986 (0.982 - 0.989)	0.979 (0.973 - 0.984)
S_{yx}	MLR	5.18	5.69
values $\times 10^5$	PCR	3.51	3.59
Best value is	PLS	3.29	3.81
nearest to zero	ANN7	3.48	4.06
	ANN10	3.41	4.13
Data Set Bias	MLR	0.0020	-0.0014
	PCR	0.0012	-0.0015
Best value is	PLS	0.0010	-0.0012
nearest to zero	ANN7	0.0008	-0.0009
	ANN10	0.0010	-0.0009

4.6 Best Analysis Methods

Given the cross-validated prediction results from the leave-one-out sets, the following table may be constructed.

Table 4. Best Techniques for Data Analysis, Based on Figures of Merit

Figure of Merit	Data Set 1		Data Set 2		Data Set 3	
	Analyte A	Analyte C	Analyte A	Analyte C	Analyte A	Analyte C
PRMSE	ANN	PLS	PLS	PLS	PLS	PCR
Slope	ANN	ANN	PCR	PLS	PLS	PLS
	(MLR)					
R^2	ANN	PLS	PLS	PLS	PLS	PCR
S_{yx}	PCR	PCR	PLS	PLS	PLS	PCR
Bias	PLS	PLS	PLS	PLS	ANN	ANN

Note that multiple linear regression appears parenthetically in Table 1. This means that its figure of merit would result in its being the best data analysis technique for the analyte A slope in Data Set 1. However, the variance in estimates of the slope from MLR results was great, and the correlation coefficient (R^2) was very low, so it should not be assumed that MLR does as well as the other techniques.

4.6.1 Artificial Neural Network Comparison

By looking at the figures of merit in Table 3, the following table (Table 5) can be constructed. It shows which of the two ANN techniques, the one with seven hidden layer neurons (ANN7), or the one with ten hidden layer neurons (ANN10), did better for that particular

analysis. It should be noted that the figures of merit from both neural networks were often nearly equal. Although the differences were small, ANN7 was better in ten instances, and ANN10 was better in fourteen. There were six instances when the figure of merit from ANN7 was equal to that from ANN10.

Table 5. Comparison of ANN Analyses, Based on Figures of Merit

Figure of Merit	Data Set 1		Data Set 2		Data Set 3	
	Analyte A	Analyte C	Analyte A	Analyte C	Analyte A	Analyte C
PRMSE	ANN7	ANN7	ANN10	ANN10	ANN10	Tie
Slope	ANN10	ANN10	ANN10	ANN10	Tie	ANN10
R ²	ANN7	ANN7	ANN10	ANN10	ANN10	Tie
S _{yx}	ANN7	ANN7	ANN10	ANN10	ANN7	Tie
Bias	ANN7	ANN7	Tie	ANN10	ANN7	Tie

4.7 Observations

Generally, principal components regression worked best for analyte C in data set 3. Partial least squares regression worked best for analyte A in data set 3, both analytes in data set 2, and analyte C in data set one. PLS did well with analyzing all data sets. The artificial neural network worked well for data set 1, and was the best technique for analyte A of data set one. Principal component(s) regression typically tracked along with PLSR, although not doing quite as well in most cases. Multilinear regression was not suitable for data analysis, especially when compared to the other techniques.

These observations are generally consistent with our previous experience with these different multivariate calibration methods. Feedforward backpropagation ANN's are more useful for developing nonlinear calibration models; but may not perform as well as PLSR or PCR for data that can be fit with multivariate linear models. This latter observation can often be attributed to the fact that the feedforward backpropagation ANN may generate a non-linear fit to calibration data that are inherently linear. This essentially creates an overfitting model, which will not predict well.

5 Conclusions

It is notable that the data analysis techniques worked as well as they did, given the non-linear nature of the data. This is a clear endorsement of the inherent advantages of multivariate calibration methods. That is, although each individual data channel has a nonlinear response, calibration techniques using information from multiple channels can obtain quantitative results. An important pre-requisite for successful calibration is a data set spanning the entire parameter space. Although this calibration data satisfied this requirement, cross-validation methods may compromise the calibration space. A better approach would involve the creation of separate, independent, calibration and prediction sets, where the calibration set spans a greater parameter space than the prediction set. This approach would particularly benefit the ANN method.

Given the deliberate difficulty of the calibration problem addressed here, the various commercially available multivariate calibration techniques worked quite well. The future step of applying these to real voltammetric systems appears promising. The most challenging

aspect will be collecting representative calibration data, which will span the anticipated parameter space.

6 Future Work

This avenue of investigation would benefit from comparative analyses of data sets that are less discrete in nature. Filling in the gaps between the one-millimolar spacing used in the data sets could force the ANN into a different fit.

Other variations on the neural net could be attempted. More neurons in the hidden layer is a standard suggestion displayed by the MATLAB™ Neural Network Toolbox whenever a neural network training cycle does not reach the target SSE. Another possibility would be to include another hidden layer in the neural network. A combined method such as used by Andersson, et al., (22) might also lead to better prediction results.

An additional avenue of study would be to study performance of the data analysis techniques with respect to non-linear composite spectra, perhaps with inter-analyte interactions causing the non-linear characteristics, such as that discussed by Falck, et al. (6).

7 References

1. DePalma, R. A., and Perone, S. P. "On-Line Pattern Recognition of Voltammetric Data: Peak Multiplicity Classification," Analytical Chemistry, 51(7), 1979, 825-828.
2. Byers, W. A., Freiser, B. S., and Perone, S. P. "Structural and Activity Characterization of Organic Compounds by Electroanalysis and Pattern Recognition," Analytical Chemistry, 55(4), 1983, 620-625.
3. Perone, S. P., and Ham, C. I. "Measurement and Control of Information Content in Electrochemical Experiments," Journal of Research of the National Bureau of Standards, 90(6), 1985, 531-539.
4. Gu, B., Wang, J., Zhou, X., Wu, X., Liu, F., and Li, Y. "Comparison of Multivariate Calibration Methods for Quantitative Analysis of Multi-component Mixture of Air Toxic Organic Compounds by FTIR," Journal of Environmental Science and Health, A33(7), 1998, 1419-1436.
5. Beltrán, J. L., Ferrer, R., and Guiteras, J. "Multivariate Calibration of Polycyclic Aromatic Hydrocarbon Mixtures from Excitation-Emission Fluorescence Spectra," Analytica Chimica Acta, 373, 1998, 311-319.
6. Chan, H., Butler, A., Falck, D. M., and Freund, M. S. "Artificial Neural Network Processing of Stripping Analysis Responses for Identifying and Quantifying Heavy Metals in the Presence of Intermetallic Compound Formation," Analytical Chemistry, 67, 1997, 2373-2378.
7. Bard, Allen J., and Faulkner, Larry R. Electrochemical Methods, N.Y: John Wiley and Sons, 1980.
8. Workman, J. Jr., and Mark, H. "Matrix Algebra and Multiple Linear Regression, Part I," Spectroscopy, 8(9), 1993, 16-20.
9. Workman, J. Jr., and Mark, H. "Matrix Algebra and Multiple Linear Regression, Part II," Spectroscopy, 9(1), 1993, 16-19.
10. Workman, J. Jr., and Mark, H. "Matrix Algebra and Multiple Linear Regression, Part III," Spectroscopy, 9(4), 1994, 18-19.
11. Workman, J. Jr., and Mark, H. "Calculating the Solution for Regression Techniques Part II: Principal Component(s) Regression made Simple," Spectroscopy, 12(6), 1997, 22.
12. Geladi, P., and Kowalski, Bruce R. "Partial Least Squares Regression: A Tutorial," Analytica Chimica Acta, 185 1986, 1-17.
13. Workman, J. Jr., and Mark, H. "Calculating the Solution for Regression Techniques Part III, Partial Least Squares Regression Made Simple," Spectroscopy, 12(9), 1997, 36-37.
14. Malinowski, E. Factor Analysis in Chemistry. 2d ed. New York: John Wiley and Sons, 1991.
15. Zupan, J., and Gasteiger, J., Neural Networks for Chemists. New York: VCH, 1993.

16. Jansson, P. A. "Neural Networks: An Overview," Analytical Chemistry, 63(6), 1991, 357A-362A.
17. Rao, V., and Rao, H. C++ Neural Networks and Fuzzy Logic, 2d ed. New York: MIS Press, 1995
18. Robb, E. Paper Presented at the ACS National Meeting, Symposium on Neural Networks Applied to Spectroscopy, Computers in Chemistry Division, San Francisco, Ca., April, 1992.
19. Rudolf, M., Reddy, D., and Feldberg, S. "A Simulator for Cyclic Voltammetric Responses," Analytical Chemistry, 66(10), 1994, 589A-600A.
20. Montgomery, Douglas C. Design and Analysis of Experiments. 3d ed. New York: John Wiley and Sons, 1991.
21. Hoel, Paul G. Introduction to Mathematical Statistics. 4th ed. New York: John Wiley and Sons, 1971.
22. Andersson, G., Kaufmann, P., and Renberg, L. "Non-Linear Modeling with a Coupled Neural Network-PLS Regression System," Journal of Chemometrics, 10, 1996, 605-614.

IMAGE EVALUATION TEST TARGET (QA-3)

APPLIED IMAGE . Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989

© 1993, Applied Image, Inc., All Rights Reserved