Factor-Based Analysis of Voltammetric Data for the Qualitative and Quantitative Evaluation of Complex Liquids

by

Suzanne Katherine Schreyer

A thesis

presented to the University of Waterloo

in fulfilment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Chemistry

Waterloo, Ontario, Canada, 2000

0-612-53513-4

Abstract

Factor-Based Analysis
for the Qualitative and Quantitative Evaluation
of Voltammetric Data of Complex Liquids

Factor-based techniques were used for the qualitative and quantitative analysis of voltammograms of complex liquids. Both normal pulse and square wave voltammetry were evaluated, at a variety of conditions, and with different electrode materials. Square wave voltammetry at a platinum electrode had the greatest variance in the resultant scores plots, so this techniques was used in all subsequent experiments.

Initially, individual scores plots were generated of fruit juices, beers, wines, coffees, and milks. For qualitative analysis, principal components analysis was used to differentiate sub-populations on scores plots. Enhanced differentiation of sub-populations was achieved by selecting out a portion of the voltammogram, and regenerating the pattern recognition plot.

For quantitative analysis, principal components regression and partial least squares algorithms were compared for their predictive ability. Square wave voltammograms of lactate, pyruvate, glucose and ethanol in beer were generated, over a known concentration range of the individual species, and prediction of the individual concentrations was done using PCR and PLS. To assess the accuracy and precision of PCR and PLS, correlation coefficients, and mean square errors were calculated. Further, PCR and PLS were used to predict individual concentrations of pyruvate, glucose and ethanol in a ternary mixture.

Acknowledgements

I would like to thank my supervisor, Dr. S.R. Mikkelsen for her help, guidance and patience throughout the course of this work. I would also like to thank the members of my committee, Dr. M. Baker, Dr. V. Karanassios, and Dr. J. Lipkowski, for their input.

I would also like to thank Colin Campbell for his invaluable assistance with MATLAB, and for all his computer advice.

Special thanks to my husband Dr. Joe Albanese, and my parents, Drs. Jarka and Val Srajer, for their support and patience.

Dedication

This thesis is dedicated to

Dr. Viola Birss,

Dr. W. Laidlaw and

Sr. Liliane Mercier,

who believed that I must.

Table of Contents

List of Tables

List of Figures

Abbreviations

Ag/AgCl	Silver/Silver chloride
ANN	Artificial Neural Network
Au	Gold
BOD	Biological Oxygen Demand
CLS	Classical Least Squares
DPSV	Dual Pulse Staircase Voltammetry
GCE	Glassy Carbon Electrode
IE	Imbedded Error
ILS	Inverse Least Squares
ISE	Ion Selective Electrode
MLR	Multiple Linear Regression
MSE	Mean Square Error
NIR	Near Infra-Red
NLL	Non-linear Least Squares
NNET	Neural Network
NP	Normal Pulse
OD	Optical Density
PC	Principal Component
PCA	Principal Components Analysis
PCR	Principal Components Regression

PLS	Partial Least Squares
PRESS	Predicted Residual Error Sum of Squares
Pt	Platinum
PVC	Polyvinylchloride
QCM	Quartz Crystal Microbalance
RE	Real Error
REV	Reduced Eigenvalues
RMS	Root Mean Square
RMSE	Root Mean Square Error
RMSEC	Root Mean Square Error of Calibration
RMSEP	Root Mean Square Error of Prediction
RRMSEC	Relative Root Mean Square Error of Calibration
RRMSEP	Relative Root Mean Square Error of Prediction
RSD	Residual Standard Deviation
SOM	Self-Organizing Map
SQV	Square Wave Voltammetry
SVD	Singular Value Decomposition
UV-Vis	Ultraviolet-visible
WE	Working Electrode
XE	Extracted Error

Matrix Notation

$\mathbf{D}$	Data matrix
$\mathbf{D}'$	Transpose of D-matrix
$\bar{\mathbf{D}}$	Properly dimensioned abstract solution
$\mathbf{D}^{-1}$	Inverse of D-matrix
$\hat{\mathbf{D}}$	Predicted matrix
rxc	Dimensions of matrix; rows by columns
x, y	Variables in matrix
$\hat{y}$	Predicted y-value
y_{ij}	y-value corresponding to i^{th} row and j^{th} column

Chapter 1 Introduction and Literature Review

1.1 Chemometrics –Definitions

Chemometrics can be generally defined as any of various methods used to extract chemically relevant information from data produced in chemical experiments.[1] These statistical and mathematical methods are used to transform chemical data into information to generate a working model for further decision making.[2] These models range from univariate analysis using classical least squares (CLS), to multivariate models such as multilinear regression (MLR), principal components analysis (PCA), or the utilization of artificial neural networks (ANN).

One advantage of these mathematical methods is the resolution of the components of a mixture in far less time than would be required for methods involving prior physical or chemical separations. Further, with modern computational systems, the analysis of mixtures using whole-spectrum methods such as factor analysis can be faster than traditional separation techniques and can also be more accurate and precise, provided that the problem of factor optimization is correctly addressed and predictions are validated through error analysis. Factor-based methods such as principal components analysis (PCA) begin with the generation of abstract axes (factors) used to transform the entire data set. Once the data matrix is decomposed factor analysis is used to determine the relative importance of these abstract components in describing the total variance across a population. A plot of the factors which contain the majority of the variance will yield groupings of the sub-populations. In this way, data can be classified according to an inherent property. Principal components regression (PCR) uses inverse least squares regression (ILS) on the principal factors in order to generate prediction models. Similarly, partial least squares regression on latent variables (PLS) can be used to generate a prediction model.

The factor-based techniques are examples of implicit models. Implicit models use empirical models to extract information from a data set in the absence of clearly defined relationships based on chemical or physical concepts. In these implicit models, validation is critical to test the reliability of the predictive model. In contrast, explicit models have clearly defined properties which are directly related to accepted concepts in chemistry or physics, such as the relationship between absorbance and concentration, described by Beer's Law.

In both cases of model development, systematic evaluation of the experimental data is essential and follows each of six stages recommended by Beebe.[2] The first stage is the examination of the data to determine any obvious errors in measurements, then the second stage is pre-processing of the data, if necessary. The third stage involves the generation of the chemometric model, and is method-dependent. After the calibration model is developed, it is validated in stage four. This stage is perhaps the most important, as it involves determining whether the model is reliable for prediction of the properties of the calibration sample set. As the software will always produce a model, the chemometrician must determine if the model is reasonable and useable. Finally, assuming that a reliable model has been defined, it can be applied to the prediction of the properties of a new sample set.

1.2 Types of Implicit and Explicit Models[3]

Numerous procedures are available to relate instrumental measurements to analyte concentrations. These calibration methods develop a model to predict analyte levels. The choice of model is one of the most important decisions the chemometrician must make, as it will define the experiment. Good experimental design requires that all important parameters are considered and any extraneous or misleading information is accounted for. Thus, knowing what kind of data to collect, and how to collect it are key steps in experimental design. For example, a complex model is not required if a simple linear relationship exists between a physical property and the chemical information of interest.

The most common example of calibration in analytical chemistry uses the simple univariate linear model, which includes a calibration step involving instrumental measurements on standard solutions, followed by a prediction step based on instrumental readings obtained with unknown samples. Even this simple model requires time-consuming data acquisition and analysis, and also requires extensive chemical and physical knowledge of the unknown solutions.

Mathematically, the univariate model is defined as

$$y_i = b_1 x_i + e_i \qquad (1)$$

where x and y are the analyte level and the instrumental measurement for the i^{th} specimen, respectively; e is the error associated with the measurement; and b is the predicted model parameter estimated by least squares regression of the instrumental measurements. For simplicity, the intercept term has been omitted.

In the inverse model. the equation becomes

$$x_i = b_2 y_i + e_i \qquad (2)$$

Limitations to both models arise when the sample and its matrix are not clearly defined. In each case. the measured parameter must be highly selective for the analyte of interest and the matrix must be non-interfering otherwise predictions are not considered reliable. Thus. when the sample is more complex. or the matrix is unknown. univariate models may fail.

For multivariate analysis involving several analytes. the calibration step involves multiple instrumental measurements on each of numerous standards. corresponding to the analytes present in the samples. Numerous measurements are made on each sample to generate calibration models which are linear sums of the absorbance of an analyte at several wavelengths. Therefore. for several analytes a series of equations are generated.

As for univariate analysis. a calibration and prediction step is required. For the calibration step. equation 1 becomes a linear sum over multiple instrumental measurements of each absorbing species. generated using independent assays.

For prediction of an analyte the prediction model will take the form of:

$$\hat{x} = b_0 + b_1 y_1 + b_2 y_2 + \ldots + b_n y_n \qquad (3)$$

where the $\hat{x}$ is the estimated concentration of the analyte of interest. y_i is the i^{th} instrumental measurement . and b_i is the i^{th} regression coefficient obtained in the calibration step. whose values differ according to the method used to develop the calibration model (MLR. PCR. or PLS).[3]

For example, the regression coefficients generated for MLR are constrained to be no greater than the number of instrumental measurements in the calibration set. For PCR and PLS, the number of instrumental measurements can be unrestricted, hence the term full-spectrum methods. Conversely, the model reduces to the simple predictive Beer's Law, when only one instrumental parameter is used to predict an analyte, as for CLS methods.

Given a well understood mechanism, CLS provides a reliable and precise method to determine analyte levels. Examination of the scatter of the predicted values from expected values, and the subsequent slope and correlation indicates the predictive capability of the model. If significant deviation is observed, an unaccounted interferent may be present, or an unexplained phenomena is decreasing the reliability of prediction. As well, the presence of outliers will reduce the predictive ability of the model.

MLR, a generalized form of CLS which operates over several instrumental measurements, is useful to develop a model capable of performing over a broader range of conditions. By choosing the optimal instrumental measurements to describe a system, the model can compensate for interferences. The total signal then becomes a linear sum for a specific analyte. However, the number of measurements is limited, as strong correlations between measurements will decrease the stability of the resulting model. To decrease instability of the model, a judicious choice of measurements, and knowledge of the system studied is still required.

MLR does not necessarily require explicit knowledge of all interference levels or matrix effects, since the multiple channels may compensate for any interferences. As these measurements are often correlated with each other, selection of the appropriate measurement channels is important to prevent instability entering into the model. Thus, knowledge about the way measurements interact, and how the various physical variables affect instrumental readings is required to select suitable experimental procedures.

1.3 Factor Analysis - Definitions

Factor analysis has its earliest beginnings in the field of behavioral studies with a paper by Pearson[4] in 1901 on orthogonal regression lines, and expanded on by Hotelling[5] in 1933. As this was the pre-computer era, the underlying assumptions and simplifications made factor analysis unreliable and irreproducible when different methods were used. Thus, it wasn't until the 1970's and later that the factor analysis method became reputable again. Unfortunately, terminology used in factor analysis has also expanded based on different methodologies and practitioners, and is only now slowly becoming standardized. As defined by Malinowski[6] factor analysis is a multivariate technique to reduce matrices to their lowest dimensionality by the use of mutually orthogonal factors and transformations. These factors are formed from a weighted linear combination of the data matrix. Principal components analysis (PCA) is one of the steps in factor analysis, and involves the determination of these orthogonal factors or eigenvectors, and their eigenvalues. The optional precursor step to the transformation procedure involves preprocessing of the data matrix, and a final regression step is used to generate prediction values. In contrast, other authors[7] use the terms factor analysis and PCA interchangeably to refer to the transformation of a data matrix into its respective eigenvalues and eigenvectors, then principal components regression (PCR) for the final regression step. In this case, preprocessing, an optional pretreatment step of the data matrix, would be considered separately.

Factor based methods such as PCR and PLS are among the most commonly used multivariate techniques. In these, a large number of instrumental measurements are used simultaneously, and the resulting model is a linear combination of the weighted instrumental measurements for the sample.

In factor based model development[6,7] the initial steps involve the decomposition of the data matrix into a set of abstract factors, which represent axes in a new coordinate system. For example, consider the entire data set as points in an *xy* plane, with each point due to a voltammogram, or the vector corresponding to a specific sample. A new plane is drawn which will attempt to encompass as many of the points as possible. In this new plane, an axis is drawn which will attempt to capture as much of the variance in the data as possible. This abstract axis is called the first principal component, or eigenvector, and the distance the points are from intersection with this axis defines their variance. The projection the data makes along the axis is defined as the scores, and the sum of the squares of the scores is defined as the eigenvalue for that eigenvector. The eigenvalue is then a measure of the total variance captured by the eigenvector, with the first eigenvalue having the largest value, and the last eigenvalue corresponding to the last eigenvector equal to zero in a noise-free case.

A second axis, orthogonal to the first, is now drawn which will attempt to contain as much of the remaining variance as possible. Subsequent axes are drawn, all mutually orthogonal, until all the variance of the sample population is encompassed. These axes, while generally not having any physical meaning, will contain the variance of the entire populations, will be used in the calibration model, and in subsequent prediction models. The determination of the principal components (PC), or factors, is called principal components analysis (factor analysis) and is a necessary precursor to principal components regression. A common algorithm used to generate the eigenvalues and eigenvectors is singular value decomposition (SVD).

Generally the first principal component contains almost all the variance associated with the total population variance, and is the most important factor. A plot of eigenvalues vs. factor number (a Scree plot) will determine the number of factors required to span most of the variance. Generally , if too few factors are kept, there will be a loss of variance leading to incomplete pattern grouping, as some of the useful signal will be lost. However, retaining too many factors will complicate the projections and incorporate noise. The issue of how many factors to retain is a vital step in the development of the prediction model, and is referred to as rank analysis.

For qualitative assessment, a plot of the scores associated with PC 1 and PC 2 is made, and any clustering of sub-populations observed indicates similarity in some variables between the analytes. This technique is commonly used to differentiate classes of species in electronic noses and electronic tongues, mostly for qualitative analysis for classification.

In PLS[1,6], factors are generated for both the X- and the Y-data matrices, where X-matrix contains the measurements generated, and the Y- matrix usually contains the concentration data. The individual factors are then rotated towards each other angle the angle between them becomes zero, in a noise-free case. The rotation is done in order to restore optimal congruence such that the two planes are now fitted on top of each other, and the corresponding points match. In a noise-free case, the points would be congruent with each other.

Thus, factor analysis or eigenanalysis will reduce a *n*-dimensional data set into *n*-eigenvectors based on a matrix of correlations. Further, eigenanalysis of the data matrix by reduction to a specified lesser number of abstract factors, will allow the resultant data matrix to be more amenable to interpretation, as in pattern recognition techniques.

Principal component analysis is concerned with finding the appropriate number of factors, while manipulation of the factors involves the use of further algorithms such as principal component regression and partial least squares. The ultimate aim of any technique of factor analysis is to define the nature and/or the source of the variance in the original data set.

For voltammetric data, the variance-covariance matrix will be generated from a raw data matrix in which each voltammogram forms a column (also referred to as a vector). Thus the pattern of the voltammogram remains intact. From the computation of the eigenvectors and eigenvalues, we can determine a number of significant underlying factors, called common factors, that will be relevant in the physical interpretation. Other factors, called unique factors, are commonly attributable to other phenomena such as noise. Again, the number of significant factors can be determined from a Scree plot, where the most significant factors will account for the majority of the variance. Each factor will consist of a weighted linear combination of individual current measurements, at potentials corresponding to the highest variance in the current measurements across sample populations.

Each sample is compared across vectors to determine the total variance associated with that sample. By summing the square of each element across the factors we can determine the amount of variance associated with each sample. The results of this manipulation should be identical to the diagonal elements of the variance-covariance matrix of the original data set. By squaring and summing over only two factors to generate a communality value, the result will not be unity, but generally approaches 0.9. That is, the first factor will contain the largest percentage of the variance present in the data, with subsequent factors accounting for correspondingly less variance. The optimal number of factors are chosen, such that those factors accounting for very little variance are discarded, while those containing the signal information are kept. In this way, the dimensionality of the original matrix is reduced.

Once the dimensionality has been reduced, and the number of significant factors has been determined, a variety of analysis and interpretation techniques are available. For quantitative analysis, further regression of the factors will yield a predictive model, while for qualitative assignment of sub-population groups, a scores plot of the scores on the first two factors is generally considered adequate.

1.3.1 Singular Value Decomposition

As mentioned, the procedure for calculating the abstract solution involves an mathematical method known as principal component analysis or principal factor analysis when the singular value decomposition method[6] (SVD) is used.

Factor analysis yields abstract solutions consisting of eigenvectors, with their associated eigenvalues that measure the relative importance of the associated vector. The larger eigenvalues are generally associated with the more significant eigenvectors. The factors can be a maximum of *s*-eigenvectors where *s* refers either to the number of rows, *r*, or the number of columns, *c*, whichever is smaller. The principal factor solution will be represented as the decomposition of the data matrix, **D**, into an abstract row matrix, **R**, and abstract column matrix, **C**, as:

$$\underset{r\times c}{\mathbf{D}} = \underset{r\times s}{\mathbf{R}}\,\underset{s\times c}{\mathbf{C}} \tag{4}$$

In SVD, the equation has the form

$$\underset{r\times c}{\mathbf{D}} = \underset{r\times s}{\mathbf{U}}\,\underset{s\times s}{\mathbf{S}}\,\underset{r\times s}{\mathbf{V'}} \tag{5}$$

where, **S** is a diagonal matrix whose elements are the square roots of the eigenvalues and,

$$\begin{aligned} \mathbf{R} &= \mathbf{US} \\ \mathbf{V}' &= \mathbf{C} \end{aligned} \tag{6}$$

Each column of **U** is an abstract orthonormal eigenvector spanning row space and each column of **V** is an abstract orthonormal eigenvector that spans column space, with both sets of eigenvectors lying within the *s*-dimensional space defined by *s*-factors. **V'** is defined as the transpose of **V**. Each *j*th eigenvector of **U** and **V** share the same *j*th eigenvalue, which represents the portion of the total variance in the data at that point. Factors are then successively ranked in decreasing order according to the amount of variance they represent, with the first factor containing the most variance of the data, up to the *s*-factor with the smallest eigenvalue, which is considered the least important.

Now, the number of the factors which contain significant information can be determined by sub-dividing the abstract factors into two sets, a primary set containing *n*-factors of the real, measurable features of the data and a second set, the null set, which will be predominately associated with the experimental error. This will allow the elimination of secondary factors, and compress the model into *n*-factors which are considered to be physically significant, with reduced error. The equation is then re-written as the properly dimensioned abstract solution when the error factors are deleted, and subsequent analyses are based on the use of this solution:

$$\underset{r\times c}{\bar{\mathbf{D}}} = \underset{r\times n}{\bar{\mathbf{R}}}\,\underset{n\times c}{\bar{\mathbf{C}}} = \underset{r\times n}{\bar{\mathbf{U}}}\,\underset{s\times n}{\bar{\mathbf{S}}}\,\underset{n\times c}{\bar{\mathbf{V}}'} \qquad (7)$$

To deduce the correct number of factors a stepwise procedure is used where each step is computed, then compared

$$\underset{r\times j}{\bar{\mathbf{R}}}\,\underset{j\times c}{\bar{\mathbf{C}}} = \underset{r\times c}{\bar{\mathbf{D}}} \overset{?}{=} \underset{r\times c}{\mathbf{D}} \qquad (8)$$

where **R** and **C** are the abstract matrices based on the *j*-factors, $\bar{\mathbf{D}}$ is the data matrix reproduction, which is compared to the original data matrix. As additional factors are employed from the first most important factor, through the *j*th factor, to *s*-factors, data reproduction becomes more accurate as greater variation is accounted for until the correct number of factors (*j*=*n*) is determined, at which point the reproduced data matrix is equal to the original data matrix. From this determination of the number of factors which describe the data set further analysis can be done utilizing this solution.

1.3.2 Rank Analysis

An important step in generating a useful calibration model is the determination of the correct number of significant factors. Keeping the optimal number of factors is important as too few factors will not adequately describe the data (thus our calibration model will not be valid), while keeping too many factors will retain noise and introduce artifacts into the model. To decide the number of factors which are to be kept, a number of tools are available, such as the use of indicator functions, cross-validation and, especially, predicted residual error sum of squares (PRESS).

Indicator functions are based upon analysis of the eigenvalues and their errors and commonly use the method of Malinowski,[6] as in the F-test on reduced eigenvalues. Eigenvalues associated with each eigenvector are equal to the amount of variance in the data that is captured by that eigenvector, and typically decline rapidly with succeeding eigenvectors over several orders of magnitude. The eigenvalues typically associated with the error eigenvalues span little variance and should be statistically equal. Given that the variance will measure the importance of the eigenvalue, it can be defined in terms of the eigenvalue, λ_j as

$$\text{variance} = \frac{\lambda_j}{\sum_{j=1}^{c} \lambda_j} \qquad (9)$$

In order to determine at which factor, the variance will be statistically equal, a Scree plot can be drawn of the residual percent variance accounted for by a given factor. The residual percent variance can be determined from the eigenvalues, as

$$\text{residual\%variance} = 100\left(\frac{\sum_{j=n+1}^{c} \lambda_j^0}{\sum_{j=1}^{c} \lambda_j}\right) \qquad (10)$$

where is the root mean square error associated with the difference between the raw data (d_{ik}) and the reproduced data (d_{ik}^*) at the *j*-factor, with *n*-primary eigenvalues, and *k*-columns and *i*-rows, as

$$\sum_{i=1}^{r} \sum_{k=1}^{c} (d_{ik}^2 - d_{ik}^{*2}) = \sum_{j=n+1}^{c} \lambda_j^0 \qquad (11)$$

A plot of the variance associated with either eigenvalues or reduced eigenvalues against the number of factors, should show a minimum, or a levelling off at the optimum rank. This point will distinguish between factors, which contain the majority of the variance and those factors containing mostly noise. Ideally, the point should be sharp, and level off to a constant level. Often however, non-ideal behaviour is observed as the reduced eigenvalues begin to increase. This sort of behaviour will complicate the decision-making process when the number of factors to keep is determined. For this reason other indicator functions should be used to gain an overall understanding of the optimal factors.

The reduced eigenvalue is defined as an eigenvalue corrected by appropriate degrees of freedom as shown

$$\mathbf{REV}_j = \lambda_j/(r-j+1)(c-j+1) \qquad (12)$$

As the reduced eigenvalues associated with the error are proportional to the standard deviation, an F-test can be used. Using a standard table, and beginning with the smallest eigenvalue, the significance level is compared to the reduced eigenvalue, until the desired significance level (5%) is obtained. Moving from the smallest eigenvalue up the table, each successive variance is added to the pool of eigenvalues, and this is compared to a standard table. The eigenvalue corresponding to the number of factors at which the pool of eigenvalues is larger than the standard table, corresponds to the optimal number of factors to be retained. The F-values correspond to the ratios of two variances obtained from independent sample pools with normal distributions, and is useful as a test statistic in comparing variances from normally distributed error eigenvalues.

A more rigorous method, to determine if the optimal number of factors has been chosen, requires that regression of the factors has been performed to generate predicted concentrations using the calibration model. The calibration model is then used to predict concentrations given an independent validation data set.[7] Prediction values are generated for every possible number of factors and the predicted residual error sum of squares (PRESS) is calculated, with the optimal number of factors corresponding to the lowest PRESS. Consequently, a plot of the PRESS values at the given factor should show a sharp decrease at the optimal factor, then a continuing decrease in error to zero.

$$\text{PRESS} = \sum (C_{predicted} - C_{expected})^2 \qquad (13)$$

If there is not an independent validation set available to calculate the PRESS, cross-validation can be performed on the original training set to simulate a validation set. Cross-validation involves an iterative leave-one-out (or leave *n*-out) procedure to predict the concentrations of the components left in the training set and a subsequent sum-squared of errors between expected and predicted values. Calibrations are generated for every possible number of factors and the resulting PRESS values are plotted against the number of factors, with the optimal number corresponding to the smallest PRESS. Often a plot of the cross-validation results shows a sharp decrease at the optimum factor, then as the number of factors increases, a rise in error is observed, as the factors containing mostly error are included.

After the optimal number of factors are determined it is expected the noise eigenvectors have been discarded. Therefore the data can be regenerated in the new coordinate system. To ensure noise has been discarded and meaningful information kept a plot of the residuals should indicate only pure random noise at the optimal factor level and no spectral or voltammetric features. Residuals may sometimes contain spectral or voltammetric features if there is some feature incorporated into the experiment, which adds an extra mode of variation into the voltammogram. This includes variation due to instrument drift, interfering compounds or an unexpected experimental feature. This could then be the trigger to investigate the sample for any of these anomalies that do not fit the sample into the calibration model and hence affect the predictive ability.

1.3.3 Principal Components Regression

Having generated the properly dimensioned abstract solution

$$\underset{r\times c}{\bar{\mathbf{D}}} = \underset{r\times n}{\bar{\mathbf{R}}}\ \underset{n\times c}{\bar{\mathbf{C}}} = \underset{r\times n}{\bar{\mathbf{U}}}\ \underset{s\times n}{\bar{\mathbf{S}}}\ \underset{n\times c}{\bar{\mathbf{V}}'} \tag{14}$$

a transformation step can be performed in order to generate real solutions from the abstract model as

$$\overline{\mathbf{D}} = \overline{\mathbf{RC}} = \{\overline{\mathbf{R}}\mathbf{T}\}\{\mathbf{T}^{-1}\overline{\mathbf{C}}\} = \hat{\mathbf{X}}\hat{\mathbf{Y}} \tag{15}$$

where the transformation matrix **T**, and its inverse $\mathbf{T}^{-1}$ are applied to generate the real solution. The transformation matrix is obtained using a least-squares method in order to match the predicted **X** which most closely matches the target **X**. Mathematically, the best transformation matrix is obtained when the deviation between the predicted and the test **X** are minimized by setting the sum of the derivatives of the squares of the differences equal to zero.

The regenerated (predicted) data set calculated from the calibration model is then compared to the original data set in order to assess the reliability. Further, having calculated the appropriate transformation matrix (comparable to a regression matrix) unknown sample concentrations can be predicted using their respective voltammograms.

1.3.4 Partial Least Squares

Partial least squares regression is a logical continuation of PCR in that abstract factors are generated for both the **X** and the **Y**-matrices which contain the concentration and the voltammetric data sets. As mentioned, noise will deflect the eigenvectors out of the theoretical plane, and as the concentration and voltammetric data sets both independently contain noise, the eigenvectors associated with these data sets are deflected in random directions. PLS compares the voltammetric vectors with the concentration vectors, assesses the angle between them and, because this angle is due to the differences in noise between the two vector, rotates the vectors back towards each other to restore optimal congruence. As the **X** and **Y**-data planes are congruent with each other this will maximize the fit of the linear regression between the projections of the **X**-factors and the projections of the **Y**-factors. As these projections (scores) of the respective data sets are directly proportional, the calibration solution will be of the form

$$\underset{nxp}{\mathbf{Y}} = \underset{nxm}{\mathbf{X}}\ \underset{mxp}{\mathbf{B}} \qquad (16)$$

where **X** is the concentration factors, **Y** contains the voltammetric factors, **B** is the matrix of proportionality constants (calibration constants) used to give the calibration matrix as

$$\mathbf{B} = \mathbf{P(P'P)^{-1}WQ'} \qquad \text{and} \qquad (17)$$

$$\hat{\mathbf{X}} = \mathbf{TB} \qquad (18)$$

Again, **T** is the transformation matrix used on **B** to generate the prediction values. In the case of PLS, the individual relationships for the matrices **X** and **Y** are calculated, and also an inner relationship is determined. This gives three relevant equations as

$$\begin{aligned}\underset{n\times m}{\mathbf{X}} &= \underset{n\times a}{\mathbf{T}}\,\underset{a\times m}{\mathbf{P}'} + \underset{n\times m}{\mathbf{E}}\\ \underset{n\times p}{\mathbf{Y}} &= \underset{n\times a}{\mathbf{U}}\,\underset{a\times p}{\mathbf{Q}'} + \underset{n\times p}{\mathbf{F}}\\ \underset{n\times a}{\mathbf{U}} &= \underset{n\times a}{\mathbf{T}}\,\underset{a\times a}{\mathbf{W}}\end{aligned} \qquad (19\text{-}21)$$

where **E** and **F** are the error matrices which are minimized by least squares, **W** contains the loading weights that span the variance, **Q** and **P** contain the respective factors or the projections of each w_i on either the voltammetric or the concentration plane, and **U** and **T** are used to hold the extracted eigenvectors until PRESS is minimized.

In PLS successive eigenvectors will be defined by (w_i, t_i, q_i, u_i, p_i), and the basis vectors are chosen so as to optimize the linear regression between both spectral and instrumental factors. As the fit between **X** and **Y** may be somewhat compromised in order to improve the regression between the data sets, the residuals of PLS tend to be greater in magnitude. Also residuals may contain some spectra-like features, as if there is any non-linearity in the spectra these non-linearities tend to be rejected into the later factors. Some of this non-linearity is now spanned by noise factors and as it tends to be in regions of strong spectral activity, some of the spectral features are observed in the residuals.

Also, eigenvalues are not calculated in PLS, but pseudo-eigenvalues may be generated in order to evaluate the amount of variance in the data that the factor models, in order to determine optimal number of factors to be used in developing the calibration solution.

1.4 Data Treatment

1.4.1 Preprocessing

Preprocessing is defined as any mathematical manipulation of the data prior to analysis.[8] It is useful to remove any irrelevant sources of variation, but it will consequently change any data given to the analysis model. There are many methods available for preprocessing of the sample data.

Mean centering is a tool applied to account for an offset in the data, and is used to preprocess row or column variables. The mean of the value of the variable is subtracted from each of the elements of the vector in order to mean-center the resulting values about zero. This has the effect of repositioning the centroid of the data to the origin of the coordinate system, and preventing points at edges from having more influence than other points. The disadvantage of using this method is that information is lost about eigenvalue magnitudes and relative error. This method is one of the most useful techniques available for factor analysis.

One of the more common preprocessing techniques is normalization of the sample in order to perform qualitative identification. All samples are put on the same scale by dividing each instrumental reading (such as the absorbance) by the sum of the squares of the absorbance readings of the entire sample. This will equalize the magnitude of each sample. While the magnitude of the distance of the data point from the origin is removed, the direction is preserved, making this useful in the preparation of reference spectra for a qualitative identification library.

Scaling or standardization of variables can be performed over either the rows or the columns of the data set so that the variation in all channels is weighted equally. Typically this is done by either dividing by a constant so that the maximum intensity is equal (range scaling), or dividing by the standard deviation in variance scaling. This will alter the weighing of the variable, making low intensity peaks more significant and high intensity peaks less. While this is useful when the dynamic range of the variables changes, but relative noise stays the same, it tends to give more value to noise, at the risk of loss of signal. For example, in variance scaling, the influence of variables where signal variation is large is reduced, while the influence where the signal variation is small tends to be increased. This can be detrimental to precision or robustness.

Another method used to preprocess row variables is variable weighting, which involves multiplying all elements in a vector by some weight. The weight is chosen either because of prior knowledge or by variable selection, where certain variables are multiplied by zero. Knowledge of the chemistry of the system and an experienced operator are necessary to use this method as those variables believed to contain the significant information can be scaled up in importance, at the risk of losing information present in variables which have been given lower weight.

It should be noted that any change to the data matrix made by using these preprocessing techniques will result in subsequent changes in the eigenvalues. Since eigenvalues are the sum of the squares of the scores, changing these will result in different roots of the simultaneous equations and change the lengths of the axis used to define the principal components. The resultant information may differ significantly from that obtained from data not preprocessed. Therefore, any data pretreatment should be based on knowledge of the resultant effects pretreatment will have on the final outcome, and a sound basis of the statistical criteria.

1.4.2 Errors

In optimizing the number of factors used in the calibration model, the chemometrician is interested in the number of factors that best model the data. Given that no data are error-free, the resulting eigenvalues produced from the factor analysis of experimental data will be larger than if the theoretical error-free data was used.[6] In effect the raw data matrix, **D** is the sum of the pure error-free data matrix, **D*** and the error matrix, **E**.

$$\mathbf{D} = \mathbf{D}^* + \mathbf{E} \qquad (22)$$

Retention of all the factors, while perfectly reproducing the raw data, will include the uncertainty into the model. Choosing the correct number of factors to include in the calibration model will decrease the prediction errors, but will not completely eliminate them, as it is impossible to remove all error.

The main factor analysis error terms used are extracted error (XE), real error (RE), and imbedded error (IE).[7] Only extracted error is experimentally measurable, but the errors can be expressed in a mutually dependant relationship as

$$(RE)^2 = (IE)^2 + (XE)^2 \qquad (23)$$

where the real error is defined by Malinowski[6] as the residual standard deviation (RSD), and referred to as total error by Kramer.[7] Given that a relationship exists between the different errors, it is possible to calculate RE and IE from XE as

$$\mathrm{RE} = \sqrt{c/(c-p)}\,\mathrm{XE} \tag{24}$$

$$\mathrm{IE} = \sqrt{p/(c-p)}\,\mathrm{XE} \tag{25}$$

where there are p principal components chosen to model the data set and c variables are measured.[6,8] The XE is given by

$$\mathrm{XE} = \mathrm{RSD}\sqrt{\frac{c-p}{c}} \tag{26}$$

Imbedded error is a measure of the difference between the pure data and the reproduced data, while real error is a measure of the difference between the pure data and the experimental data. Given that factor analysis discards the noise factors, the IE will normally be less than the RE.

The extracted error is identical to the root-mean-square (RMS) error[6] of prediction and is one of the indicator functions used to determine the optimal number of factors. Other indicator functions based on eigenvalues include analysis of the imbedded error, which will decrease until the optimal number of factors, then increase again as the noise factors are fitted back into the model.

Root mean square error (RMSE) and residual standard deviation (RSD) are closely related, as the RMSE calculates the difference between the raw data and the factor-regenerated data, while RSD measures the difference between the raw data and data containing no experimental error. Hence RMSE tends to be less that RSD, and its use as an indicator function is not recommended.

$$RMS = \left(\frac{c-p}{c}\right)^{1/2} (RSD) \tag{27}$$

$$RMSEP = \sqrt{\frac{\sum (y-\hat{y})^2}{N_{pred}}} \tag{28}$$

RMSE is commonly used in the regeneration step as a measure of reliability of the calibration solution. As RMSE is measured as the difference between the raw data and the factor regenerated data, the optimal choice of eigenvectors has a direct effect on this error. For example, using an excessive number of eigenvectors will minimize the RMSE, but will also minimize the extracted error. Conversely, deleting an excessive amount of eigenvectors will remove significant information from the model, and incorrectly describe the important variables. When this occurs, the RMSE tends to be too large. To determine the RMSE associated with regenerated and predicted data, and to assess the validity of the resultant calibration solution, the RMSEC (root mean square error of calibration), and the RMSEP (root mean square error of prediction) are calculated.

$$RMSEC = \sqrt{\frac{\sum (y-\hat{y})^2}{N_{cal} - p}} \tag{29}$$

where the sum of the squares of the differences between actual and predicted calibration or prediction values are divided by either the difference between the number of calibration samples (N_{cal}) and the number of factors used (RMSEC), or by the number of prediction samples (N_{pred}) in RMSEP. After regeneration is complete, prediction of validation samples is performed using the calibration solution, and assessing the validity of the model using RMSE is done.

Regeneration is the process where the portion of the variance that displaces the data out of the ideal noise-free plane is discarded. Since noise is assumed to be isotropic, the direction each point is displaced is completely random. Thus eigenvectors are slightly displaced from the noise-free case, and the eigenvalues associated with the eigenvectors do not decrease to zero for the last eigenvalue. Therefore by examination of the eigenvalues, the error eigenvectors can be discarded. A plot of the residuals associated with retention of a given number of eigenvectors should show only noise, when the optimal number of factors has been chosen. Any remaining noise is spanned by the retained eigenvectors, and can be described by the imbedded error, which will be equal to the real error when the residuals are discarded.

1.5 Sensors and Multivariate Analysis

Initial development of chemical sensors focussed on sensors selective to a particular chemical entity or property, with the goal of specificity. Signal analysis is relatively simple with these selective sensors and often linear or Nernstian relationships exist between the chemical concentration and the instrumental measurement. For example, pH electrodes utilize a conversion of emf to a pH scale, and an ion selective electrode is selective for a certain ion in solution. For these cases, matrix effects and interferences must be known and compensated for, otherwise inaccurate readings are obtained.

In recent years (1980 and onward) non-selective sensor arrays combined with mathematical manipulations have come into use for the analysis of more complex liquids and gases. Previously, many of these same liquids and gases could not be measured accurately due to a lack of selectivity or sensitivity of available sensors. These non-selective sensors and sensor arrays use pattern recognition and multivariate calibration techniques to analyse either complex gas media as electronic noses or complex liquid media as electronic tongues, for identification or for quantitation of components.

1.5.1 Voltammetry

Previous work in sensor arrays, coupled with an appropriate data analysis technique has focused on semiconductor arrays, or ion selective electrodes as the signal generating device. Voltammetry offers the advantage of high sensitivity, and versatility as the potential range, electrode material and waveform can be chosen to optimize experimental conditions. Square wave voltammetry, a small amplitude controlled potential technique, provides several advantages for sensitive and rapid detection. Small amplitude techniques tend to be more sensitive and precise, as the signal can be more easily distinguished from the background, charging currents are minimized and steady-state measurements can be made.[9] In square wave voltammetry a small amplitude pulse (square wave) of frequency (f) and amplitude dE, is superimposed onto a constant dc-potential ramp of increasing potential, E. The large amplitude potential sets the surface concentration at the electrode. and the small amplitude excitation periodically perturbs the surface concentration. The instantaneous faradaic current is proportional to the surface concentration of redox active species. The base potential increases by dE for each full cycle of the square wave. The current is measured at end of each half-cycle, and the difference between the forward and the reverse current is the current density plotted on a voltammogram with respect to the potential. The increment in square wave voltammetry sensitivity is due to the measurement at each half-cycle, to give the difference between forward and reverse currents.

Typically, the current is recorded for redox active species in solution, but any changes on the electrode surface will also result in a change in current. As the change in current in a diffusion controlled experiment tends to be proportional to the concentration of bulk species, any species which tend to adsorb onto an electrode will also result in a current change. The more common electro-active species are those containing nitro, thiol, carbonyl and double bonded functional groups which can undergo reduction.[10] For the complex liquids used in this study, common groups present such as sugars and other carbohydrates[11], ethanol[12], ascorbates[13] or citrates and molecular oxygen may be present in the matrix, and can give a signal. As well, adsorption of redox inactive species will also affect the voltammetric signal.

In a related study,[12] an amperometric technique was used to determine levels of ethanol, and common sugars at an platinum electrode. The authors found a combination of peaks, with ethanol response at -0.32 V, and sugars giving two peaks at -0.70 V and -0.23 V. The peak centered at -0.70 was larger than the other peaks observed. Further, for the electroanalysis of ascorbic acid using cyclic voltammetry on a Pt electrode, a peak centered at 0.30 V was observed, which tended to decrease in peak current due to the slow desorption of the redox products from the metal surface.[14] This parallels the results obtained using square wave voltammetry for this study, where two peaks, a large peak at -0.60 V, and smaller peak at 0.30 V were observed. Given the different conditions used in these studies, and the different methods, it is difficult to form a conclusion as to the electrochemical process occurring on the voltammograms, but it is believed, that a combination of ethanol and sugar reduction, combined with surface adsorption of species is responsible for the peaks generated.

The application of a combination square wave voltammetry and PCA, was the first step toward development of a multi-array system for the detection, and quantitation of species present in a complex matrix. This first step towards an electronic tongue, after determination of the optimal conditions and limitations, could be used for in a variety of media, without resorting to any preliminary separation of compounds present in the matrix.

1.5.2 Electronic Noses

In mammals the ability to discriminate odours is a property of the olfactory system as a whole. That is, both primary and secondary neurons work together to process a signal from non-specific sensors.[15] The process typically encodes a pattern of signals corresponding to a particular odour. Pattern classification techniques should thus be able to simulate this process using mathematical methods to recognize a pattern corresponding to a particular odour. Mathematical manipulation of the complex signals and pattern classification will maximize the differences between the individual components of the signal for different samples and allow discrimination of specific odour constituents. In this sense, the electronic nose will be a mimic of the neural process involved in the biological sense of smell.

Interest in development of an electronic nose arose due to the relatively expensive or inefficient means used to identify odours. For example, trained experts are required as primary evaluators of perfumes, wines or foodstuffs. These experts (termed noses) are useful only for short periods of time due to saturation of the human olfactory system, and are expensive due to their rarity. Conventional analytical instruments such as gas chromatography and GC-mass spectrometry can be used to sample odours, but these techniques tend to be expensive and can suffer from a poor level of detection depending on the type of odor or the matrix constituents. Hence, there is a demand for a low-cost, rapid and portable system for odor detection and identification of flavor constituents.

The term electronic nose first appeared in the 1980's and was used initially at the 8th International Congress of European Chemoreception Research Organization (1987).[16] The first conference dedicated to electronic noses was held in 1990 as part of the NATO Advanced Research Workshops in Iceland.[17] Due to its relatively recent introduction, a comprehensive definition for the term "electronic nose" has not been clearly stated but for the most part can be defined as an array of electronic sensors of partial or low selectivity which, when coupled to an appropriate pattern recognition system, is capable of recognizing and quantifying simple and complex odours.[18] The term electronic nose was applied to this type of instrument due to its ability to mimic the olfactory system, which typically consists of cells of low selectivity and uses neural processing to increase the sensitivity by several orders of magnitude, allowing discrimination between several thousands of odours. The concept of an electronic nose as a combination of a sensor array and an intelligent data processing device for classification began with the work of Persaud and Dodd[18] in 1982. The limiting step in development of these systems has been data analysis. Research in electronic noses has really parallelled the development of computer systems, and, due to the need for partial selectivity, development in semiconductor sensor technology.

An electronic nose requires an array of partially selective sensors that respond to a broad range of gases, so a large body of work in this area has been done using semiconductor materials. These semiconductors range from inorganic oxides and catalytic materials to sensors utilizing integrated thin films,[19] oxides, and conducting polymeric materials. For example, a multisensor array of conducting polymers was used to detect diacetyl in beer, a component associated with off-taste.[20] Data analysis of the odorant response has also ranged from unsupervised pattern recognition to classify the odours in fuel cells,[21] to more advanced methods using supervised learning artificial neural networks.[22] In most cases the objective has been to correlate results from the sensor array to expected responses of the human olfactory system.

Persaud and Dodd[18] constructed an artificial olfactory system to model a mammalian nose using commercially available semiconductor gas sensors to test responses to a wide variety of odours. The voltage changes with odorant concentration were monitored over three gas sensors, and ratios of the responses of the three sensors comprised the data set. A comparison was made between the artificial nose and the sheep olfactory system, and it was determined that the electronic nose was able to mimic the mammalian system at a superficial level.

Tin-oxide semiconductors are commonly used in gas sensor arrays, as improvements to selectivity of multicomponent analysis and drift compensation are possible when combined with an appropriate pattern recognition system.[23] Examples of the use of semiconductor systems for analysis of odorants include the separation and classification of gas samples and flavour samples using a four-sensor array of doped thin-film-silicon-based micro gas sensors.[24] Sensitivity of the sensor to the gas was monitored as the ratio of the responses of the sensors to the odorant to the response in air, and principal component analysis and neural network analysis were used for pattern recognition. A pattern recognition plot generated using an ANN gave a recognition probability of 100% over varying concentrations of 12 gas samples and 93% recognition probability when was used to classify 6 flavour samples into their respective clusters using PCA.

An array of five semiconductor gas sensors coated with different materials was used to classify vintage years of wine[25] using PCA to generate a pattern recognition plot. By selecting the array components it was possible to completely differentiate the vintages of the wines, and to identify the years in which a barricatura process was used to age the wine, as these years formed a unique sub-population group.

Organic electrically conducting polymers such as polypyrrole and polythiophene were used as sensor coatings in semiconductor arrays to measure the response to pig malodours and food odours[26] which are complex gas mixtures. Pattern recognition by principal components analysis and artificial neural networks correlated fingerprints of the odours, and were used to classify specific chemicals responsible for the odour. Different individual chemicals were also quantified with correlation coefficients of 0.89 to 0.96. In classifying a particular odour, electronic noses are not concerned primarily with all the underlying constituents of that odour, but rather with detecting the global effect of all the chemical species present to characterize the odour.[27] Therefore the cross-sensitivities of the array components are important in determining an overall pattern of the particular odour, and ultimately classifying it in a manner similar to the human olfactory system.

Presently, artificial neural networks are commonly used for data analysis in the electronic nose, due particularly to their inherent similarity to the human neural system. Neural networks are not always necessary, however. If classification of odours is desired, a linear pattern recognition technique such as PCA is sufficient for data analysis and often considered more robust than neural networks.[28] For quantification of odour constituents, neural networks offer advantages due to the non-linearity of sensor responses to some odour constituents, which are better fitted by a non-linear mathematical model. However neural networks require a large training data set which increases with the complexity of the array and the sample complexity. An alternative to this is a self-organized map (SOM) which requires less computational complexity while still modelling the biological process of learning and associative classification. A hybrid neural network has been reported that used a generalized perceptron network trained by a back-propagation algorithm and a SOM for the recognition of patterns for binary mixtures.[29] A sensor array of six quartz-crystal microbalances (QCM) coated with different polymers monitored the frequency response of 3 classes of binary gas mixtures with varying gas concentrations and passed the sensor outputs to a neural network. Of the three classes of gas samples, one class was correctly identified, while the other two classes had an identification rate of more than 80%.

QCM sensors are versatile for the identification of gas stream components, with sensitivity of analysis depending on the coating of the QCM. For example, an array of calixarene-coated QCM's combined with pattern recognition methods from PCR and an artificial neural network (ANN) distinguished between pentane, methanol, hexane and chloroform present in a gas stream.[30] Further, an epoxy-coated QCM was used as the sensing device for wine recognition[31] which, combined with a principal component plot, was able to correctly (100%) separate out white, red and rose wines into their respective sub-population groupings

Additional work on electronic noses has focused on the nonlinear relationship of the regression technique to the sensor array and the use of mathematical techniques applied to the neural network in order to either improve the prediction of the results or to reduce the size of the required training set. One study has focused on the use of non-parametric techniques[32] (smoothing methodology) to give an approximate relationship between the data from the sensor array and the calculated results. Conductance measurements for mixtures of acetone and methanol vapours on an interdigitated sensor were subjected to a reverse calibration where the weights remained constant but the values entering the neural net were adapted until stable output values were achieved. It was determined that the results obtained using this method were closer to expected values than using conventional neural network analysis, in which the weights are varied and inputs remain constant.

Further work has been done on a sparse data set used for calibration, with different computational methods to determine which is the most efficient and robust algorithm when the size of the calibration set is minimized.[33] Seven methods, two linear and five non-linear, were evaluated for their predictive ability with mixtures of octane and toluene vapours tested by bulk-acoustic wave quartz crystal non-linear gas sensors coated with six different polysiloxanes. Based on root mean square errors of prediction, artificial neural networks and least squares estimation showed similar performance with a RMS error of 0.3 to 0.7% for the least squares method and 0.4 to 0.8% for the ANN.

A more theoretical approach has been to design a sensor array consisting of a multiplicity of semiconducting oxide electrodes with differing spacing in relationship to their varying thickness of sensing material.[34] Although not tested experimentally, the authors speculate that, based on reaction diffusion effects, the response of the transistor should allow gases in a mixture to be distinguished based on the differing reactions of the gases and this effect on their diffusion rates.

More practically, neural network analysis was applied to predict gas levels using responses generated by surface acoustic wave sensors. In one study[35] the detection of differing NO_2 concentrations by frequency shifts at metallophthalocyanine coated sensors gave prediction results of 98.9% on the training set and 82.8% on the prediction set. Alternatively an attempt was made to mimic the mammalian olfactory system by coating an array of twelve thickness shear mode acoustic wave sensors with various adsorptive materials. Frequency responses were measured for various organic compounds such as esters, ketones and aldehydes in a continuing study to distinguish aromas from various fruits and essential oils.[36] Using pattern recognition techniques, organic classes were separated into their respective clusters, and the different essential oils and fruit aromas had good separation into their respective groups without overlap of sub-population groupings.

Identification of groups is not restricted to foods, as these same techniques can be used to identify functional groups based on correlation to known spectral libraries. For example, one study[37] that used near infrared gas phase analysis of 40 samples combined with PCA was correctly able to classify 95% of the samples into aromatic and non-aromatic groups; the PCA scores were then fed as inputs into a neural network, and used to identify the functional groups present resulting in 95% correct identification for aromatics, 98% for hydroxy and halogens and 98% for carbonyl groups.

Another study[38] measured the fluorescence emission spectra ofNile Red immobilized in a polymer array in the presence of nine organic vapors (amyl alcohol, amyl acetate, butanol, butyl acetate, pentanol, pentyl acetate, benzene, toluene, xylene) and used neural network analysis to generate the fluorescent fingerprint of each species resulting in correct prediction of 99.5% in the training set and 90% in the prediction set.

A more complex study involved the use of fiber optic chemosensors to measure organic odours.[39] Both fluorescence amplitude and temporal variation were measured in order to reproduce actual olfactory responses which vary with time. Neural network analysis of the fluorescence signals from six fiber optic sensors correctly identified 71% of the test patterns, and was considered a first approximation of an artificial olfactory system to a biological model as it assumed a Gaussian shape for both the temporal variation and the amplitude of the spikes due to varying concentrations of the organic species.

1.5.3 Electronic Tongues

An electronic tongue is commonly defined as a non-selective multisensor array for the detection and quantitation of components liquids, combined with appropriate mathematical methods to process the signals.[40] Common uses of the electronic tongue have been in the areas of food, environmental and process analysis. For food analysis, some examples involved the classification of coffee from its infra-red spectrum,[41] to monitoring the process at a sugar plant for quality control.[42] Qualitative differentiation of beverages[43]into respective sub-populations was reported as well as the quantitation of adulteration of orange juice samples. Further, quantitative analysis of fungicide treatments in the resulting wines was also reported.[44]

For process control, sensors coupled with appropriate multivariate analysis have been used from on-line monitoring of plastic waste,[45] to monitoring of stack emissions.[46] Numerous review of the industrial use of sensor arrays and chemometric techniques for analyte monitoring[47] and process control[48] are also available.

The first work on multisensor arrays combined with regression analysis using partial least squares was the work by Otto and Thomas[49] on the simultaneous analysis of calcium, magnesium, potassium and sodium ions at typical physiological concentrations. Different concentrations and combinations of the ions were examined using four ion selective electrodes (ISE), and prediction results using classical least squares and PLS were compared.

The ISEs used were commercially available sodium, potassium and calcium electrodes. Calcium and magnesium were also measured with an ISE made of calcium bis{di[4-(1,1,3,3-tetramethylbutyl)phenyl]phosphate} immobilized in polyvinylchloride and either di-n-octylphenylphosphonate (DOPP) for the calcium determination or decanol/DOPP for magnesium determination. Prediction errors were only slightly lower for PLS over CLS when synthetic intracellular fluids were measured, however when body fluid concentrations of the ions were measured using an over-determined system of 5 or 8 sensors in the array, the prediction errors using PLS dropped significantly from CLS.

This work was expanded by Beebe et. al. [50] to compare predictions of sodium and potassium levels from ion selective electrodes using nonlinear projection pursuit regression and the partial least squares technique of Otto and Thomas. Prediction errors using both methods were comparable, leaving the authors to conclude that a future existed for the use of non-selective sensors coupled with an appropriate analysis method. Further work used artificial neural networks to process the data from an array of ion selective electrodes for the determination of calcium and copper ions in mixtures and the simultaneous determination of potassium, calcium, nitrate and chloride ions.[51] Using a combination of a glass pH electrode with ISEs, simultaneous determination of ions was possible with a mean prediction error of 8% for a binary mixture and 6% for the quaternary mixture. Unfortunately due to the slow speed of computers in 1990, it took between 24 and 48 hours for training.

The rapid evolution of computer technology in the early 1990's resulted in rapid growth in computer processing speeds, and subsequently caused an increase in the use of chemometric techniques for the analysis of various liquid systems. Neural network analysis was applied to partially selective ISE arrays for the determination of sodium, potassium and calcium ions in a model flow injection system[52] where the addition of noise, baseline shift and peak height reduction was successfully modelled for 44 out of 56 possible ion combinations with an error less than 10% for prediction of single ion levels in the various mixtures.

Heavy metal and inorganic ions have been quantitated in multicomponent mixtures using an ISE array consisting of chalcogenide glass membrane electrodes[53] with various data processing methods (MLR, PLS, and ANN) to determine the most appropriate method. The array contained partially selective electrodes, and it was assumed that interference occurred. Therefore, MLR was initially used to estimate the degree of non-linearity of the sensor response and PLS was used for the quantitative processing of data. For data exhibiting extensive nonlinearity, ANN was found to be the method most likely to correctly fit response and quantitative information together. The series of steps included pre-processing of the data response matrix through pattern recognition by separation into separate classes, then finally identification and calibration of the data to the unknown composition. For all three methods (MLR, PLS, and ANN), the relative errors of prediction were used to determine the optimal method. MLR results were the poorest, while the smallest errors were observed when ANN was used. As ANN allows for non-linearity in the prediction of data this is not a surprising outcome, as significant non-linearity was observed due to interfering species.

The work in this paper was expanded on to produce an electronic tongue, where the sensor array was used to quantify ions in polluted river water.[54] An array of 22 electrodes consisting of chalcogenide glass doped electrodes and conventional ISEs was used and a data set of 150 solutions of varying ion levels was split into two for use as the calibration and the prediction sets. As in the previous study, MLR, PLS, ANN were the methods used for data analysis along with nonlinear least squares (NLL). Again MLR was used to test for departures from linearity, as errors in this method arise from nonlinearity. PLS also requires linear relationships but prediction errors were smaller using this analysis method. NLL allows for the use of non-linear data, so errors can be minimized to about the level of PLS. ANN results indicated slightly lower error in predicting concentrations as compared to PLS; this was expected due to the ability of ANN to be trained on the target data during the training period. In all cases of data analysis by NLL, PLS and ANN the errors were approximately similar and ranged from about 1% to 15%.

Further work on food analysis involved the combination of pattern recognition coupled with partially specific sensors for the classification of beverages in an electronic tongue. An early work to develop a taste sensor for beer[55] or tomatoes[56] where multisensor arrays were used with lipid membrane transducers to develop a pattern recognition plot for the classification of different types of either beer or tomatoes onto what the authors termed a taste map. The taste map had subjective portions attached to the objective classifications which correlated to human taste senses. Beverages generally fell into patterns correlating with the human taste sensation. Further work correlating the taste senses of bitter, sweet, sour, and salty to specific foods to generate a taste map was done using an eight channel lipid membrane coated electrode system and measuring the electrochemical potential.[57] The measured potentials were then related to the areas of the taste map using pattern recognition.

A taste map relating amino acids to taste senses was also produced, as well as taste maps for beers, mineral waters, coffees, sake and tomatoes. The authors were able to show which channels correlated with which taste senses and reproduced this with the various samples.

Evaluative classification of different beverages has been done to distinguish different types of beverages - tea, coffee, juices, soft drinks, and beer- as a precursor to an electronic tongue useful for quality control in the food industry.[58] An array of 18-21 potentiometric sensors was used with various beverage samples. Principal components analysis was used to generate a pattern recognition plot, and artificial neural networks were used for monitoring the aging of juice over time. For the qualitative classification, good discrimination was observed among all the different types of beverages, as the samples fell into appropriate sub-population groupings. Aging of orange juice was monitored and the resultant plot of expected versus true aging time fell along a linear model.

An array of 29 different chemical sensors including chalcogenide glass, PVC membrane, metal sensors (Pt, Sb), and conventional ion selective electrodes was used along with principal components analysis to generate pattern recognition plots for wines and mineral waters[59] for qualitative analysis. Further, quantitative analysis of components of the waters and wines were also done using PCR. Concentrations of some ions (pH, fluoride, chloride, sodium, potassium, bicarbonate) present in the waters and ethanol and organic acids in the wines were determined with relative errors of 2 to 16% for the waters and 1 to 15% for the wines.

An electronic tongue using large and small amplitude pulses as the excitation signal was used to generate voltammetric scans from various beverages, and pattern recognition was used to classify the beverages into appropriate sub-population groups.[60] As well, the aging of juice was monitored over time, and a principal components plot was obtained of the scores over time, clearly showing the changes in scores with time.

Another pulse electrochemical technique, dual pulse staircase voltammetry (DPSV), was used for the simultaneous measurement of the quantities of ethanol, fructose and glucose at an unmodified platinum electrode, using neural networks to analyse the data generated.[61] Initially it was found that the peaks did not merely grow as the concentration of each species increased, indicating some interference of each analyte with the others. Peak response was shown to contain responses due to each analyte, without merely being additive, as the total current response was lower than if individual responses were summed. A possible reason can be electrode fouling by one species in solution, thereby interfering with the oxidation of the other species at higher potentials, as peak area is reduced with surface reduction, or the formation of an insulating layer caused by adsorption of species onto the electrode. Also saturation was postulated as another reason at higher concentrations, thereby reducing sensitivity of response. The use of multivariate techniques means that this type of voltammetric method can still be used to generate quantitative information, as multivariate techniques allow for simultaneous separation of the ternary data into respective quantities without the need for pre-separation by chemical or physical means. Comparison of ANN with PCR and PLS showed that ANN had reduced error of prediction due to its inherently greater ability to cope with varying blank responses, nonlinearity and interanalyte interference.

Although chemometric techniques have become widely used for analysis of food, and for classification of beverage samples according to their intrinsic properties, a cautionary note should be sounded for using these techniques without a full understanding of the methodology involved. Frequent use of factor analysis as a "black box" where prediction values are obtained can lead to serious misconceptions when improper data selection of the input variables, and inappropriate or inadequate validation is performed.[62]

1.5.4 Chemometric Analysis of Other Instrumental Data

The use of chemometric techniques for data analysis are not restricted to potentiometric input data, but other instrumental data can be used. Spectrophotometric data is the most commonly used method for generating data sets, but mass spectral or chromatographic methods are also common, as seen in the following examples.

The simultaneous quantitation of cobalt, copper and nickel in alloy samples using spectrophotometry, and PLS has a distinct advantage for analysis of alloys, as no pre-separation of the ions is required.[63]

The resolution and quantitation of pesticides in mixtures is another area where chemometric methods are useful. Ternary mixtures of pesticides were examined by spectrophotometry, then the resulting spectra were interpreted by various multivariate methods, such as CLS, PCR, and PLS.[64] Using UV-visible spectrometry, a comparison of the various mathematical techniques for the quantification of carbofuran, carbaryl and fenamiphos in a ternary mixture showed that all three methods were adequate to predict quantities of the three analytes, with PLS giving a lower root mean square error of prediction than PCR, which was slightly lower than using CLS.

Alternatively, high performance liquid chromatography coupled with PLS was used for the detection and quantitation of folpet, procymidone and triazophos in a ternary mixture[65] and iprodione, chlorothalonil, folpet, procymidone and triazophos in a five-component mixture[66] using UV-vis diode array detection. A multi-component spectrum was obtained, and analysis by PLS gave more accurate quantitation of the pesticide levels over that by PCR using the root mean square error of prediction. Using PLS analysis and testing river samples for spiked levels of the pesticides, recoveries of 80 - 110%, depending on the type of pesticide, were obtained during prediction.

Quantitative analysis of pyrolysis mass spectral data of lysozyme, DNA, and RNA in glycogen using artificial neural networks, and comparing the results to those obtained by PCR and PLS, as well as using these methods to estimate the percentages of bacteria in a ternary mixture of *S. aureus*, *B. subtilis* and *E. coli* was one of the first studies done to directly compare the predictions of the three methods.[67] Using PCR and PLS, prediction errors ranged from 1 to 6% and there were indications errors tended to be higher due to non-linear relationships between the spectra. Neural networks, taking into account nonlinear relationships, would then be more accurate for this type of data analysis and resulted in prediction errors of 0.5% for the test data.

Further work in monitoring the fermentation process has ranged from process monitoring to detection of recombinant proteins. The brewing process at Labatt Brewers was monitored for ethanol production using ANNs, to improve fermentation prediction.[68] Monitoring of bacteria to detect biomass, and successfully quantify microbial cell suspensions, using PCR and ANN, has been reported for yeast,[69] *Bacillus thuringiensis,*[70] *Penicillium chrysogenum,*[71] and *Escherichia coli.*[72]

Dow Chemical has been using chemometric techniques for quality control and for quantitation and identification of raw material and waste water streams since 1988.[73] Their techniques include using near infrared spectroscopy (NIR) and SIMCA for quality control probes of incoming materials for either quick acceptance or rejection of the sample; NIR and PLS to analyse a caustic stream for acceptable salt concentration ranges; NIR and CLS to predict olefin concentrations in another process stream; and analysis of organics in a waste stream using proton-NMR and PLS to model the BOD_5 levels after a 15-minute response time.

Further applications of chemometrics are described in a summary paper that shows the importance of chemometric stratagems in determining the important factors for enamine synthesis, monitoring and improving a crystallization process, and quality control in cheese making.[74]

One of the newest uses of chemometric techniques, especially hierarchal clustering analysis has been in gene expression profiling. Patterns of gene expression can be deciphered using clustering algorithms which recognize the underlying organization, or correlation present in a data set. Thus genes can be classified according to a basic congruence in their pattern of expression. Groupings generated by cluster analysis techniques may be used to determine class membership and functionally related elements of a data set.

For example, clustering algorithms have been used successfully to cluster genes of similar function in *S. cerevisiae*[75]. Alon et al. (1999), have employed hierarchal cluster analysis to detect groups of related genes in several types of tissues[76]. Their findings suggest a rudimentary organization of gene expression within tissues. The same investigators, using a two-way clustering algorithm, were able differentiate between malignant tumors and normal tissue on the basis of gene expression patterns. Recently, hierarchal clustering analysis was employed by Alizadeh et al. (2000), to classify B-cell lymphomas on the basis of their gene expression profiles[77]. Construction of a two-cluster self-organizing map has also been used to distinguish between two types of leukemia based on gene expression profiles[78]. Factor analysis, using partial least squares (PLS) on latent variables has been compared to principal components analysis (PCA) in the discrimination of cancerous and normal patients based on ICP-AES analysis of hair samples[79]. These reports support the use of pattern recognition as an effective methodology in classifying tissues according to the groups of genes they express.

1.6 Thesis Organization

In this chapter an overview has been presented of the mathematical methods used in obtaining research results covered in the following chapters, and a review of some of the chemometric methods, with particular emphasis on the factor-based methods, that have been applied for data analysis has been presented. Beginning with the work done by Winquist et al on the use of an amperometric technique for discrimination of complex liquids, two methods were used, under varying conditions and with different electrode materials to attempt to generate scores plots to differentiate between different beverages.

Chapter 2 covers the materials and methods used to generate these scores plots, and also discusses the materials and methods used for quantitative analyses.

Chapter 3 includes all qualitative results and the discussion of the amperometric techniques employed, the conditions used, and the electrode materials surveyed. Based on a random selection of easily obtainable beverages, the discriminating ability of these techniques, and the subsequent ability to generate scores plots which differentiated sub-populations is presented. From the preliminary survey, square wave voltammetry with a platinum (Pt) working electrode was chosen as the better combination. Amperometric scans were generated of several types of beverages, and using PCA to generate scores plots on the first two principal components, discrimination of these different beverages was obtained. Having determined that PCA is adequate for qualitative analysis of complex liquids, a more rigorous approach was proposed to use PCR for prediction of analytes in a complex liquid.

Chapters 4 and 5 cover the quantitative results obtained using both principal components regression and partial least squares. Rank analysis is also included to illustrate how the optimal number of factors was chosen. Chapter 4 is concerned with all results obtained on individual component analysis in complex media, such as orange juice and beer, while Chapter 5 contains research results of ternary mixtures of ethanol, glucose and pyruvate in beer, and the results obtained from both PCR and PLS in determining the individual concentrations in the ternary solutions. Chapter 5 was a logical extension of the work shown in Chapter 4. Initially, individual analyte concentrations in de-alcoholized beer were predicted using both PCR and PLS. Having ascertained that reliable and accurate prediction results were obtained for glucose, lactate, pyruvate and ethanol, the next obvious step was to combine the analytes and ascertain whether the simultaneous determination of individual analyte concentrations was possible. As shown, accurate, simultaneous prediction of pyruvate, ethanol and glucose is possible in a beer sample.

Chapter 2 Materials and Experimental Methods

2.1 Qualitative Analysis

2.1.1 Chemicals

Sodium phosphate, monobasic (Sigma, reagent grade) and sodium phosphate, dibasic (Fisher, reagent grade) were used to prepare phosphate buffers. Gold electrodes were cleaned in 3% hydrofluoric acid made from 49.1% hydrofluoric acid (Baker, ACS reagent grade) and saturated chromic acid made from concentrated sulphuric acid (BDH, reagent grade) and potassium dichromate (BDH, reagent grade). Platinum electrodes were cleaned in methanol (BDH, reagent drum grade). All electrodes were polished using polishing alumina (1.0 µm, Buehler Micropolish II) in an aqueous slurry on polishing cloths (Bioanalytical Systems, West Lafayette, IN).

The growth media were prepared from glucose, potassium phosphate, monobasic (Fisher, reagent grade), potassium phosphate, dibasic (Aldrich, 98+%), trisodium citrate (BDH, 99.0%), magnesium sulfate (BDH, ACS reagent grade), calcium chloride (Aldrich, 98+%), ammonium sulfate (Aldrich, 99+%) and ammonium chloride (Aldrich, 99.5+%). The growth media were made for me by Gabriele Hager.

All solutions were prepared from distilled, deionized water (Nanopure).

2.1.2 Samples

The samples consisted of 14 different brands of beer: Labatt Blue (1), Dave's Honey Brown (2), Sleemans Pale Ale (3), Sleemans Irish Ale (4), Waterloo Dark (5), Algonquin Honey Brown (6), Rickert's Red (7), President's Choice Brew, De-alcoholized(8); Ruddles Ale(9); Pilsner Urquell(10); Radegast Czech Beer(11); Holsten Festbock(12); Old English Malt(13); Crest Lager(14); Wines and Liquors: De Paysage Blanc de Blanc, French White, 13% (34), Inniskillin Vidal, Canada, White, 10.5% (15), Inniskillin Riesling, Canada, White, 12% (16), Inniskillin Chardonnay, Canada, White, 12% (17), Inniskillin Gamay Noir, Canada, Red, 12% (18), Inniskillin Cabernet Sauvignon, Canada, Red, 12% (19), Inniskillin Old Vines Foch, Canada, Red, 12% (20), Inniskillin Vidal Ice Wine, Canada, White, 10.5% (21), St. Remy Napoleon Brandy, France, 40% (22), Karlovarska Becherovka, Czech Herb Liquor, 40% (23), Gossamer Bay Chardonnay, California, White, 13% (24), Ernest & Julio Gallo Chardonnay, California, White, 13% (25), Chateau Roc de Minvielle Bordeaux, France, White,11.5% (26), La Cour Pavillon Bordeaux, France, White,11.5% (27), Freixenet Traditional Method Cava Sparkling Wine, Spain, 11.5% (28), The Balvenie Single Malt Scotch Whiskey, Aged 10 years, 40% (29), Libertas Pinotage, South Africa, Red, 12.5% (30), Valpolicella Classico, Italian Red, 12% (31), Balbi Vineyard Malbec Syrah, Argentina, Red,13% (32), Heritage Zinfandel, California Red, 13.5% (33), Kittling Ridge Estates Ice Wine and Brandy, Canada, 17% (34), Kittling Ridge Estates Gewurztraminer, Canada, White, 12% (35)Wellesley Apple Cider (36); 12 types of fruit juice: Minute Maid Apple Juice (37), Navel oranges, fresh squeezed(38), Zehr's fresh squeezed orange juice(39), Tropicana Pure Premium Orange Juice, not from concentrate(40), Minute Maid Orange Juice (41), Minute Maid Low Acid Orange Juice (42), Old South Orange Juice (43), Old South Pulp-Free Orange Juice (44), President's Choice Orange Juice (45), President's Choice No Pulp Orange Juice (46), No Name Brand Orange Juice (47); 10 types of coffee: Mother Parker's Mocha Java (48),Mother Parker's Hazelnut Vanilla (49),Mother Parker's Dutch Chocolate (50), Mother Parker's Columbian (51), Mother Parker's Irish Cream (52), Mother Parker's Butter Pecan (53), Staff room Coffee (Mother Parker's) (54), Second Cup Royal Blend (55), Starbuck's Espresso (56), Tim Horton's Coffee (57); 3 types of milk: Neilson 1% milk (58), Neilson 3% milk (59), Neilson 10% cream (60); Evian Spring Water (61); .

The orange juices from concentrate were diluted as per package directions (1:3) with distilled, deionized water. The navel oranges were cut and squeezed just before use. All other samples were used as received. No other pretreatment of samples was performed prior to analysis, unless otherwise noted.

For the bacterial study, samples of *Escherichia coli JM105*, *Bacillus subtilis*, *Staphylococcus aureus*, and *Saccharomyces cerevisiae* were obtained from Technical Services, Department of Biology on agar plates and grown into growth medium overnight. Then 1 mL of the overnight culture was introduced into 50 mL of growth medium for monitoring the growth curve and running concurrent square wave voltammograms and optical density (OD) at 600 nm.

2.2 Quantitative Analysis

2.2.1 Chemicals

Pyruvic acid (99+%), D-(+)-Glucose(ACS reagent grade), L-(+)-Lactate Monohydrate (ACS reagent grade) were all purchased from Sigma and used as received. Ethanol (99.98%, Chromatography grade) was purchased from EM Science.

For the assay to determine the amount of lactate, pyruvate or glucose present in President's Choice (PC) Brew, standards kits were purchased from Sigma Diagnostics for the quantitative determination of each component using an enzymatic assay and measuring the absorbance at 340 nm. The lactate assay contained lactate dehydrogenase, glycine buffer, nicotinamide adenine dinucleotide, Grade III, and lactate standard solution. The glucose kit contained glucose hexokinase reagent. The pyruvate kit contained Tris(hydroxymethyl)aminomethane, 1.5 M and sodium azide, 0.05% as Trizma Base solution; nicotinamide adenine dinucleotide, Grade III; pyruvic acid standard solution; and lactate dehydrogenase.

2.2.2 Samples

Dilution experiments were done using Tropicana Pure Premium Orange Juice and diluting with 0.050 M phosphate buffer (pH=7.02).

President's Choice De-alcoholized Beer Beverage (nominally 0.5% alc/vol) was used in all experiments when glucose, lactate, ethanol or pyruvic acid were added. When ternary additions of ethanol, glucose or pyruvic acid were added to 15.00 mL PC Beer, 0.050 M phosphate buffer (pH 4.57) was added to make up volume to 20.00 mL.

2.3 Instrumentation

Measurements of pH were performed using a Corning pH meter (Model 430).

Measurements of OD were performed using a Cary 1 double-beam uv-visible spectrophotometer set to read at 600 nm. Samples were pipetted into polystyrene disposable cuvettes to record OD, then voltammetry was performed on the same samples.

Electrochemical experiments were performed using an EG&G Instruments Potentiostat/Galvanostat (Model 263A) using a standard three electrode cell with either platinum, glassy carbon or gold working electrodes (BAS), a silver/silver chloride (3 M NaCl) reference (BAS) and a coiled NiChrome wire auxiliary electrode.

2.4 Procedures

2.4.1 Sample Treatment

Before any runs platinum and glassy carbon electrodes were polished using a slurry of 1μm alumina in water on a polishing cloth, and sonicated in and rinsed with water. Between samples, the platinum electrode was sonicated in reagent grade methanol, and rinsed with Nanopure water, and the glassy carbon electrode was re-polished.

Between samples, gold working electrodes were cleaned three times in saturated chromic acid followed by 3% hydrofluoric acid. Electrodes were rinsed between each wash with Nanopure water and were sonicated in water prior to use.

All samples were used as received, with no dilution or prior de-aeration, unless otherwise stated, and all samples were used from freshly opened containers. Experiments were performed at room temperature (25±2 °C).

Voltammograms of samples were run in random order, as samples became available. Preliminary scans were generated of a small number of the samples, and as more samples became available, further voltammograms were run. For the quantitative analysis, voltammograms were generated over the entire concentration range, then these were split into training (calibration), and validation dat sets.

Square wave voltammetry was performed from 1300 to -800 mV (v.s. Ag/AgCl) at a frequency of 5.00 Hz, and pulse height of 50 mV, unless indicated otherwise. Normal pulse voltammetry was performed from 900 to -500 mV.

2.4.2 Mathematical Treatment

Raw data were converted to spreadsheet format using Quattro Pro 8 (Corel Corporation, Ottawa, Ont., 1996-1999), averaged (n=3 runs per sample) and converted to a Lotus file for incorporation into MATLAB Ver. 5.3.1 (The MathWorks, Natick, MA., 1994-2000) programs. Each vector file (consisting of the averaged voltammogram for one sample) was incorporated into matrix files for further analysis. The matrix of measurements was then split into a training, and a validation data set. Respective concentrations corresponding to each vector were also saved in a different file.

Factor analysis (PCA, PCR and PLS) was performed using the Chemometrics Toolbox of MATLAB (Version 2.3, The MathWorks, Inc., Natick, MA, 1998).

PCA involved generating the eigenvectors and scores of the training set. To determine the optimal number of factors, rank analysis was performed on the training data set. First, the reduced eigenvalues (REV) according to the method of Malinowski[6] were obtained, and plotted against the number of factors. Another indicator function, the *F*-test was also used to assess the optimal number of factors. Finally, a cross validation based on a successive leave-one-out process was performed on the training data set, using the excluded data as a validation set. The number of factors corresponding to the lowest PRESS from the cross-validation analysis was determined to give a better estimate of the optimal number of factors, than the previous indicator functions.

Finally, having decided on the optimal number of factors, the data were regenerated using the optimal number of factors and a residual plot was also generated. This would show only random noise if the number of factors had been chosen correctly. Finally, PCR was performed using the calibration matrix to re-generate the training set concentrations, and to predict the validation set concentrations.

Since validation data was available, Kramer[7] recommends a further test to determine if the optimal number of factors has been chosen. The PRESS of the validation data were calculated based on the number of factors. The optimal number should agree with that obtained using only the calibration data set. However, rank analysis was performed based on the optimal number of factors determined from cross-validation of the calibration data set.

Graphing of calibration and prediction values and generation of all XY graphs, and re-generation of scores plots was done using GraphPad Prism version 3.0 for Windows (GraphPad Software, San Diego, CA., 1999). The predicted values were exported to Prism where they are plotted against the actual values on an XY plot. Linear regression was performed on each plot, and correlation coefficients calculated. Linearity based on the r^2 value and correlation coefficients were all calculated using Prism.

To determine the r^2 value, the regression model is compared to the null hypothesis model, as

$$r^2 = 1 - \frac{SS_{reg}}{SS_{tot}} \qquad (30)$$

where SS_{reg} is the sum of squares of the vertical distances of the best fit linear regression line, and SS_{tot} is determined from the null hypothesis as the sum of squares of the vertical distances of the points from the horizontal line which passes through the mean of all y-values. The correlation coefficient[80] was calculated from a comparison of the x- and y-data values, where y-values are the predicted, and x-values are the expected.

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} \qquad (31)$$

Quantitative analysis was also performed using PLS2 and compared to PCR via the above mentioned XY plots. PLS uses a different algorithm to generate regression vectors. It generates abstract factors based on both the voltammetric data and the concentration data, then rotates the two towards each other to minimize prediction errors. The PLS algorithm of MATLAB is used to generate the weights and abstract factors. As before, rank analysis was performed to generate the optimal number of factors. The residuals were plotted to ensure only noise is present, but in the case of PLS there may still be traces of the original voltammograms in the residuals because PLS compromises the fit of the voltammogram factors in attempting to fit the regression between the voltammogram data and the concentration data. Finally prediction of the calculated values for the calibration and the prediction sets was done and these values were exported to Prism as described previously.

In both qualitative and quantitative analysis it is possible to generate scores plots which are two-factor plots of the resulting scores generated from PCA. These are useful in classifying sub-populations of data sets to find qualitative correlations between the groupings. The scores for factor one (PC 1) and factor two (PC 2) are exported to Prism in order to plot the results.

2.5 Methods

2.5.1 Optimization of Voltammetric Conditions

To optimize the signals obtained from square wave voltammetry, pulse height, frequency and filter conditions were varied using 1% milk sample, and scanning from 1.3 to -0.800 V (vs. Ag/AgCl) using a Pt working electrode. Pulse height and frequency were initially maintained constant at 25 mV and 2 Hz, while scans were run with filter off, filter at 5.3 Hz, and 590 Hz. Then the filter was turned on and at both 5.3 Hz and 590 Hz, frequency was set at 2, 5, and 10 Hz while pulse height was constant at 25 mV. Then frequency was alternated between 2 and 5 Hz, filter was set to 590 Hz, and pulse height varied between 25 and 50 mV. Determination of pulse height and frequency was repeated using spring water. The most noise-free signal chosen was under conditions of filter at 590 Hz, pulse height at 50 mV and frequency at 5 Hz. All scans were subsequently run under these conditions.

To determine the best electrode to use as the working electrode, platinum (Pt), glassy carbon (GCE), and gold (Au) electrodes were used for scans set at 590 Hz filter, 5 Hz frequency and 50 mV pulse height of samples of 1%, 3.25%, half-and-half milk, Spring water, Minute Maid Apple juice, Algonquin Honey Brown, Rickert's Red, Waterloo Dark, Minute Maid orange, Minute Maid Low Acid orange, Old South, Old South Pulp free, President's Choice, President's Choice Pulp Free, and No Name orange juices. Scans of each sample were run in triplicate, imported into Quattro Pro, averaged, saved as Lotus files for import into MATLAB Chemometrics Toolbox where factor analysis was performed. PCA yielded two-factor scores plots for each electrode, and it was determined that platinum was the best working electrode as it gave the most distinct sub-population groupings.

Using the 590 Hz filter, and the same samples as described above, normal pulse voltammetry was run between 0 to 0.900 V, -0.400 to 0.900 V, -0.500 to 0.900 V, -0.400 to 1.300 V; and the reverse of these potentials was also run, using platinum, glassy carbon and gold working electrodes. Data treatment was as described previously, and resulting scores plots were compared. These results showed that square wave was the better technique to separate out sub-populations. All subsequent experiments were run using square wave voltammetry.

Results from these and other optimization studies are presented in Chapter 3.

2.5.2 Pretreatment of Samples

Orange juice samples were scanned without pretreatment as well as with 15-minute deaeration, and after filtration to remove pulp solids. Voltammetric scans showed no obvious differences, therefore future runs did not include any pretreatment of the samples.

2.5.3 Bacterial Culture Study

After sterile inoculation, bacterial cultures were grown overnight in a shaker (200 rpm) at 30°C for *B. subtilis*, *E. coli*, *St. aureus*, and 24°C for *S. cerevisiae*. One mL of overnight culture was pipetted into 50 mL of fresh growth medium in a fresh shake flask and returned to grow in the shaker, and incubation time was recorded. Approximately one hour later 2 mL were taken from the shake flask, and time was recorded for each subsequent sampling. Each sample was measured for OD at 600 nm, and scanned using square wave voltammetry, in triplicate. Further scans were run at half-hour intervals until an OD reading indicated exponential growth was finished. Voltammetric scans were imported into Quattro Pro, replicates were averaged, vectors were combined into matrices, and transferred to MATLAB for pattern recognition analysis. Two analyses were performed. First, scores plots of the individual cultures were generated in two-factor space, and then data from all four cultures were incorporated together in one matrix in an attempt to differentiate the different sub-populations. Finally, principal components regression was performed on the individual species matrices to attempt to predict OD_{600} values from the calibration data set and overlay this with OD results obtained using UV-vis spectroscopy to generate growth curves.

2.5.4 Dilution Study

Tropicana Pure Premium Orange Juice was successively diluted with 0.050 M phosphate buffer (pH 7.02), and square wave voltammograms were generated. These were imported into Quattro Pro, triplicate scans were averaged, and the final matrix was split into a training set (9 vectors) and a validation set (5 vectors). Factor analysis using PCA and PLS was performed in the Chemometrics Toolbox of MATLAB, resulting in a scores plot after pattern recognition analysis and generating a calibration model and a prediction model. The resulting calibration and prediction values were exported to Prism and compared to the actual dilution values by plotting on an XY graph.

2.5.5 PC Brew and Ethanol, Lactate, Pyruvate, and Glucose

A series of experiments were made by adding increasing levels of ethanol, lactate, pyruvate or glucose to PC Brew. Individual amounts of lactate, pyruvate or glucose were added to 15.00 mL of PC Brew. Approximately 15 samples, per analyte, were run using square wave voltammetry. Concentrations were recorded, based on the added amounts in the 15.00 mL. Ethanol was successively added to PC Brew in volume increments. In all cases, a blank (PC Brew only) was also run. Over the twelve to fifteen samples run, 5 samples were removed to be used as a validation data set. These were chosen to be representative over the entire concentration range scanned. Training sets were designed as random data sets in order to ascertain if PCR and PLS could be used to predict both training set and validation data set concentrations. In some cases, a further 3 to 5 samples were withheld to be used as unknown samples, for the calibration solution to predict the analyte levels.

The concentration ranges were chosen to be, roughly, representative of actual amounts in beer samples[81,82] which were pyruvate (0.2 to 1.2 mM), lactate (0.5 to 2.0 mM), glucose (0.2 to 55 mM), and ethanol (0.5 to 12%). In all cases the dynamic range tested was greater than literature values. The analytes, themselves, were chosen as representative of the glycolytic pathway by which glucose is converted to ethanol.[83] Pyruvate is a mid-stage product of this pathway, and under anaerobic conditions, lacatate (instead of pyruvate) will form.

Voltammograms were generated, the resulting data set was split into training(8 to 10 samples) and prediction(5 samples) sets, and after factor analysis, predicted concentration values were plotted against the true values, as described previously.

Plots were generated of the calibration and prediction values based on the optimal number of factors. As well, the number of factors used in the prediction were varied to determine the effect on the final predicted concentrations. The effect that the number of factors had on the final prediction model was done by comparing the correlation coefficients generated from the actual and the predicted values. To compare the factor-based predictions to the actual values, RMSE of calibration (RMSEC) and prediction of the validation set (RMSEP) was also determined.

2.5.6 Ternary Solutions of Ethanol, Pyruvate and Glucose in PC Brew

Varying levels of ethanol, pyruvate and glucose were added to 15.00 mL of PC Brew, then 0.05 M phosphate buffer (pH 4.57) was added to make up to 20.00 mL. Final recorded concentration levels were calculated based on amount added in the 20.00 mL. As before, the analyte concentrations were representative of the amounts present in beer samples, and comprised a random data set. Further, a larger data set was designed to cover the entire possible range from 0 to 1 (normalized values), and using structured combinations of the three analytes. The concentration ranges again, were based on the actual concentration ranges of analyte present in beer samples.

As mentioned in the previous section, samples were run using square wave voltammetry, then scans were split into a training set and a validation set, and rank analysis was performed to determine the optimal number of factors, and finally both PCR and PLS were used to generate calibration and prediction matrices. Values calculated from these algorithms were compared to the actual values as described previously. Again, different numbers of factors were used to generate the calibration and prediction sets and the resulting values were compared to the actual values.

Chapter 3 - Qualitative Analysis

3.1 Results and Discussion

3.1.1 Preliminary Assessment

The initial requirements for this project were the identification of conditions for a relatively noise free signal, giving good signal-to-noise ration, for a variety of complex liquids. Relatively little work had been done on the combination of voltammetry and PCA (see Chapter 1, Literature Review), yet the combination of a low-level detection technique, and multivariate data analysis could be a powerful analytical method. Voltammetry is a useful technique for low-level quantification of a wide range of possible species. Signals are generated not only from reducible or oxidizable species, but also from chemi- or physiosorbed species on the electrode. It also has the added advantage of simplicity and speed. At a solid electrode, oxidation of nitrogen and oxygen containing organic compounds can be performed, which makes this technique advantageous for the types of complex liquids discussed in this work.

Initially a survey was done to determine which combination of electrode and voltammetric technique would generate the most useful voltammogram in terms of resolution, signal-to-noise ratio, and variance in a scores plot. Electrodes used were platinum(Pt), glassy carbon(GCE) and gold (Au), which were prepared as described in the Materials and Methods chapter. Voltammetric techniques utilized were normal pulse voltammetry (forward and reverse scanned) and square wave voltammetry. Both techniques were tested with all electrodes. Initially, untreated milk samples (1%) were run with GCE and Pt as the working electrodes in square wave and normal pulse voltammetry.

For square wave voltammetry, several scans of this sample were run to determine the range, varying from 1.3 V to 1 V starting potential (vs Ag/AgCl), and ending at between -0.4, -0.6 or -0.800 V. With both Pt and GCE, a scan from 1.3 to -0.800 V, while covering a wide range, resulted in a good baseline, and two regions where reduction may be observed.

Scan conditions were then optimized for Pt by varying frequency, pulse height and low pass filter conditions. First, keeping the frequency at 2 Hz and the pulse height at 25 mV, the filter was switched from off (A), to on at 590 Hz (B), to on at 5.3 Hz(C), as shown in Figure 3.1. The 590 Hz filter gave a less noisy signal than when the filter was off, as well as a smoother signal over the small peak area from 500 to -400 mV than the 5.3 Hz filter, so subsequent tests were done using the 590 Hz filter. The 5.3 Hz filter is expected to filter some of the signal as well as the noise.

Next, the frequency was varied from 2 to 10 Hz (Figure 3.2), keeping other conditions constant. It was decided to use the 5 Hz frequency as the 2 Hz signal lost detail, while the 10 Hz signal was too noisy. Figure 3.3 shows the effect of pulse height, where 50 mV pulse height (B) was better than 25 mV (A) which tended to give noisier signals also. To determine if this combination was unique to the sample, the entire sequence was repeated using untreated spring water (see Figure 3.4), a sample not expected to give a large signal. The small signal generated was not noisy, hence these conditions should be adequate to generate useable voltammograms for a wide variety of samples. The best combination then was a 50 mV pulse generated at 5 Hz. This was adequate to perturb the double layer for a sufficient period in order to generate redox activity, without introducing noise to degrade signal quality.

For all subsequent square wave voltammetric experiments, scans were therefore run from a potential of 1.3 to -0.800 V, using a frequency of 5 Hz, pulse height of 50 mV and keeping the filter on at 590 Hz. Based on these results, the GCE was tested under these conditions and the collection of square wave voltammograms of a wide sample population was begun using the GCE and the Pt electrodes. This will be presented in a subsequent section.

Normal pulse voltammetry scan conditions were also varied. Figures 3.5 and 3.6 show voltammograms run from +0.900 V to -0.400 V, and from -0.400 V to +0.900 V, respectively. Both experiments were run using the GCE and the Pt working electrodes, with the GCE giving noisier and much weaker signals for all samples. In all cases, for normal pulse voltammetry, the scan rate was 5 mV/sec and pulse width was 0.050 sec with step time of 1.000 sec. From the normal pulse voltammetric results it quickly became obvious that signals generated were not as well resolved, even at optimized operating conditions with the Pt electrode. Signals were weak, and poorly differentiated between different samples, especially with the carbon electrode. A possible explanation for the poorer results obtained using normal pulse techniques, is the increased sensitivity of square wave voltammetry due to the decreased non-faradaic (charging) current. Also, GCE could be subject to fouling (for both voltammetric techniques), resulting in a decreased signal, and poorer differentiation of signals between samples.

Pretreatment of samples by deaeration and filtering was investigated using square wave voltammetry, the Pt electrode, and either Waterloo Dark beer or Minute Maid orange juice sample. Figure 3.7 shows the effect of a 15 minute deaeration to remove carbonation from the beer sample. A small positive shift and intensity reduction may be observed, which may be sufficient to change the outcome in subsequent factor analysis. Thus, in subsequent experiments, the carbonated samples (beers and champagne) were pretreated by sonication for 15 minutes. Next, solids were removed from an orange juice sample by filtration. No obvious difference was found in the voltammograms, so further experiments did not use pre-filtering.

The gold electrode was then examined using the same scan conditions and the same sample set run with GCE and Pt. The difficulty with Au was immediately apparent as extensive cleaning was required due to fouling of the electrode. However, square wave and normal pulse scans were run with Pt, GCE, and Au electrodes on samples of orange juices, 3 beers (Waterloo Dark, Honey Brown, Richards Red), 3 types of milk, spring water, apple juice, and coffees to determine if scores plots would differentiate the sub-populations. Despite the cleaning procedure done between sample scans, signals were weaker, in terms of current recorded, than with either Pt or GCE. As well, little differentiation was observed between sample classes.

Figure 3.1 Square wave voltammograms of 1% milk with frequency 2 Hz, pulse height 25 mV, and filter off (A), 5.3 Hz (B) or 590 Hz (C)

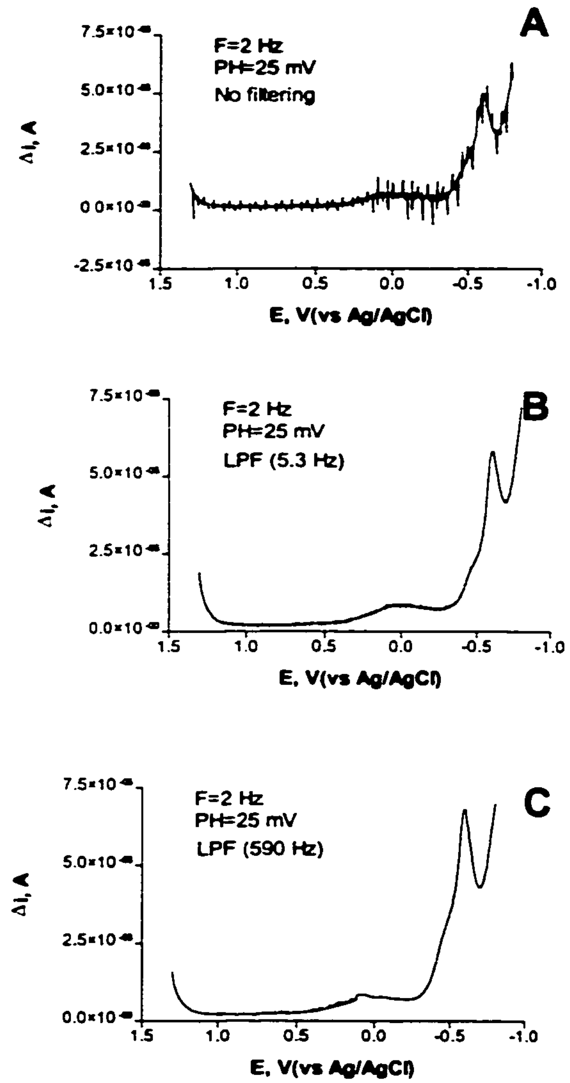


Figure 3.2 Square wave voltammograms of 1% milk with pulse height 25 mV, filter of 590 Hz and frequency of applied pulse 2Hz (A), 5 Hz (B) or 10 Hz (C)

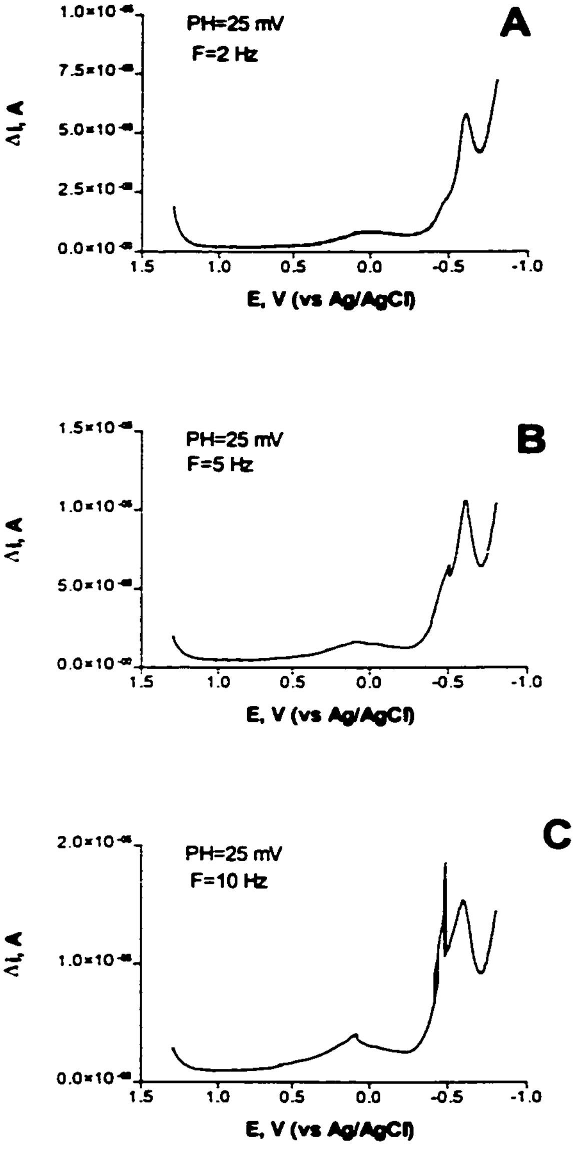


Figure 3.3 Square wave voltammograms of 1% milk with frequency of pulse 5 Hz, filter of 590 Hz, and pulse height of 25 mV (A) or 50 mV (B).

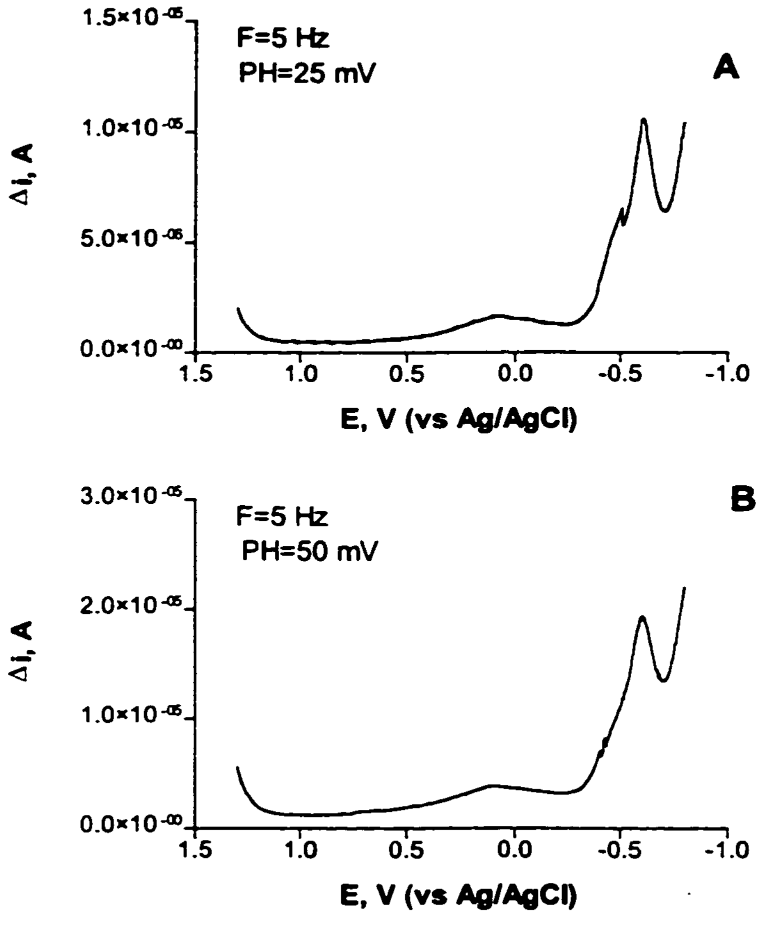


Figure 3.4 Square wave voltammograms with frequency of pulse 5 Hz, filter of 590 Hz, and pulse height of 50 mV for 1% milk and spring water.

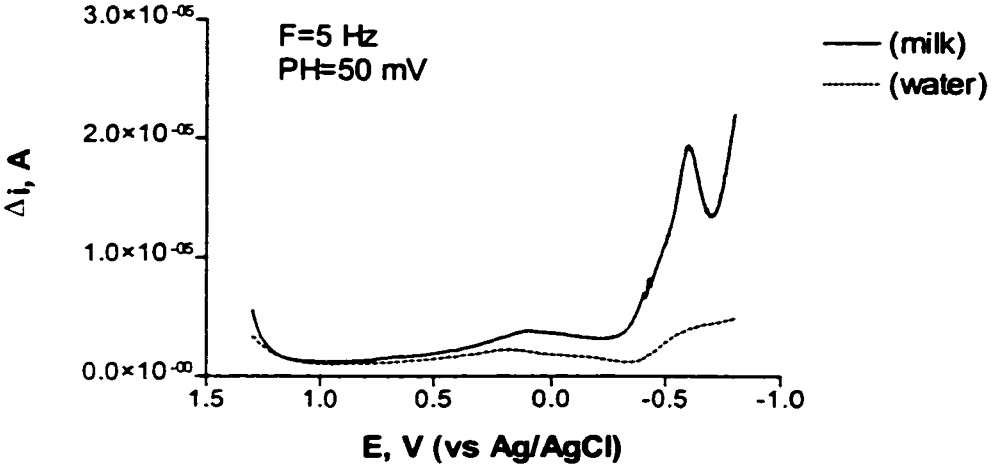


Figure 3.5 Normal pulse voltammograms from 900 to –500 mV of milk, juice, water and beer samples obtained using a GCE (A) and Pt electrode (B).

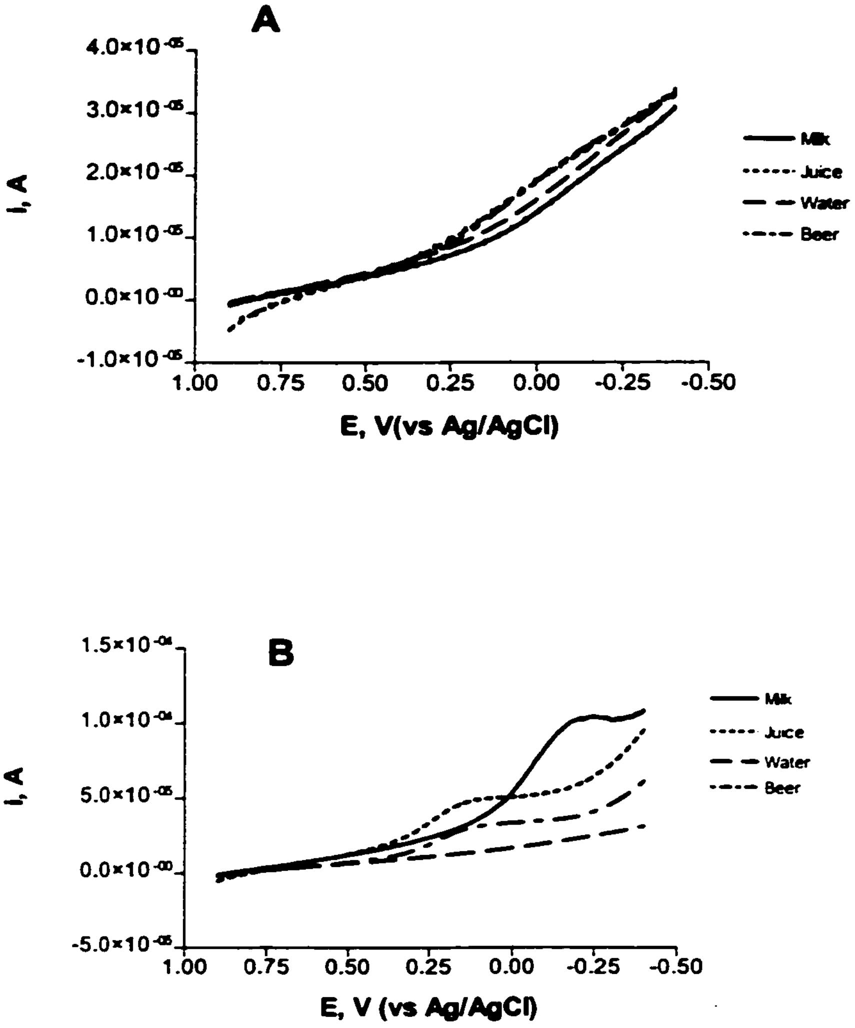


Figure 3.6 Normal pulse voltammograms from –400 to 900 mV of milk, juice, water and beer samples obtained using a GCE (A) and Pt electrode (B).

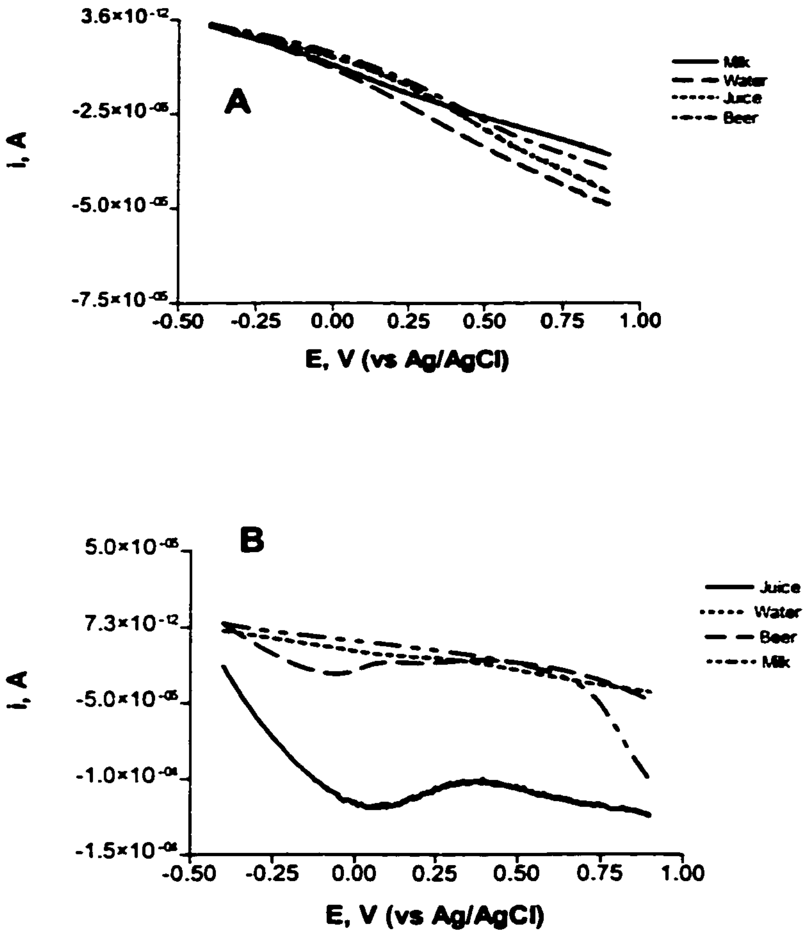
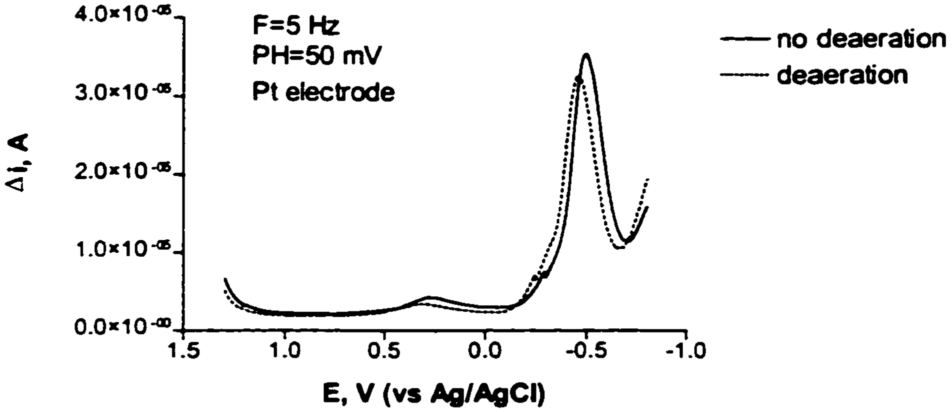


Figure 3.7 Square wave voltammogram of Waterloo Dark Beer run at 5 Hz frequency, pulse height of 50 mV, before and after sonication to remove carbonation.



3.1.2 Sample population analysis

After the optimal conditions were determined for the two voltammetric techniques, voltammograms of each sample were generated with both square wave or normal pulse voltammetry, using Pt, GC, and Au electrodes. These voltammograms were then transferred into a spreadsheet, and scores plots were generated for each technique and electrode material. No data pretreatment was performed. Scores plots were used to assess the capability of the different electrode materials and the two different amperometric techniques to differentiate the juices, beer, water, and milks into their respective classes.

Normal pulse voltammetry results are shown in Figures 3.8 for the Pt electrode, 3.9 for GCE, and 3.10 for Au. Square wave voltammetry scores plots are shown in Figures 3.11 for Pt, 3.12 for GCE, and 3.13 for Au. Initially, the same number and type of beer, juices, water and milks was run to compare electrodes and techniques. As observed from the respective scores plots, square wave voltammetry and the Pt electrode were able to differentiate the classes with the greatest spread between classes. The scores plot generated using the Au electrode and square wave voltammetry also was able to differentiate the sub-populations, except for the beer samples, but the variance among classes was not as good as when the Pt electrode was used. Further, with the Pt electrode, the sub-populations were separated into distinct clusters, except for the coffees and milks, and the groupings tended to be more tightly clustered in their respective classes. As the Au electrode also suffers from fouling the best combination was Pt and square wave voltammetry.

With the Pt electrode, and square wave voltammetry, the variance between classes was wide, and subsequently clear clustering of the sub-populations was observed (Figure 3.11). Further studies using the combination of Pt electrode and square wave voltammetry were then performed, adding samples as they became available. The aim in this portion of the project was to collect voltammograms from as many representative samples as possible.

Figure 3.8 Scores plot of milks, beer, and juices obtained using normal pulse voltammetry with a Pt WE.

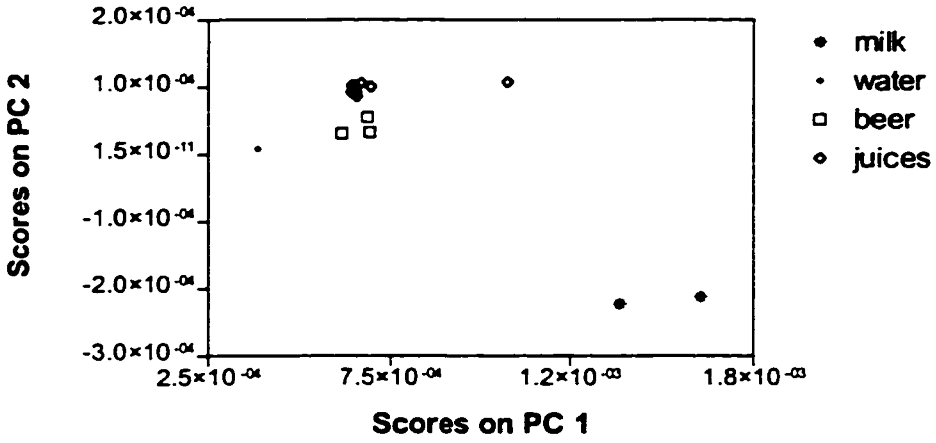


Figure 3.9 Scores plot of milks, beer, and juices obtained using normal pulse voltammetry with a GCE.

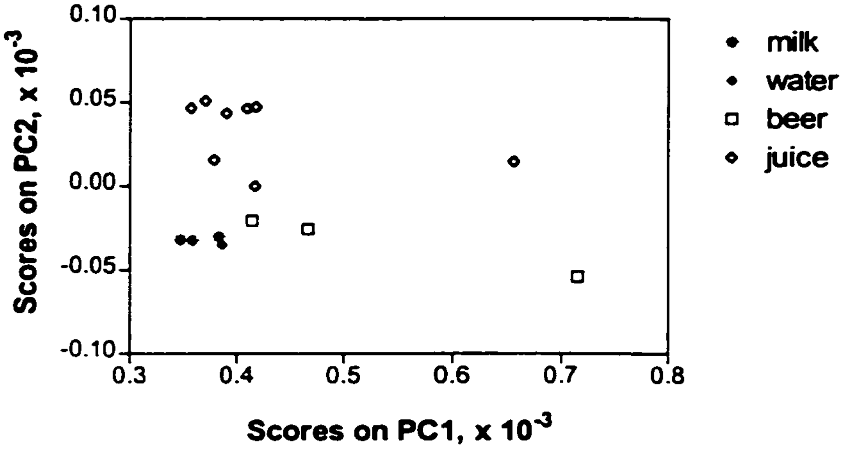


Figure 3.10 Scores plot of milks, beer, and juices obtained using normal pulse voltammetry with a Au WE.

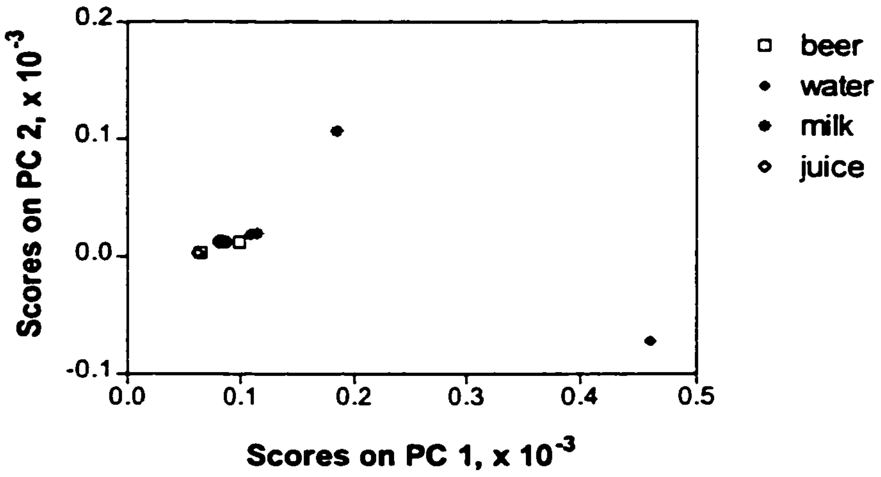


Figure 3.11 Scores plot of milks, beer, and juices obtained using square wave voltammetry with a Pt WE.

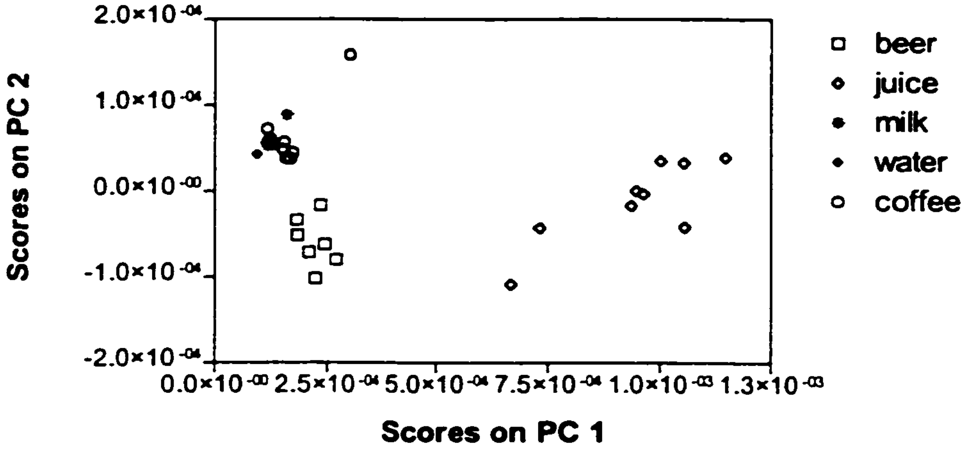


Figure 3.12 Scores plot of milks, beer, and juices obtained using square wave voltammetry with a GCE.

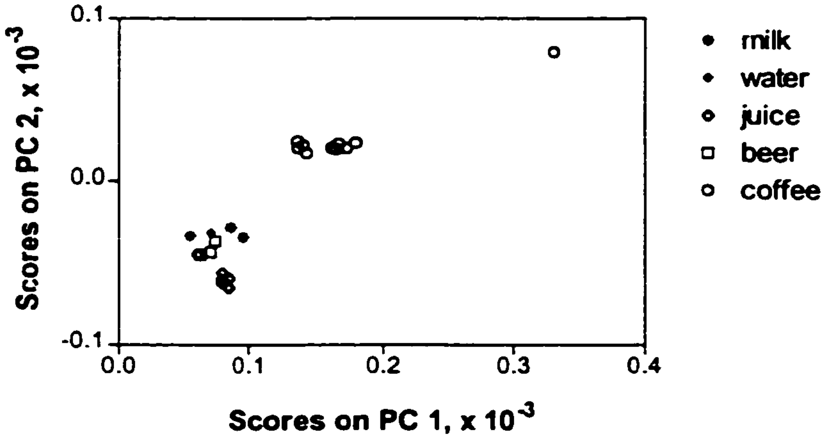
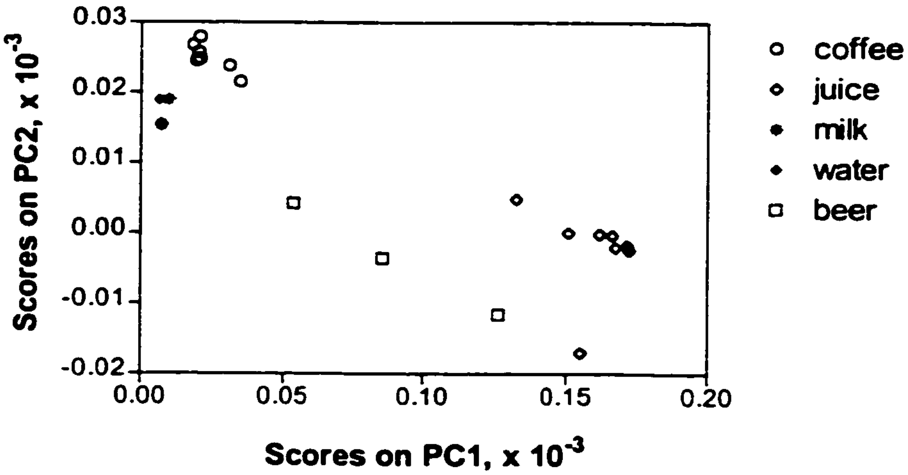


Figure 3.13 Scores plot of milks, beer, and juices obtained using square wave voltammetry with a Au WE.



3.1.3 Pattern Recognition Analysis of Individual Classes

Fourteen types of beer were examined using square wave voltammetry and the Pt working electrode, with two examples shown in Figure 3.14. Each voltammogram was loaded as a vector (1051x1) into a matrix (1051x14). Using MATLAB factor analysis, matrices of eigenvectors and eigenvalues were generated. For a qualitative scores plot only the two principal factors which account for the majority of the variance were plotted, and variances for these factors (98.22% for PC 1, 1.09% for PC2) were noted. A scores plot for the beers (Figure 3.15) was generated, and shows a wide spread or variance between the samples. No significant clustering of scores according to beer type occurred. From the respective voltammograms generated, it was apparent that differences between the beers should translate into distinctiveness on the scores plots, as was observed. Even the two honey brown beers (2 and 6) do not cluster near each other, but maintain enough of a distance to be distinct. It is possible that the apparent lack of clustering was due to the relatively small data set used where no attempt was made to test similar beers, or perhaps the samples don't cluster in this space.

This procedure was repeated for coffee samples with the generation of eigenvalues and eigenvectors from a matrix of the square wave voltammograms. A typical square wave voltammogram is shown in Figure 3.16. The resultant scores plot (Figure 3.17) showed two regions at the opposing ends of the plot; the first due to all the coffee samples except the espresso, and the other due to the espresso. For the scores plot, PC 1 contributed 97.31% of the variance and PC 2 contributed 1.74%.

From the voltammograms (Figure 3.16), the Royal Blend and the Mother Parkers blend show similar peaks, while it can be seen that the Starbucks espresso shows a more pronounced large peak at about -0.600 V. This clearly differentiates the espresso from the other coffees, and indeed Starbucks is an outlier in the resulting scores plot (Figure 3.17). Another interesting observation in the large cluster is the apparent split between those coffee samples which were flavored (top) and non-flavored (bottom) types. Flavored coffees are typically ground with a flavoring oil added, so this may account for the split within the cluster. However, a definite conclusion based on these lines would require more experimentation using a larger population base than given here, as this difference could also be due to the location from which the coffees were obtained, as all the flavored coffees were obtained from one location, while the other types came from various vendors.

The voltammograms of the juices tested (10 orange juices, apple cider and apple juice) are shown in Figure 3.18 . The scores plot for the juices (Figure 3.19) differentiates between a larger grouping of all the orange juice samples made from concentrate, the fresh squeezed oranges (38), Tropicana fresh squeezed orange juice(40) and, in the opposite quadrant a clear separation between the apple cider(36) and the apple juice(37). In this population, PC 1 contributes 99.21% of the total variance and PC 2 contributes only 0.42%.

The voltammograms (Figure 3.18) show an observable difference between the apple juice and the orange juice (both Minute Maid brands). This difference is clearly shown on the scores plot (Figure 3.19) with apple juice (37) clearly separate from orange juice (41) as expected from the large difference in their peak heights on the voltammogram which translates into large variance on the scores plot. From the scores plot the apple cider (36) is also separated from the larger group of orange juices as well as from the apple juice. Tropicana fresh squeezed orange juice (40) is also separated from the other brands, lying closest in distance and hence closest in voltammetric similarity to the fresh squeezed orange juice (38). It is interesting to note that the fresh squeezed orange juice from Zehr's Supermarket (39) actually lies within the grouping of orange juices reconstituted from concentrate. In Figure 3.19, differences are also observed between the apple beverages and the orange juices which form separate groups, as well as another grouping of orange juices reconstituted from concentrates which also tend to cluster together. Possibly the concentration or reconstitution process of the orange juices from concentrate causes similar changes in all these samples.

Milk samples and spring water were run together (Figure 3.20) due to the small sample size. Variance present in the first two factors is 98.66% for PC 1 and 0.79% for PC 2. The scores plot (Figure 3.21) shows spring water (61) at the opposite end of 1% milk (58), with the whole milk (59) and cream (60) clustered together.

The voltammograms (Figure 3.20) show the expected differences between milks and water, which are clearly observable on the scores plot (Figure 3.21). From this plot, water (61) is quite dissimilar from the 1% milk and surprising more similar to the whole milk and cream. However, this difference in distances can be attributed to the voltammetric results, which show the largest differences in peak positions to occur between 1% milk and water. Water, as expected, shows little activity on the voltammogram.

Finally, a variety of wines and liquors were scanned in the same manner, with representative voltammograms shown in Figure 3.22. From the voltammogram, brandy appears to give no signal on the current scale shown. A much smaller scale, shows brandy (and whiskey) with two broad peaks, at the same relative potentials as the other alcoholic samples. It is postulated that the higher alcohol content gives the lower and broader peaks, but no further examination was made. The resulting scores plot is shown in Figure 3.23. In this scores plot, PC 1 contributes 95.39% of the variance and PC 2 contributes 3.95%.

The voltammograms of these samples (Figure 3.22) suggest that there should be correspondingly greater differences between the samples on the scores plots, as each type of wine or liquor clearly gives a distinct signal. This is shown on the scores plot (Figure 3.23) with liquors clearly separate from the other types on opposing sides of the plot; a commercially-available combination of ice wine and brandy (34) falls between the wines and the brandy, as expected since it would contain characteristics of both. Unfortunately, white wines can not be distinguished from red wines from the scores plots. Other more complex matrix effects must be responsible for the separations and groupings. There is a grouping of Inniskillin wines (15-19) in one cluster, however this cluster also contains wines from other vineyards. Thus conclusions about wine type cannot be formed from this scores plot, except for gross differences observed between wines and liquors.

Figure 3.14 Square wave voltammogram of Dave's Honey Brown and Sleemans Ale, using a frequency of 5 Hz, pulse height of 50 mV and a Pt WE.

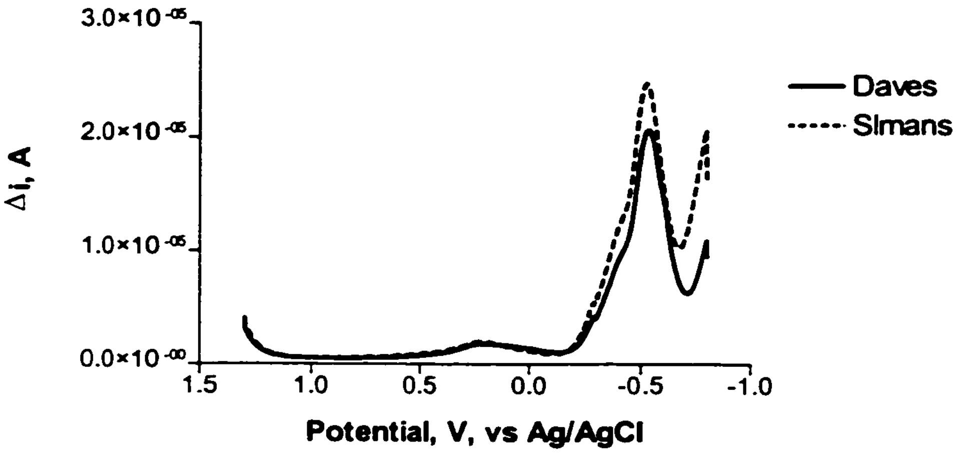


Figure 3.15 Scores plot for beer samples.

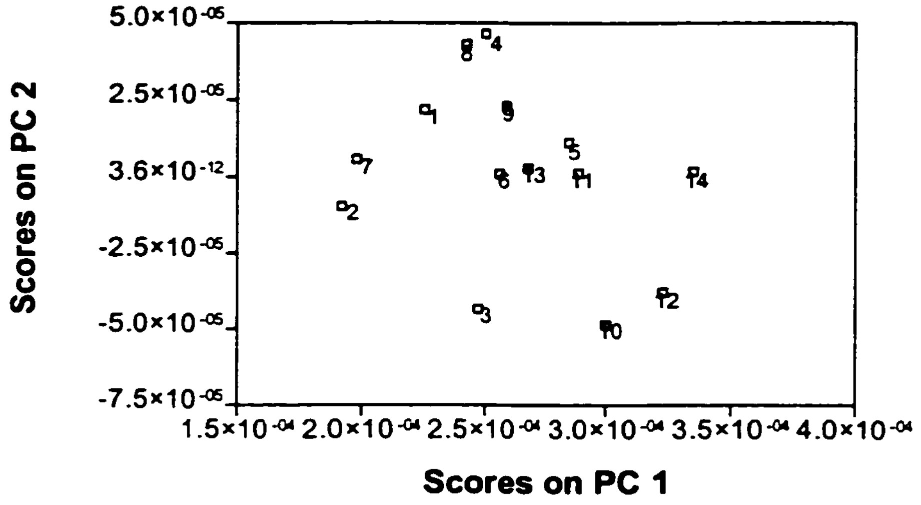


Figure 3.16 Square wave voltammogram of three coffees (Staffroom, Royal Blend, Starbucks espresso), using a frequency of 5 Hz, pulse height of 50 mV and a Pt WE.

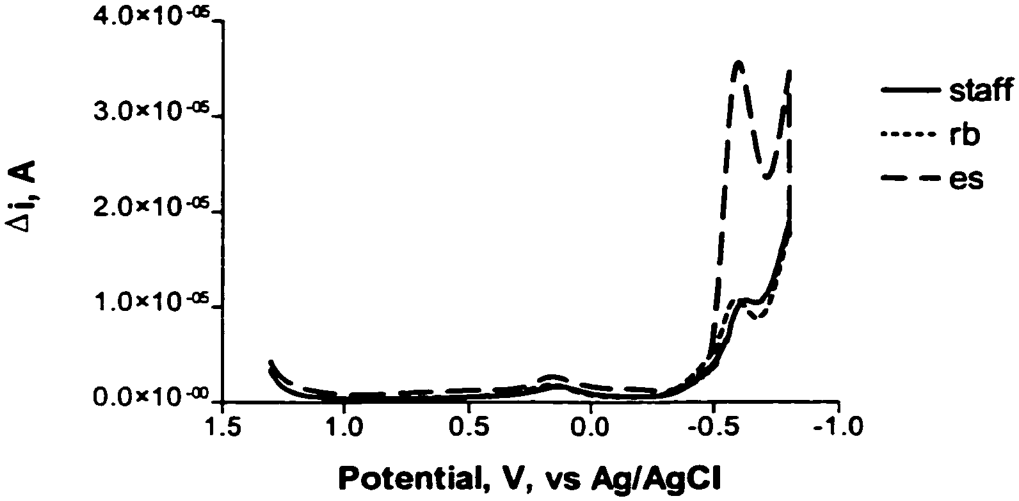


Figure 3.17 Scores plot for coffee samples.

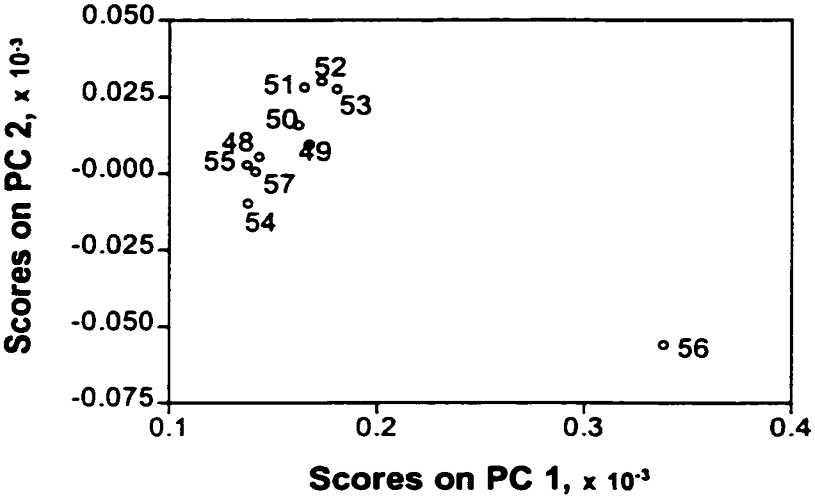


Figure 3.18 Square wave voltammogram of Minute Maid apple and orange juice, using a frequency of 5 Hz, pulse height of 50 mV and a Pt WE.

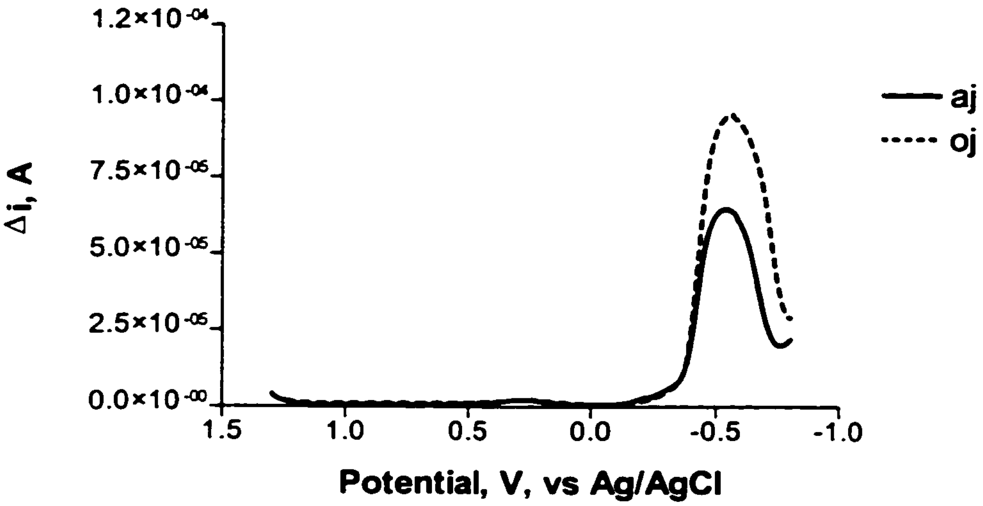


Figure 3.19 Scores plot for juice samples.

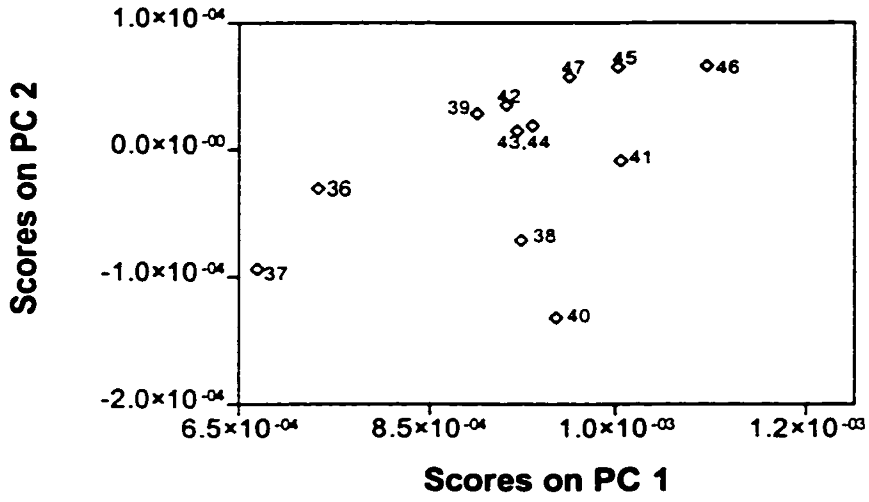


Figure 3.20 Square wave voltammogram of 1% milk, cream, and spring water, using a frequency of 5 Hz, pulse height of 50 mV and a Pt WE.

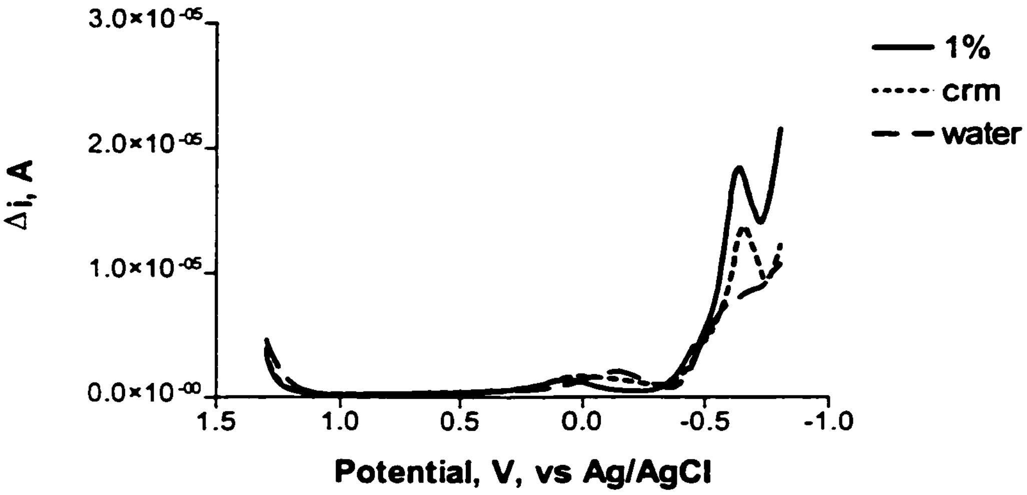


Figure 3.21 Scores plot for milk and water samples.

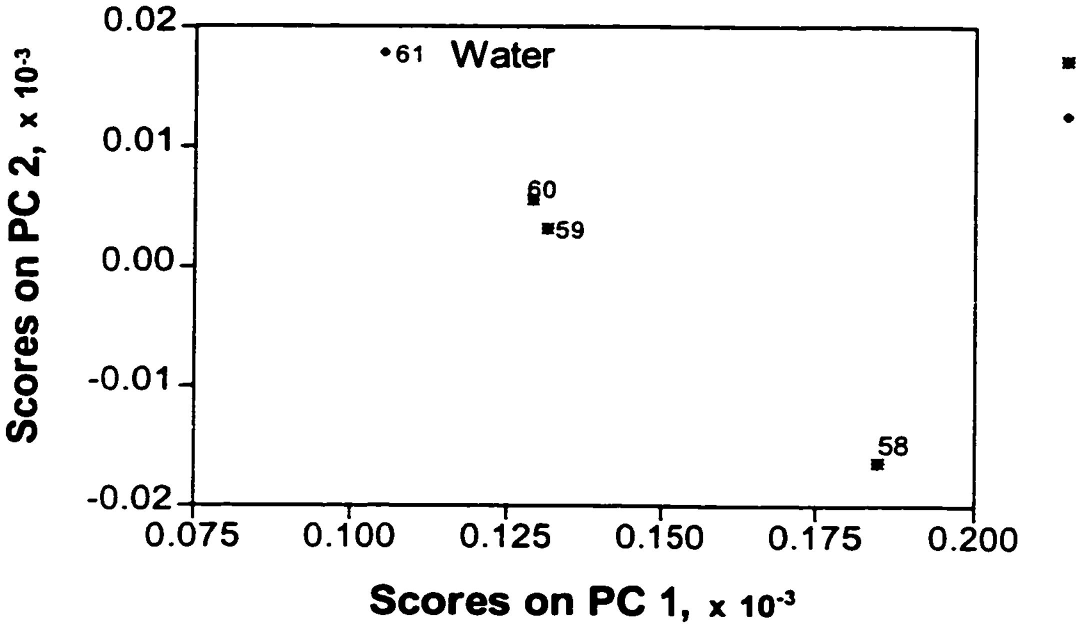


Figure 3.22 Square wave voltammogram of Inniskillin Vidal (white), Gamay Noir (red), Ice wine, and brandy, using a frequency of 5 Hz, pulse height of 50 mV and a Pt WE.

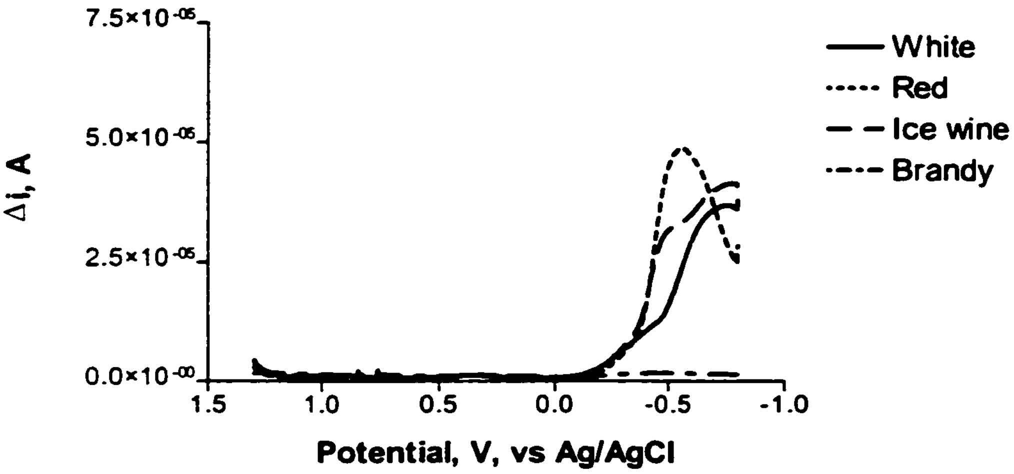
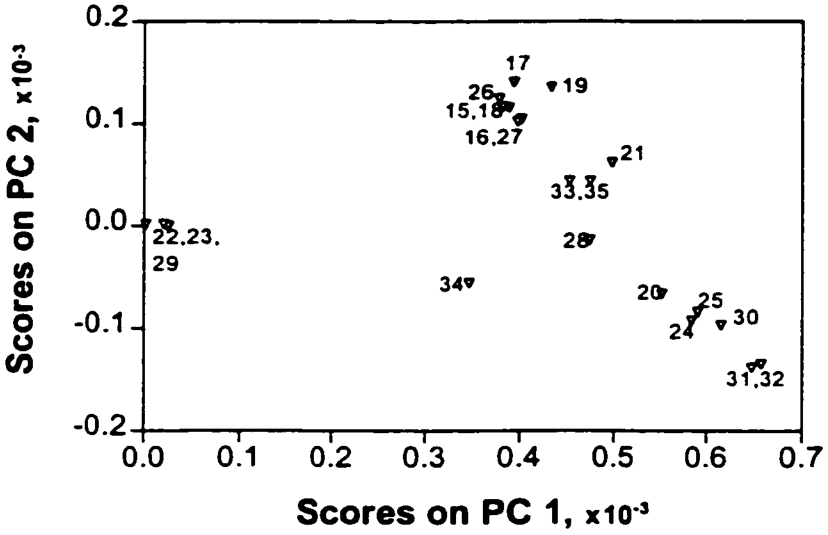


Figure 3.23 Scores plot for wines and liquors samples.



3.1.4 Pattern Recognition Studies

Sample vectors from entire populations were incorporated into a larger matrix (1051x61), and PCA was performed to generate the resultant scores on the first two principal components. The scores plot shown in Figure 3.24 is based on these two factors (PC 1 has 96.24% of the variance and PC 2 has 2.42%). Clustering of the sub-populations shows distinct sub-population groupings for juices, beers, and liquors, while a diffuse grouping can be seen for the wines, and a cluster exists due to the coffees and milks. Although distinct groupings are observed with the wines present in this plot, clear demarcation of the groups is not ideal as the wines constitute a grouping with a large variance. Thus, for the interest of clarity, the wines were removed from the matrix, and a new scores plot was generated (Figure 3.25). After factor analysis, the scores plot of the two most significant principal components (variance of 97.44% for PC 1, 1.47% for PC 2) showed clear separation between the sub-populations which tended to cluster together in their respective sample classes. The coffee and milk groups were still overlapped except for Starbucks espresso which is some distance away.

In order to determine if the groupings were valid[84] and reproducible the matrix was split into two, with each smaller matrix containing different voltammograms from each sub-population. Eigenvalues and eigenvectors were generated using one matrix, and a scores plot was made as shown in Figure 3.26 (A). These same eigenvectors were used to generate the scores plot of the second matrix as shown in Figure 3.26 (B). The eigenvectors of the first data set form the axes of the new abstract space. The respective samples in the dat set are then plotted as scores on these eigenvectors. The assumption is then, that similar samples should project similarly onto the eigenvectors. Hence, by generating the scores of the second data set, and plotting these on the eigenvectors (or first two PCs) of the first data set, similar samples should cluster in the same grouping in both data sets. An alternative to splitting the data sets into two, would be to plot replicates of a data set to ascertain if they cluster similarly. However, the first method is more rigorous test of the validity of the model to cluster samples into respective classes.

In both plots the groupings due to the sub-populations fall into the same regions as before, and the overlay of the two plots regenerates the large pattern recognition plot (Figure 3.25).

Selection of the currents in the two major peak regions was then done on the voltammograms to attempt to separate the coffee and milk populations. By selecting the region between 100 and 400 mV, and excluding all alcoholic samples (beers , wines, liquors), a clear separation of sub-populations was observed with the desired separation of coffees and milks (Figure 3.27). For comparison, a scores plot containing only the non-alcoholic liquids was generated using the entire voltammogram, as shown in Figure 3.28. In this plot the coffees and milks still overlap. From the voltammograms - the entire scanned region shown in Figure 3.29 (A) and the variable selected region shown in Figure 3.29 (B) - the difference between samples is shown, with the largest differences occurring in the large peaks which would account for the separation observed in the scores plot for the entire voltammogram (Figure 3.28). To separate out the coffees and milks, more subtle differences found in the smaller peak are required and variable selection eliminates the overwhelming influence of the large peaks near -0.500 V. It appears that subtle differences between the non-alcoholic samples in the selected small peak region have sufficient variance to effectively separate out the sample classes.

The resulting plot (Figure 3.27) not only separated the coffees and the milks, but also separated the coffees according to their types such as flavored and non-flavored. From the voltammograms which compare the scans obtained for the different samples (Figure 3.30) it appears that the large peak shifts in position between the different liquids, but coffees and milks are closest to each other in terms of peak potential. When only the variable-selected portion is shown (Figure 3.30, B) peak differences are more distinct for the coffees and milks while they tend to converge for the beers and wines. This could explain why the coffees and milks separate into distinct sub-populations when variable selection is used, as the peaks are now more distinct.

In the presence of the large peak, these subtle differences are swamped, as the variance present due to the large peak effectively controls the resulting separations. An examination of the large peak region (Figure 3.29A) shows the differences between representative samples, and over a much larger scale than observed in the small peak region (B). It is obvious that the large peak region controls the subsequent variance determinations, which is corroborated when only the large peak region is used, the resulting scores plots give the same sub-population groupings as when the entire voltammogram is used.

Figure 3.24 Scores plot of the entire matrix.

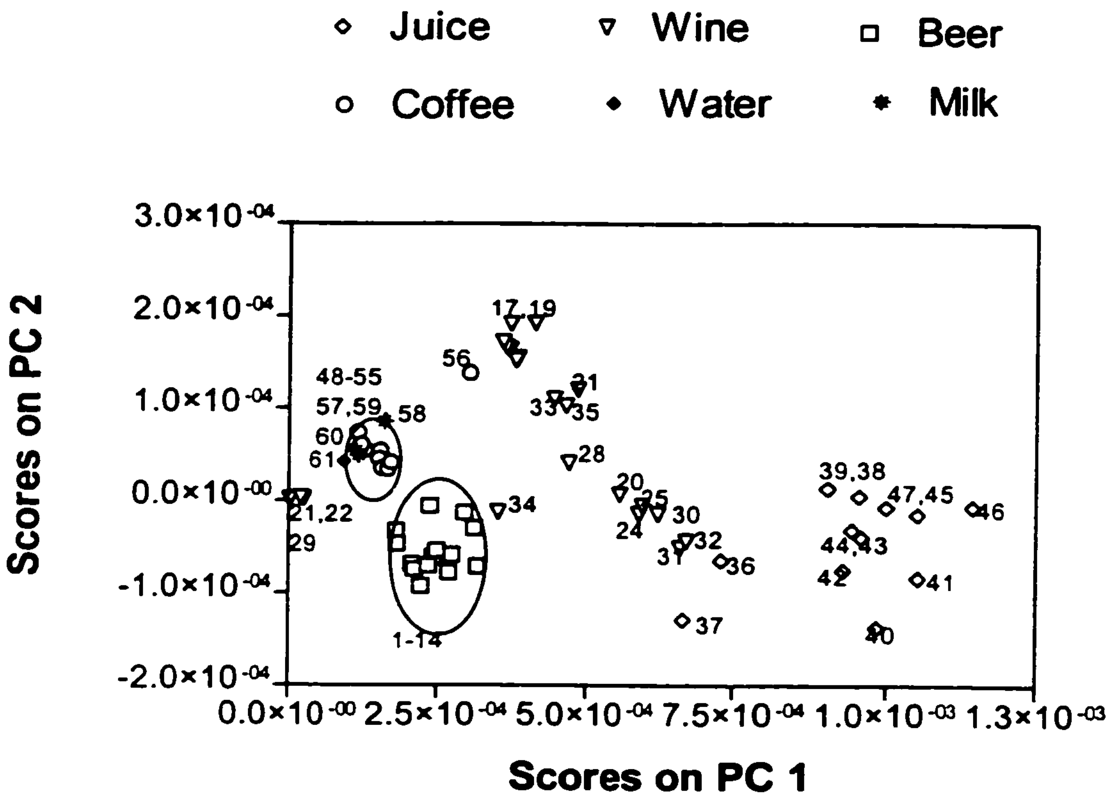


Figure 3.25 Scores plot of the entire matrix, excluding wines and liquors.

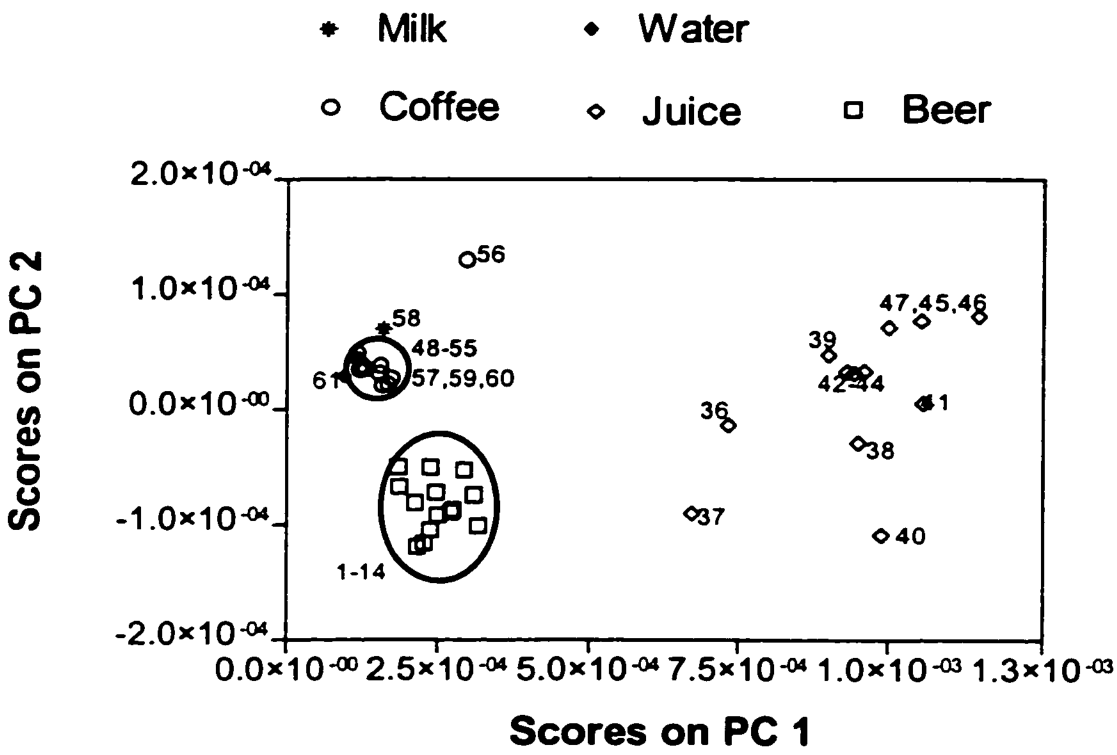


Figure 3.26 Validation plots of the two separate sub-sets generated by splitting the large data matrix into two separate matrices, generating a scores plot for one set (A), then using the eigenvectors of (A) to generate the scores plot for the second set (B).

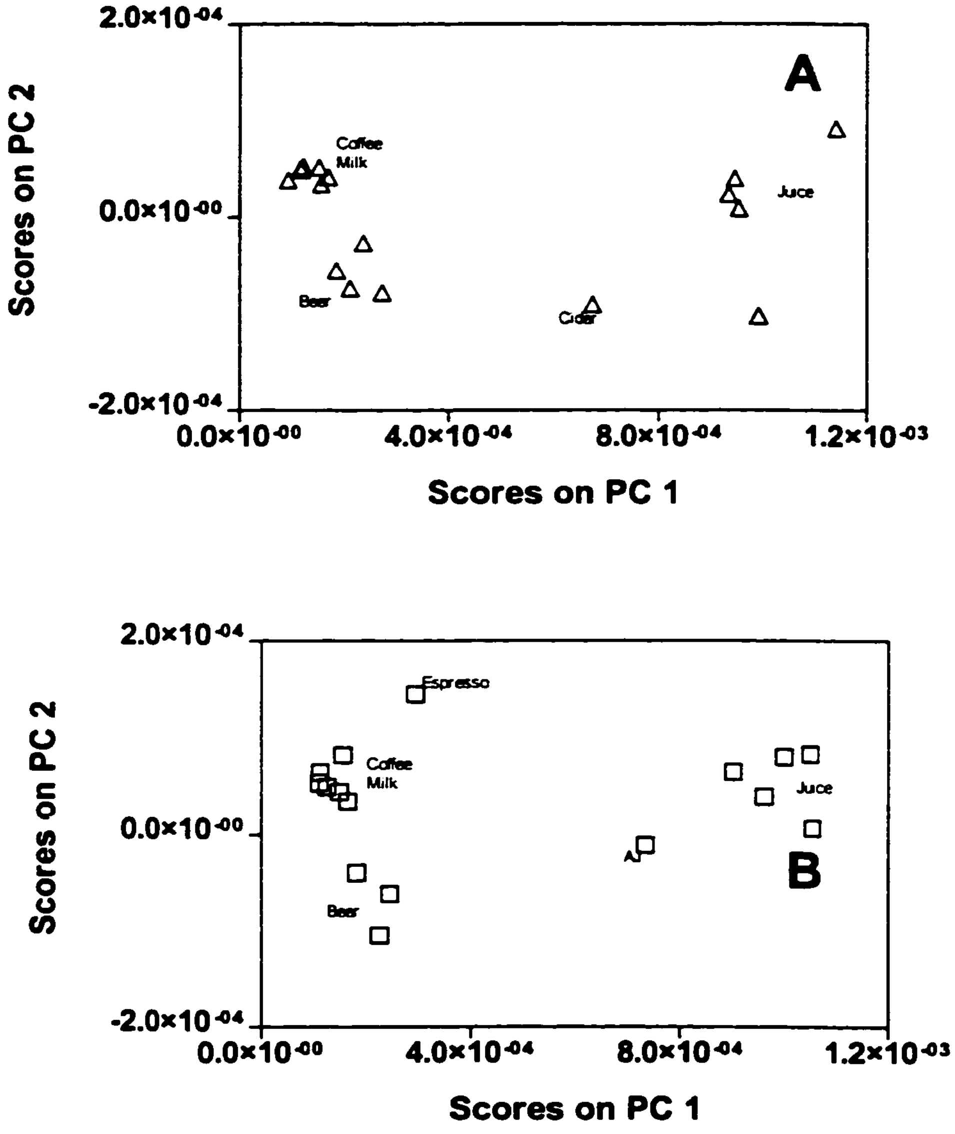


Figure 3.27 Scores plot generated from a matrix containing voltammograms for juices, coffee, milks and water, using selected potential range from 100 to 400 mV.

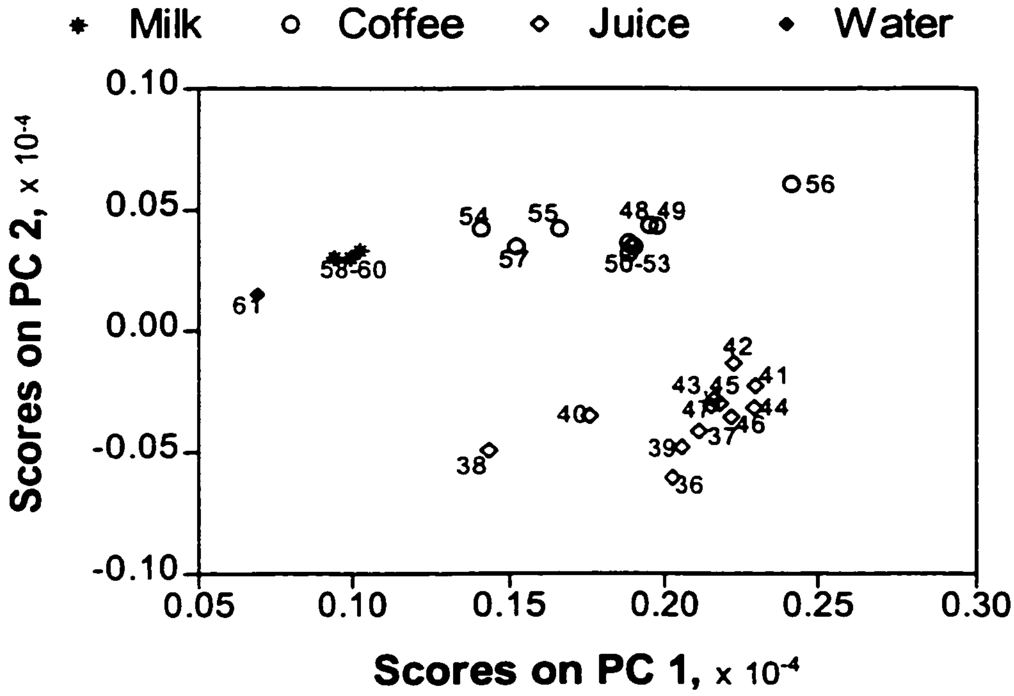


Figure 3.28 Scores plot generated from a matrix containing voltammograms for juices, coffee, milks and water, using the entire potential range from 1300 to -800 mV.

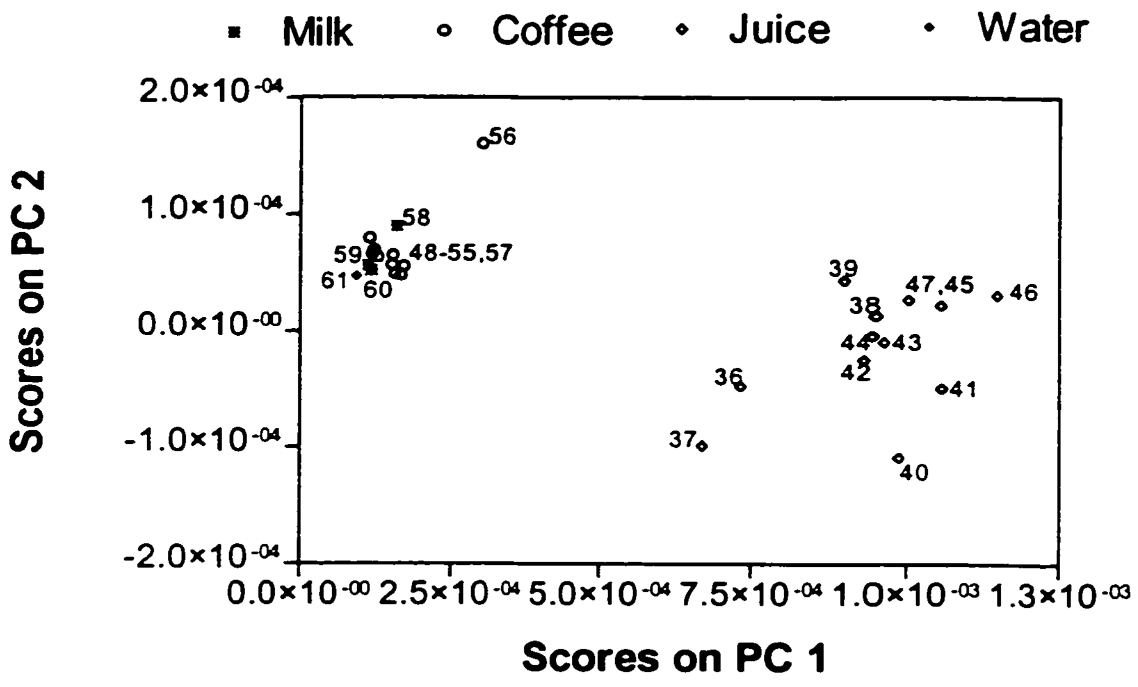
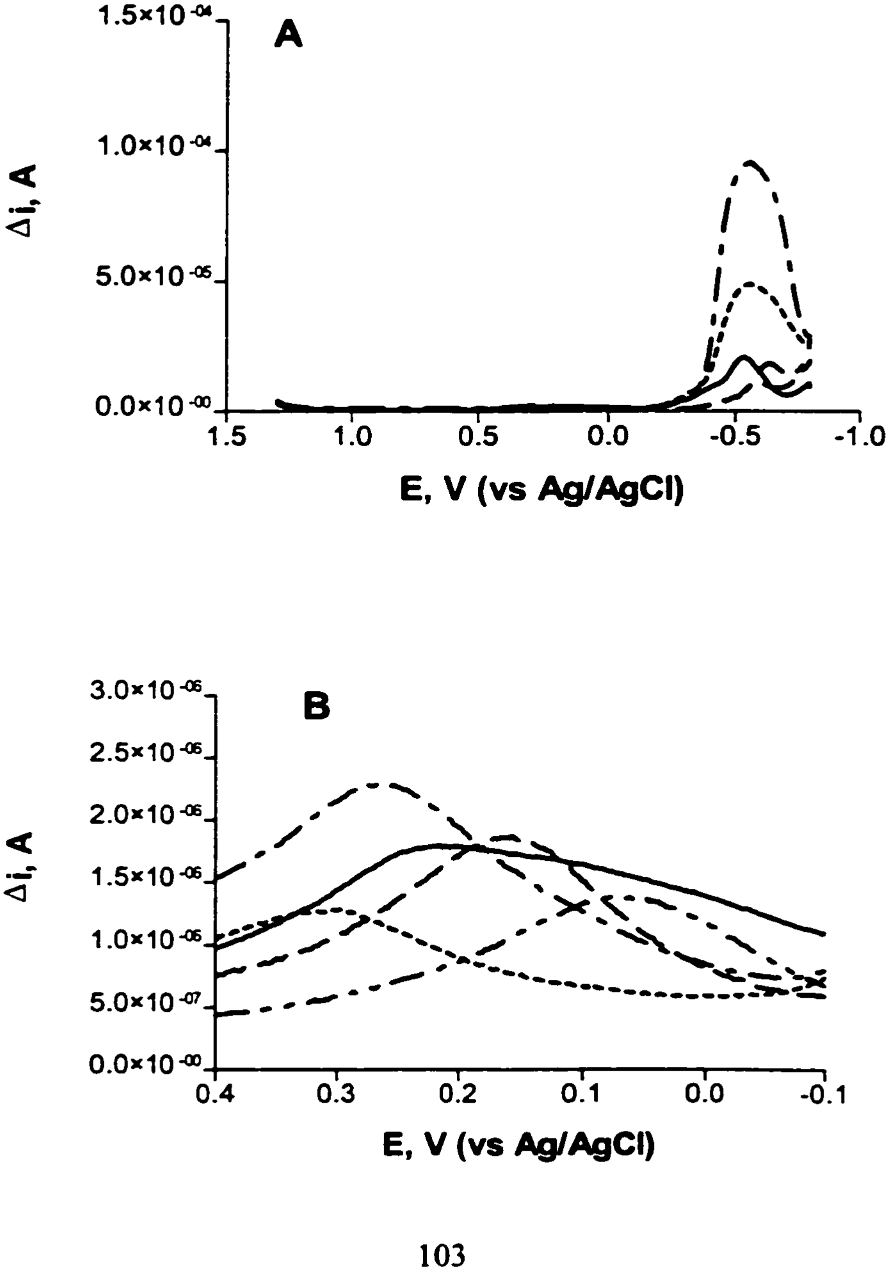


Figure 3.29 Square wave voltammograms using a Pt WE covering the entire potential range (A), and the selected region from 100 to 400 mV (B) for Waterloo Dark beer, Inniskillin Vidal red wine, Royal Blend coffee, 1% milk, and Minute Maid orange juice.



3.1.5 Bacterial Cultures

Postulating that it may be possible to differentiate microorganisms based on voltammograms recorded at different stages of growth, cultures *B. subtilis*, *S. aureus*, and *E. coli* were investigated. Both OD(600) readings and square wave voltammetric scans were recorded over time to generate growth curves at various points corresponding to the lag phase, exponential growth, and the plateau or stationary phase. The voltammograms were not distinguishable between species, or growth phases. For the resultant scores plot representative vectors corresponding to the iag, the growth, and the stationary phase for each of the bacterial cultures were plotted with 97.88% variance for PC 1 and 1.42% for PC 2. The scores plot does not show differentiation of the cultures or of the phases, although for *S. aureus* and *E. coli* an upward shift from lag, through growth to stationary phase was observed. Differentiation of species or growth phases was not possible on the scores plots, due to the indistinguishability of the representative voltammograms. Differentiation of the bacteria may be possible if the voltammograms of species were done at set stages of the growth curve, as species do show variance between each other at the different phases. Scores of the species at the different phases are separate, but a more rigorous study needs to be performed. This could involve the pre-separation of the individual bacterial species, or a pre-concentration step to increase the detectable signal, and lessen the signal from the matrix (mostly sugars, which could increase the signal significantly, and mask the bacterial signal). As well, repetitive scans, and an increased number of bacterial types, could further improve the resultant scores plots in terms of sub-population differentiation.

Chapter 4 Quantitative Analysis of Individual Analytes

Quantitative analysis of components in different complex media involved a multi-step procedure. After generation of the data set (split into calibration and validation), rank analysis was performed on the calibration set. Based on the optimal number of factors, the calibration solution was determined, and used in regression analysis to predict unknown analyte levels in the validation set. The validity and accuracy of the model was determined by calculating the root mean square error of calibration and prediction, the relative root mean square error based on the range (RRMSE), and the correlation coefficient between predicted and known analyte levels.

This Chapter presents the early experiments and data analyses done to determine if the combination of square wave voltammetry and factor analysis was an effective combination for the prediction of analyte levels in a complex liquid. Determination of biomass in bacterial cultures was an extension of the work presented in the previous Chapter (qualitative analysis of bacteria). The early series of experiments done, using the addition of a phosphate buffer to orange juice, were performed to assess the predictive capability of PCR and PLS. Following the success of these experiments, the determination of glucose, lactate, pyruvate and ethanol in beer involved a more rigorous examination of the capabilities of these models.

4.1 Results and Discussion

4.1.1 Determination of Biomass in Bacterial Cultures

Three microorganisms (*Staphylococcus aureus, Bacillus subtilis, Escherichia coli JM105)* were grown separately in shake flasks and monitored by turbidity, or optical density, at 600 nm. Square wave voltammetry was performed on samples taken at the same time, using the same conditions for optimal pattern recognition as described in Chapter 3. Turbidity (OD_{600} values) were used as the expected or true values. Voltammograms were analysed by factor analysis to determine the most significant factors, which were used to generate the calibration solution. The predicted turbidity values were compared to the measured turbidity values, and the error in slope deviation from 1.000 was obtained. Further, both the expected and the predicted OD_{600} values were plotted versus time to obtain the usual growth curve, and show a close correlation between the predicted and the expected values.

To test the ability of the calibration model to predict the turbidities corresponding to unknown voltammograms, the matrix for each culture was divided, with 6 vectors used for the calibration set and the remainder (4) used for the validation set. Again, after factor analysis to determine the optimal number of factors to be used for lowest prediction errors, the predicted OD_{600} values for both the calibration and the prediction sets were obtained, and are plotted on the growth curves as shown in Figures 4.1 to 4.3. Five factors were optimal for predicting unknown OD_{600} values for both PCR and PLS analyses, based on prediction errors from cross-validation. From this, predictions of turbidity were made for the remaining four vectors which had been taken out of the original matrices and used in the validation set.

The resulting predictions, along with those found from the calibration set and those determined experimentally are given in Tables 4.1 to 4.3. Part A of the tables gives the OD_{600} values and part B give the resulting correlation coefficients and the 95% confidence interval (interval within which 95% of the results lie).

Results from Chapter 3 from the qualitative determination of bacterial cultures at various stages of the growth had been inconclusive. The scores plotted on the first two factors were inadequate to accurately predict the growth stage of the bacteria. By performing a rank analysis, and predicting turbidity values based on the higher number of optimal factors, either PCR or PLS might be used for prediction of bacterial growth stages. This preliminary study indicated promise for the use of multivariate data analysis for biomass quantitation as OD_{600} predicted values were closely correlated to the measured turbidity values. Examination of the RMSE of calibration or prediction gave small errors, but a relative RMSE which varied depending on the bacterial type. This makes the predicted values accurate, but varying over a large range (to 15% RRMSE), hence affecting the precision. Due to the small data set (6 calibration samples) used to develop the calibration solution, and the retention of five factors from cross-validation making this multilinear regression, rather than PCR, it was difficult to assess the ability of PCR to accurately and precisely predict analyte levels. While multivariate methods appear to be promising in prediction, a larger data set was required. Hence, a different media (orange juice) was analysed to predict adulteration of the sample using PCR. This media, while still complex in its matrix, does not contain as many uncertain variables as a fermentation broth, and was used to assess the potential reliability of multivariate calibration solutions for further work.

Table 4.1 Determination of OD values of *E.coli* from UV-Vis readings at 600 nm, and predicted OD values from PCR of the calibration and validation set.

Time, min.	OD(600 nm)	OD(calibration)	OD(validation)
67	0.135	0.145	
186	0.521		0.433
210	0.735	0.732	
236	0.844		0.793
264	1.187	1.149	
294	2.001	2.063	
334	2.775		2.880
359	3.174	3.081	
384	3.431		3.425
439	3.904	3.669	

Table 4.1 (B) Error Analysis

E.coli	OD(cal)	OD(val)
Pearson r	0.952	0.988
RMSE	0.263	0.073
RRMSE, %	7.465	2.444

Table 4.2 (A) Determination of OD values of *B.subtilis* from UV-Vis readings at 600 nm, and predicted OD values from PCR of the calibration and validation set.

Time, min.	OD(600 nm)	OD(cal)	OD(val)
60.0	0.1767	0.1794	
95.0	0.1938		0.1625
150.0	0.2538	0.2495	
195.0	0.4284	0.4338	
240.0	0.6252		0.661
270.0	0.8764	0.8578	
300.0	1.127	1.1534	
355.0	1.2389		1.0746
400.0	1.3027	1.2901	
445.0	1.4559		1.1752

Table 4.2 (B) Error Analysis

B.subtilis	OD(cal)	OD(val)
Pearson r	0.958	0.987
RMSE	0.035	0.164
RRMSE, %	3.192	16.229

Table 4.3 (A) Determination of OD values of *S.aureus* from UV-Vis readings at 600 nm, and predicted OD values from PCR of the calibration and validation set.

Time, min.	OD(600 nm)	OD(cal)	OD(val)
60.0	0.2029	0.2859	
95.0	0.2191		0.1768
150.0	0.4152	0.4388	
195.0	0.6196	0.6542	
240.0	0.8161		0.8011
270.0	1.114	1.119	
300.0	1.4169	1.186	
355.0	1.6726		1.6687
400.0	1.9098	2.0147	
445.0	2.003		1.9345

Table 4.3 (B) Error Analysis

S.aureus	OD(cal)	OD(val)
Pearson r	0.970	0.990
RMSE	0.270	0.041
RRMSE, %	15.627	2.332

Figure 4.1 Growth curve for *E.coli JM 105* showing OD values found from UV-Vis spectroscopy, and obtained from prediction of calibration and validation data sets using PCR.

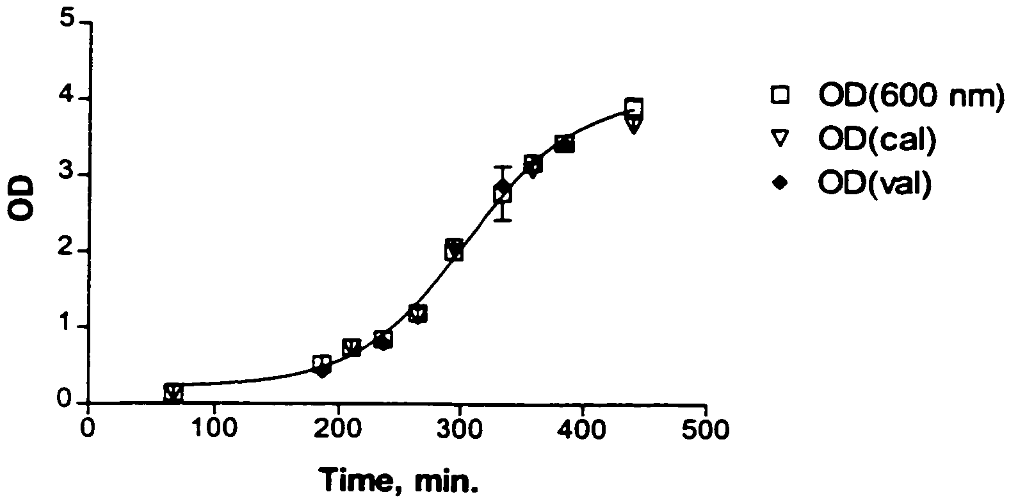


Figure 4.2 Growth curve for *B.subtilis* showing OD values found from UV-Vis spectroscopy, and obtained from prediction of calibration and validation data sets using PCR.

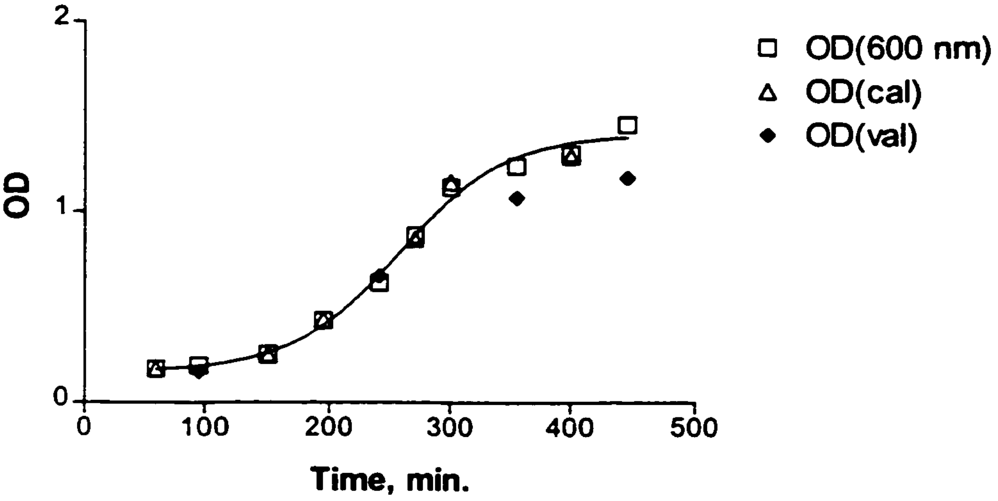
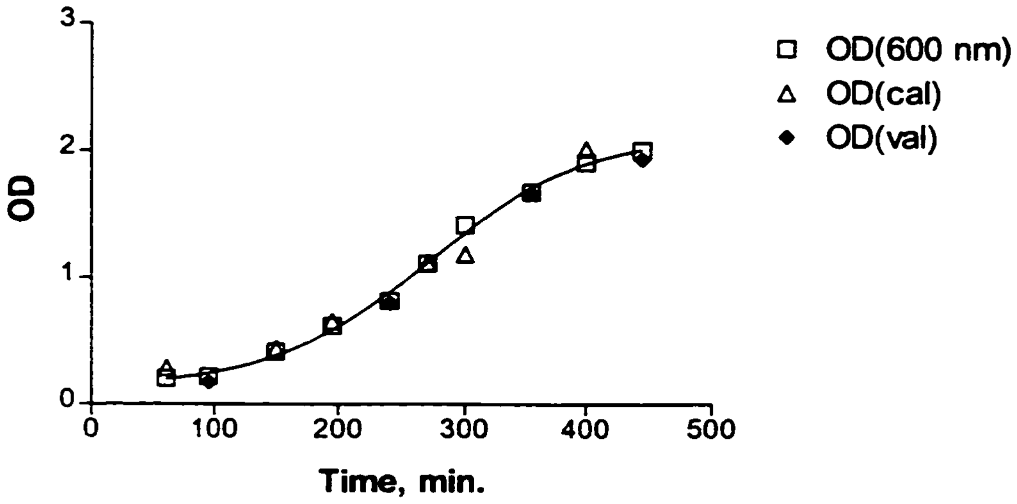


Figure 4.3 Growth curve for *S.aureus* showing OD values found from UV-Vis spectroscopy, and obtained from prediction of calibration and validation data sets using PCR.



4.1.2 Orange Juice Dilution Studies

Successive dilutions of Tropicana Pure Premium Orange Juice with 0.050 M phosphate buffer (pH=7.02) were performed , and square wave voltammograms of each dilution were generated. These were divided into the calibration (16 points) and validation (5 points) data sets. Validation points were taken randomly from the original data set, and covered the range of possible values. The calibration data set is given in Table 4.4, and validation data in Table 4.5. From cross-validation, it was determined that 4 factors would minimize the PRESS, as the cross-validation errors drop sharply at this number of factors. After 4 factors, very little variation is observed in the errors, indicating that these factors contain mostly noise. After generating the calibration model (Figure 4.4A), prediction of the volume fractions of orange juice (Figure 4.4B) was done resulting in a prediction error of 2.1% as deviation from a perfect slope.

Studies of orange juice dilutions showed good correlation between the predicted and expected volume fractions in the validation set, with correlation coefficients of 0.987 (PCR) and 0.978 (PLS). From the plot of the resulting predictions (Figure 4.4), only slight deviations are observed from an ideal slope of 1.000. Further, high correlation between the expected and predicted dilutions was obtained, as correlation coefficients were greater than 0.98 (PCR) for prediction of the validation set. This was reinforced by the low errors calculated from RMSEP, and the relative precision, where the deviations of the predicted dilutions varied from 3 to 5% (RRMSEP) only. From this preliminary work, either PCR or PLS show high predictive ability of the calibration solution, with both good accuracy and precision. From this, the potential reliability of either PCR and PLS for further use was promising, and these methods were tested for the predictive capability of common individual analytes in another complex medium, PC Brew.

A common fermentation process which produces a complex liquid is the production of beer. The process proceeds from conversion of glucose by *S. cerevisiae* to the end-product of ethanol. During the glycolysis process (described in Chapter 2), pyruvate or lactate may also be produced. The end result is a complex liquid which not only contains glucose, ethanol, pyruvate and lactate, but myriad other species. Thus, accurate measurement of the desired components can be difficult if other species interfere or signals are masked by noise. If the combination of square wave voltammetry, and PCR could prediction levels of these analyte in this complex broth, it would provide a fast, reliable method for process control.

Table 4.4 (A) Predicted dilutions of orange juice, diluted with phosphate buffer using PCR and PLS (four factors retained) for prediction of calibration data.

Actual Dilution Value	Predicted- PCR	Predicted-PLS
1.000	1.029	1.027
0.909	0.857	0.867
0.830	0.813	0.811
0.769	0.786	0.786
0.714	0.710	0.708
0.667	0.679	0.679
0.630	0.621	0.616
0.556	0.577	0.576
0.500	0.520	0.521
0.417	0.393	0.394
0.357	0.348	0.349
0.333	0.343	0.342
0.313	0.315	0.315
0.278	0.290	0.292
0.250	0.244	0.244
0.217	0.212	0.211

Table 4.4 (B) Error Analysis

Calibration Set	PCA	PLS
Slope	0.995±0.022	0.996±0.020
Pearson r	0.997	0.997
RMSE	0.023	0.021
RRMSE, %	2.781	2.560

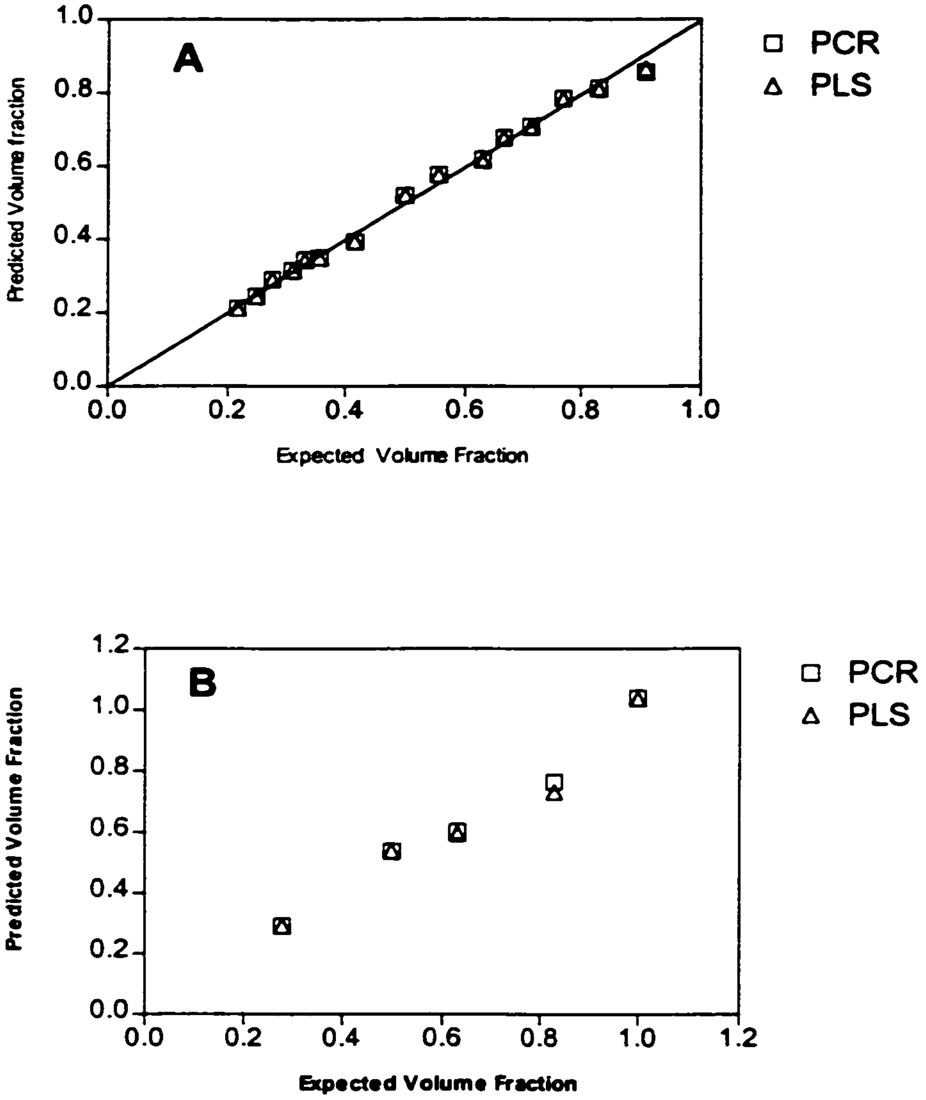
Table 4.5 (A) Predicted dilution values of orange juice from the validation set using either PCR or PLS on the validation data set.

Actual Dilution Value	Predicted-PCR	Predicted-PLS
1.0000	1.0387	1.0364
0.8300	0.7646	0.7301
0.6300	0.6019	0.5946
0.5000	0.5364	0.5381
0.2800	0.2903	0.2915

Table 4.5 (B) Error Analysis

Validation Set	PCA	PLS
Slope	0.998±0.091	0.954±0.12
Pearson r	0.987	0.978
RMSE	0.026	0.034
RRMSE, %	3.449	4.609

Figure 4.4 Plots of predicted volume fractions of Tropicana Pure Premium Orange Juice dilute with buffer. Predicted results obtained from PCR and PLS, for the calibration data (A) and validation data (B).



4.1.3 Quantitation of Individual Chemical Constituents

For these series of experiments, President's Choice De-Alcoholized Brew (0.5% alcohol) was used exclusively, and addition of either ethanol, lactate, pyruvate or glucose into the PC brew was done. A complex media, PC Brew de-alcoholized Beer, allowed a more rigorous evaluation of the use of factor analysis for quantitation. After each addition, square wave voltammetric scans were generated in triplicate, the data were averaged, and imported into a spreadsheet format. Each analyte matrix was divided into two: the larger matrix contained from 9 to 13 columns and was used to produce the calibration model, while the smaller matrix contained 5 columns and was used as the validation or prediction set.

Data analysis consisted of transforming the data set into the abstract solution, selection of the number of latent variables by plotting the reduced eigenvalues present at a factor number and by determination of the PRESS, then prediction of analyte levels in the validation sets. Rank analysis was performed using the calibration data set. A representative example of rank analysis is presented for ethanol.

Increasing additions of ethanol to PC Brew were used to generate a calibration data set (9 points), with the validation data set collected one week later (8 points). A plot of the eigenvalues and reduced eigenvalues at a given number of factors, generated using PCA are shown in Figure 4.5. Very little change in variance is observed, with a slight change in slope at factor 3. This is more noticeable in the plot of reduced eigenvalues (A). A further indicator function was used to determine optimal rank; cross-validation using the calibration data was done. The resultant analysis (Figure 4.6) suggests four factors may generate the lowest PRESS on the calibration set. Both PCA and PLS results are shown, and as observed very little change in slope is observed for either past two factors, indicating that predictive ability may suffer when this model is used. Further it is observed that at 8 factors for PCR, a large jump in PRESS occurs, as when all the factors are included, the experimental error is added back in. Errors in the predictive capability of PCA are larger than for PLS. This could be due to the ability of PLS to minimize angles between the X- and Y-data sets, so as to optimize congruency. Thus, PLS attempts to optimize the fit between the data and the concentration matrices, while PCA does not. Subsequent residual for four factors, gave only noise indicating that the significant information was contained in the first four factors. Due to the small change in slope, making selection of the optimal number of factors difficult, residual analysis was also done for two factors. The plot of the residuals showed retention of the voltammetric signal, indicating that at two factors retained, signal would be discarded, and the resultant calibration model would not be optimized. Therefore, four factors were retained, and this model was used in subsequent regression calculations. This model was used in the subsequent PCR to regenerate the calibration ethanol levels, and to predict ethanol levels in the validation data set. Comparison of the predicted values were then made to the actual values. These results are presented in Table 4.6 (calibration data) and Table 4.7 (validation data), and plotted in Figure 4.7.

These Tables also show results predicted using PLS. Determination of the number of factors to retain for PLS was done in a similar manner as for PCA, except that eigenvalues generated using PLS are termed pseudo-eigenvalues, and provide a measure of the variance spanned by the given latent variable. From residual analysis, when four latent variables were chosen, only noise was observed. Therefore, prediction of ethanol levels using PLS was also done using the four factors (or latent variables in PLS).

Further, data manipulation of the voltammograms to select out the two potential regions that correspond to the two peaks was done. This type of selection was tested to improve correlations between predicted and expected results by choosing which peaks in the voltammograms would be analysed by both PCR and PLS. The small peak in the voltammogram (100 to 400 mV unless otherwise noted) was used for lactate, pyruvate and glucose studies, while the large peak (-100 to -800 mV) was used for the ethanol matrix. The choice of the small peak for the non-alcoholic chemicals, and the large peak for the ethanol was due mostly to the results of Chapter 3, where variable selection of the small peak allowed good separation of the non-alcoholic populations. Because the large peak contributes the majority of the variance to the factor analysis, selection of the small peak provides an idea of whether a smaller voltammetric range is adequate for quantitation by factor analysis.

For the analysis of ethanol percentages, based on the plot of reduced eigenvalues, cross-validation and PRESS, four factors contain the meaningful information, while the remaining factors mostly contribute noise. So, based on four factors, the prediction values were generated for the calibration and the validation sets and are shown in Table 4.6 and Table 4.7, respectively. The results are also plotted in Figure 4.7.

From the correlation coefficients, there is greater correlation between expected and predicted values when the entire voltammogram is used. When voltammograms containing only the large peak are used in subsequent calibration and prediction, a decrease in correlation (from 0.98 drops to 0.91) is observed. The poor correlation observed when only selected regions of the voltammogram are used is illustrated in Figure 4.7. In the validation set (B), deviation from the ideal slope of 1.000 is obvious for the selected portions, indicating poor predictive ability of PCR and PLS for the selected ethanol levels. PCR on the entire voltammogram gave the best fit between expected and predicted values, with a correlation coefficient of 0.976. As well, prediction results obtained using PLS have poorer correlations than PCR, indicating that for this set of analyses, PCR is the more accurate method to predict ethanol levels.

From analysis of the validation results, the RMSEP and the RRMSEP, the poor predictive ability shown from the plots is evident from the high relative error over the range, from 17% to 67%. This shows the poor precision evident in the plots, which tend to deviate from the ideal slope. Given the error between the predicted values determined from the factor-based solution and the actual values (as shown in the RMSEP), accuracy of prediction is also a problem. PCR (RMSEP=1.467%) gave the lowest predictive error, while PLS (RMSEP=3.786%) obviously has poor accuracy of prediction. Both of the selected results were unable to predict ethanol levels either accurately or precisely. Possibly, the difficulty in correctly choosing the optimal number of factors explains this lose in accuracy, since the RMSE is a measure of the difference between actual and factor-based predictions, and as it was difficult to correctly identify the optimal number of factors, an increase in retaining the experimental error would result in a poorer predictive solution.

Table 4.6 (A) Prediction of ethanol percentage in PC Brew using PCR and PLS to predict on the calibration set, and using variable selection (vs from -0.100 to -0.800 V).

Actual, %	PCR	PLS	PCR(vs)	PLS(vs)
0.500	0.459	0.514	-0.382	-0.131
3.400	2.813	3.118	4.534	4.237
4.700	4.999	4.945	5.179	4.906
6.000	6.542	5.917	6.717	6.700
7.200	7.063	7.245	7.926	7.955
8.300	8.980	8.500	10.619	10.695
9.400	9.828	9.547	7.387	7.465
10.500	10.073	10.535	8.979	9.051
11.500	10.716	11.174	10.045	10.186

Table 4.6 (B) Error Analysis

Ethanol-Calibration	PCR	PLS	PCR (vs)	PLS (vs)
Pearson r	0.989	0.998	0.912	0.921
RMSE, %	0.662	0.252	1.854	1.770
RRMSE, %	6.454	2.365	16.856	16.267

Table 4.7 (A) Prediction of ethanol percentage in PC Brew using PCR and PLS to predict on the validation set, and using variable selection (vs from -0.100 to -0.800 V).

Actual, %	PCR	PLS	PCR (vs)	PLS (vs)
0.500	-0.615	6.214	3.868	3.380
3.360	2.208	7.342	6.115	6.073
3.900	2.287	6.303	4.999	4.865
4.940	4.472	9.612	8.991	8.867
5.950	5.471	9.960	8.490	8.381
7.400	6.315	10.919	9.940	9.950
8.780	7.372	11.813	8.938	8.922
10.500	7.719	11.461	11.562	11.776

Table 4.7 (B) Error Analysis

Ethanol- Validation	PCR	PLS	PCR (vs)	PLS (vs)
Pearson r	0.976	0.915	0.918	0.926
RMSE, %	1.467	3.786	2.515	2.400
RRMSE, %	17.240	67.618	32.687	28.582

Figure 4.5 Contribution of the reduced eigenvalues (A) and eigenvalues (B) to total variance based on the number of factors in PCA and PLS. PLS eigenvalues are “pseudo-eigenvalues”.

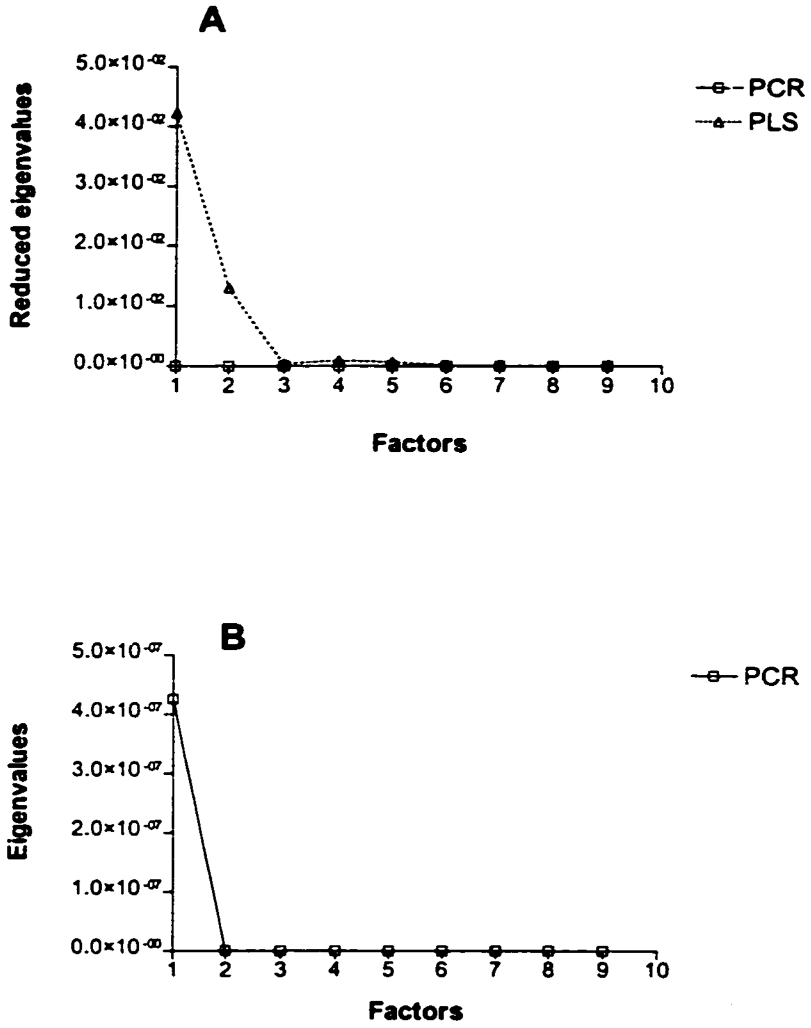


Figure 4.6 Predictive errors obtained from cross-validation of the calibration set. Results are for ethanol.

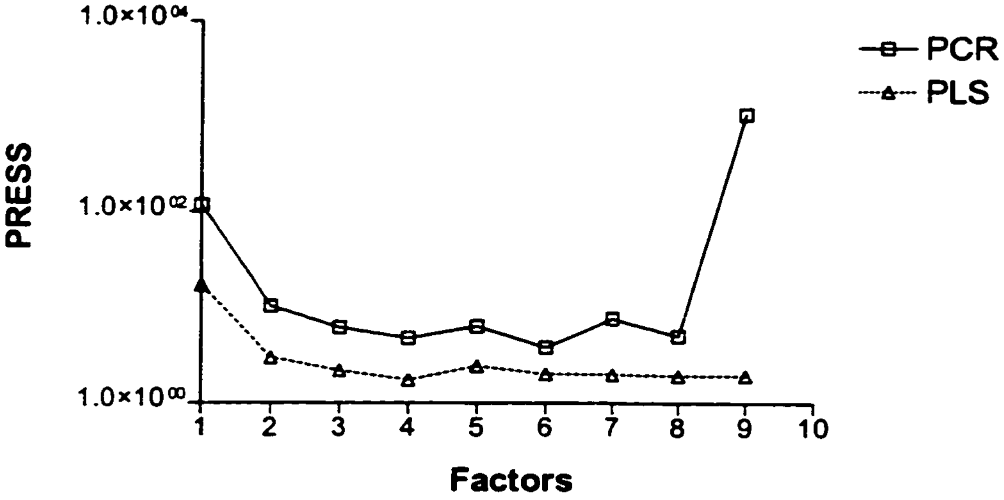
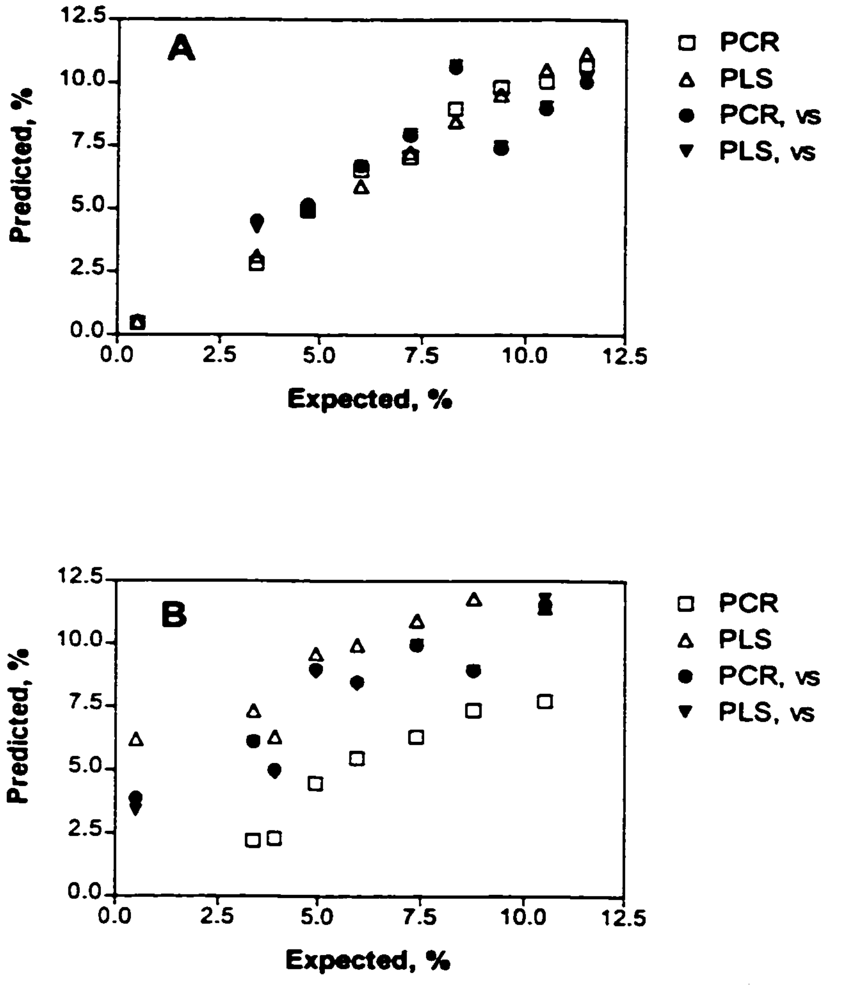


Figure 4.7 Plot of predicted ethanol values obtained for the calibration (A) and validation (B) data sets using PCR and PLS. Results shown for entire voltammograms and selected region (vs=-100 to -800 mV)



A similar series of experiments were performed with the addition of lactate to a fresh sample of PC Brew. Voltammograms obtained for the eight calibration, and the five validation data sets were used either in their entirety, or the currents corresponding to the potential region between 0.100 and 0.400 mV were selected. Based on rank analysis of the calibration set, six factors were found to contain the significant information (lowest PRESS from cross-validation), and predicted values of lactate concentration for the calibration and validation set were determined. The results are given in Table 4.8 for the calibration set and Table 4.9 for the validation set. From the calibration set analysis, it appears that using the entire voltammogram gives higher accuracy and improved precision with correlations greater that 0.99 for both PCR and PLS. Plots of these results (Figure 4.8) bear out the excellent correlation observed, as results appear classically linear for the results obtained over the entire voltammogram. Figure 4.9, the variable selected data, still retains good linearity (about 0.96 overali), with good correlation (about 0.98). However, subsequent experiments utilized the entire voltammogram in the resulting matrices due to the slightly higher correlations obtained, and the relative ease of using an entire voltammogram over pre-seiecting out regions.

The excellent correlation obtained was also evident from the RMSE calculations, with RMSEP showing the high accuracy of the calibration solution (both PCR and PLS had RMSEP of 0.002 M). Given the RRMSEP of 1.453% for PCR and PLS, good precision and accuracy of the predicted results was observed. The selected region also gave good, but slightly higher, predictive errors over a narrow range (RMSEP about 0.01 M).

Table 4.8 (A) **Prediction of lactate concentrations from calibration set data using** PCR and PLS and also having variable selection (vs from 100 to 400 mV).

Actual, M	PCR	PLS	PCR (vs)	PLS (vs)
0.237	0.238	0.237	0.228	0.235
0.215	0.214	0.215	0.204	0.206
0.169	0.167	0.167	0.174	0.170
0.140	0.144	0.142	0.155	0.151
0.103	0.102	0.102	0.120	0.115
0.085	0.085	0.085	0.066	0.070
0.055	0.055	0.056	0.046	0.047
0.045	0.045	0.045	0.060	0.058

Table 4.8 (B) Error Analysis

Lactate Calibration	PCR	PLS	PCR (vs)	PLS (vs)
Pearson r	1.000	1.000	0.980	0.988
RMSE, M	0.004	0.002	0.026	0.020
RRMSE, %	1.824	1.189	14.420	10.730

Table 4.9 (A) Prediction of lactate concentrations from validation set data using PCR and PLS and selecting for variable selection (vs from 100 to 400 mV).

Actual, M	PCR	PLS	PCR (vs)	PLS (vs)
0.198	0.199	0.198	0.200	0.198
0.158	0.161	0.160	0.179	0.174
0.125	0.121	0.121	0.150	0.142
0.099	0.097	0.097	0.103	0.101
0.051	0.051	0.051	0.040	0.043

Table 4.9 (B) Error Analysis

Lactate Validation	PCR	PLS	PCR (vs)	PLS (vs)
Pearson r	0.999	0.999	0.979	0.986
RMSE, M	0.002	0.002	0.015	0.011
RRMSE, %	1.453	1.453	9.616	7.276

Figure 4.8 Results obtained using PCR and PLS to predict lactate concentrations of the calibration set (A) and the validation set (B). Results are from using the entire voltammogram.

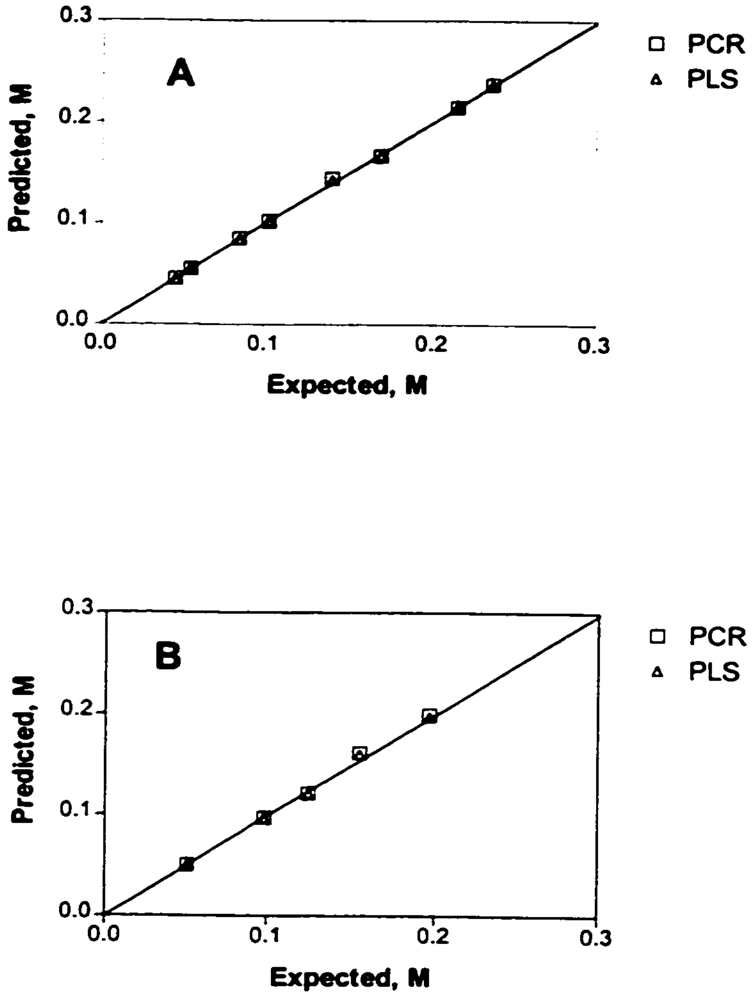
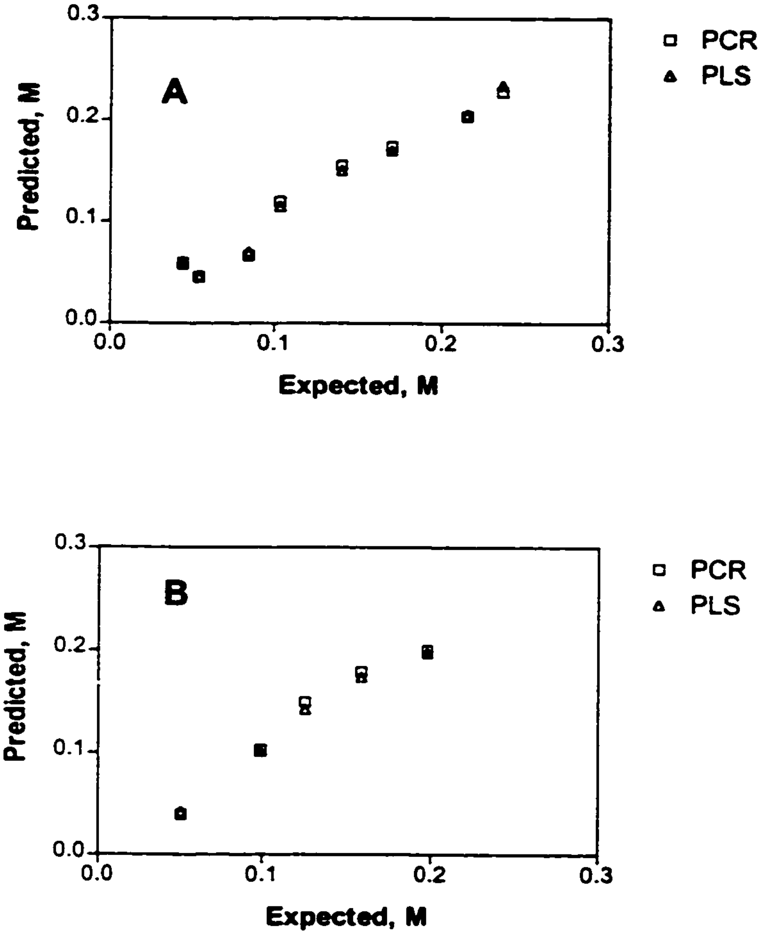


Figure 4.9 Results obtained using PCR and PLS to predict lactate concentrations of the calibration set (A) and the validation set (B). Results are from using selected potential regions of the voltammogram (vs=100 to 400 mV).



Similarly, voltammograms were obtained for addition of pyruvate to PC Brew, and these were divided into the calibration (10 points) and the validation (5) set. Again, prediction of pyruvate concentration was done using the entire voltammogram and the selected region from 100 to 400 mV. After determination that four factors would contain the meaningful information from cross-validation of the calibration set, the calibration and the validation prediction results were found (Table 4.10 and 4.11). These results again indicate that poorer correlations are seen when the variable selected region is used. Using the entire voltammogram, factor analysis and subsequent prediction of the validation results indicate that both PCR and PLS done on the entire voltammogram gave the best correlation (0.977 and 0.954, respectively) and were relatively precise This is shown on the correlation plot of Figure 4.10 (B) which showed good agreement between the expected and predicted concentrations of the validation set. Due to the large deviations observed from the selected regions, these were not plotted.

Analysis of the RMSEP shows the good agreement between predicted and actual values, generated using this calibration solution over the entire voltammogram. These values range over about 10% from expected, so some loss of precision was obtained. However, the model was adequate to predict pyruvate levels in the media. RMSEP values obtained for the selected regions show the poor accuracy and precision with RMSEP approximately double and a range of predicted errors over 24%. As seen in previous experiments, no improvement in predictive capability of the calibration solution was observed by using only a portion of the voltammogram.

Table 4.10 (A) Prediction of pyruvate concentration of the calibration set using PCR and PLS (4 factors) and comparing to prediction of concentrations calculated from variable selection of the voltammogram (vs from 0.100 to 0.400 V).

True, mM	PCR	PLS	PCR (vs)	PLS (vs)
21.070	20.965	21.173	21.016	21.193
28.525	28.355	28.655	26.570	27.068
26.804	25.542	25.683	25.871	26.963
26.040	26.447	26.197	26.207	27.117
25.212	25.700	25.664	26.916	25.447
24.459	24.155	24.197	22.743	23.199
23.830	24.994	24.790	25.697	24.925
22.680	22.237	22.213	22.260	22.237
21.910	22.155	21.959	21.568	21.837
21.786	21.721	21.743	23.279	22.270

Table 4.10 (B) Error Analysis

Pyruvate Calibration	PCR	PLS	PCR (vs)	PLS (vs)
Pearson r	0.964	0.974	0.835	0.936
RMSE, mM	0.790	0.674	1.662	1.049
RRMSE, %	10.690	9.008	28.175	17.706

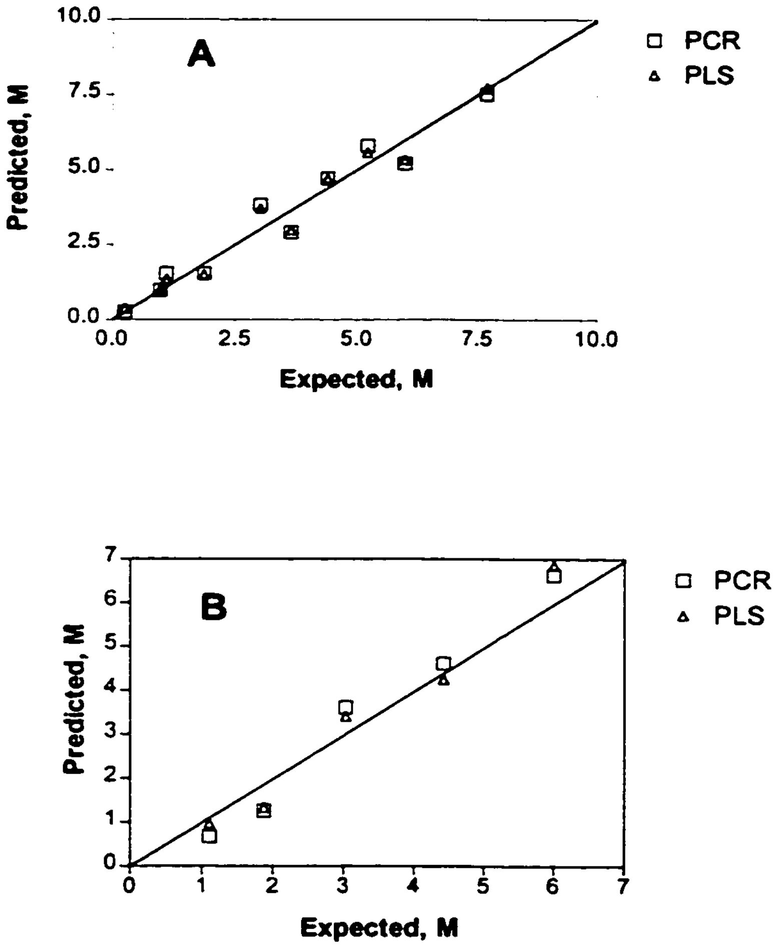
Table 4.11 (A) Prediction of pyruvate concentration of the validation set using PCR and PLS (4 factors) and comparing to prediction of concentrations calculated from variable selection of the voltammogram (vs from 0.100 to 0.400 V).

Actual, mM	PCR	PLS	PCR (vs)	PLS (vs)
27.280	27.503	27.815	25.422	25.637
25.560	25.181	25.051	26.109	25.781
24.310	23.933	23.669	24.385	23.899
22.820	22.052	22.119	22.928	22.671
21.950	22.350	22.654	22.405	22.397

Table 4.11 (B) Error Analysis

Pyruvate Validation	PCR	PLS	PCR (vs)	PLS (vs)
Pearson r	0.977	0.954	0.900	0.944
RMSE, mM	0.466	0.623	0.892	0.792
RRMSE, %	8.547	10.945	24.081	23.415

Figure 4.10 Results obtained using PCR and PLS to predict pyruvate concentrations for calibration (A) and validation (B) data sets.



As for lactate and pyruvate, square wave voltammograms were obtained for addition of glucose to PC Brew, to give the calibration (10 points) and validation (5 points) data sets. From cross-validation of the calibration data set, a possible five factors would be optimal for predictive capability. Again, the predicted results of using the entire voltammogram are compared to those obtained using the selected region from 100 to 500 mV (Table 4.12 and 4.13). From the correlation coefficients obtained, better correlation is obtained when the entire voltammogram is used, as the correlation coefficients for the selected region are half of the entire region results (0.928 and 0.495, PCR). While the correlation coefficients shown for PCR (0.928) and PLS (0.933) using the entire voltammogram, are good, a plot of the expected and predicted results in Figure 4.11 reveals significant problems with glucose prediction of the validation set when only a selected portion of the voltammogram was used. Very poor precision is observed, as significant deviations from linearity occur. This was clear from the RRMSEP which ranged from 57 to 62%, indicating that the calibration solution is inadequate to accurately predict glucose concentrations.

Conversely, when the entire voltammogram was used, the model was adequate, with a RMSEP of 0.5 mM for both PCR and PLS, and a RRMSEP of about 14%. Though the range of predicted values do suffer from precision, based on the RRMSEP, the model was still adequate to accurately predict glucose values.

Table 4.12 (A) **Predicted glucose concentration using PCR and PLS and comparing to variable selected (vs from 0.1 to 0.4 V) regions, using the calibration data.**

Actual, mM	PCR	PLS	PCR (vs)	PLS (vs)
16.580	16.237	16.319	16.055	16.652
16.040	16.087	16.083	15.637	15.935
14.870	15.180	15.077	14.303	14.975
13.920	13.931	13.966	13.881	13.621
13.310	12.807	12.954	14.955	13.447
12.900	13.171	13.135	13.393	13.019
12.460	12.757	12.711	12.482	12.399
11.970	12.011	11.974	10.925	11.924
11.650	11.558	11.521	11.620	11.679
10.590	10.555	10.551	10.924	10.641

Table 4.12 (B) Error Analysis

Glucose Calibration	PCR	PLS	PCR (vs)	PLS (vs)
Pearson r	0.990	0.994	0.925	0.998
RMSE, mM	0.358	0.275	0.991	0.178
RRMSE, %	6.306	4.770	19.323	2.966

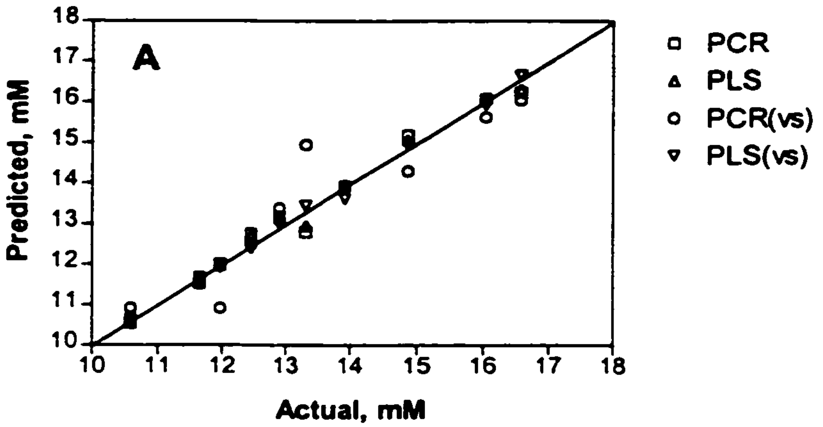
Table 4.13 (A) **Predicted glucose concentration using PCR and PLS and comparing to variable selected (vs from 0.1 to 0.4 V) regions, using the validation data.**

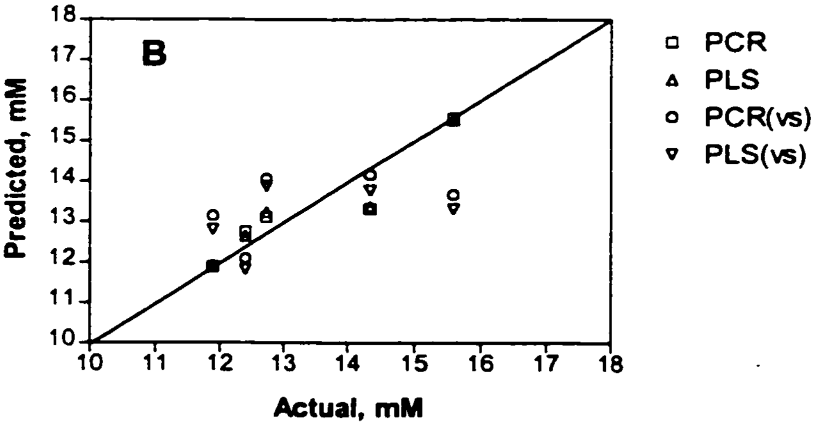
Actual, mM	PCR	PLS	PCR (vs)	PLS (vs)
15.580	15.556	15.518	13.672	13.319
14.340	13.314	13.373	14.161	13.785
12.730	13.115	13.233	14.061	13.882
12.400	12.753	12.648	12.092	11.825
11.890	11.885	11.916	13.142	12.802

Table 4.13 (B) Error Analysis

Glucose Validation	PCR	PLS	PCR (vs)	PLS (vs)
Pearson r	0.928	0.933	0.495	0.451
RMSE, mM	0.515	0.501	1.192	1.258
RRMSE, %	14.033	13.902	57.650	61.127

Figure 4.11 Results obtained using PCR and PLS to predict glucose concentrations for calibration (A) and validation (B) data sets. Results are for the entire voltammogram, and for selected potential region (vs=100 to 400 mV).





Despite repeated experiments, problems of predicting glucose with precision and accuracy continued, as poor correlations and precision continued. As the range of glucose concentrations was limited, the range was expanded to ascertain if this would improve the predictive capability. New voltammograms were obtained from thirteen glucose concentrations (10 to 75 mM) and divided into calibration (8 points) and validation (5 points) data sets. Cross-validation of the calibration set indicated that either three or seven factors would minimize prediction errors. Using both three and seven factors, the predicted glucose concentrations of the calibration and validation data sets were determined, and the results were compared. Table 4.14 (calibration) and 4.15 (validation) show the predicted concentrations, and Figure 4.12 and 4.13 the resultant correlation plots. As observed using seven factors, the calibration data set was perfectly correlated. However, as this is based on cross-validation determinations, seven factors should fit the data set perfectly. This also has the effect of returning noise back to the model, which is seen in the lower correlation coefficients of the validation data set. Using three factors, the correlation coefficients are 0.830 (PCR) and 0.792 (PLS), as opposed to poorer correlation using seven factors (0.599 for PCR and 0.596 for PLS). The perfect fit of the calibration data for seven factors, also suggests overfitting, so three factors would be the optimal number to be used in predicting the validation data. Plots of the expected and predicted results again suggest overfitting at seven factors (Figure 4.13), as the calibration data is perfectly linear, and the validation data has poor precision and accuracy. This was also evident in the RMSEC, and RRMSEC, where at seven factors, the model was unable to accurately regenerate concentration values. At three factors (Figure 4.12), the validation data is correlated, except at the highest concentration. Although the problem of poor precision, as the range of values tend to fluctuate, and the loss in accuracy (RMSEP of 29 mM, and RRMSEP of 40%) indicate that the calibration solution for 3 factors also suffers from the lack of precision and accuracy that were present in any models used for glucose prediction. The Figures also indicate that deviations from linearity occur, especially at the higher concentrations. This would affect the entire model accuracy, as possibly if the range of tested glucose concentrations was kept at a lower range, the model obtained would improve in accuracy.

Without further testing, it is difficult to determine if the final point, which shows significant deviation, is real. For example, a possible explanation for the drop in predicted concentration could be due to saturation at the electrode, resulting in a lower signal than expected.

Table 4.14 (A) **Predicted glucose concentrations found by PCR and PLS using either 3 or 7 factors of the calibration set data.**

Actual mM	PCR (3)	PLS (3)	PCR (7)	PLS (7)
75.480	43.828	49.057	75.518	75.493
48.760	42.548	43.137	48.095	48.643
40.090	42.624	39.671	41.231	40.307
30.510	35.419	35.984	29.907	30.365
22.800	21.916	19.538	22.157	22.724
17.070	16.984	18.301	18.103	17.269
14.510	13.870	10.689	14.197	14.421
10.590	43.077	43.616	10.595	10.586

Table 4.14 (B) Error Analysis

Glucose Calibration	PCR (3)	PLS (3)	PCR (7)	PLS (7)
Pearson r	0.598	0.658	0.999	1.000
RMSE, mM	0.859	0.165	46.125	43.329
RRMSE, %	1.322	0.254	153.96	112.93

Table 4.15 (A) Predicted glucose concentrations found by PCR and PLS using either 3 or 7 factors of the validation set data.

Actual mM	PCR (3)	PLS (3)	PCR (7)	PLS (7)
64.670	68.606	69.808	39.651	40.625
37.630	101.802	103.517	40.059	43.346
28.620	32.246	32.106	30.271	30.443
22.060	30.526	31.064	18.070	15.176
15.910	29.238	30.199	16.910	16.303

Table 4.15 (B) Error Analysis

Glucose Validation	PCR (3)	PLS (3)	PCR (7)	PLS (7)
Pearson r	0.830	0.792	0.599	0.597
RMSE mM	29.651	30.545	11.415	11.503
RRMSE %	40.862	41.661	49.310	40.837

Figure 4.12 Results obtained using PCR and PLS to predict glucose concentrations for calibration (A) and validation (B) data sets. Results are for the entire voltammogram, retaining three factors.

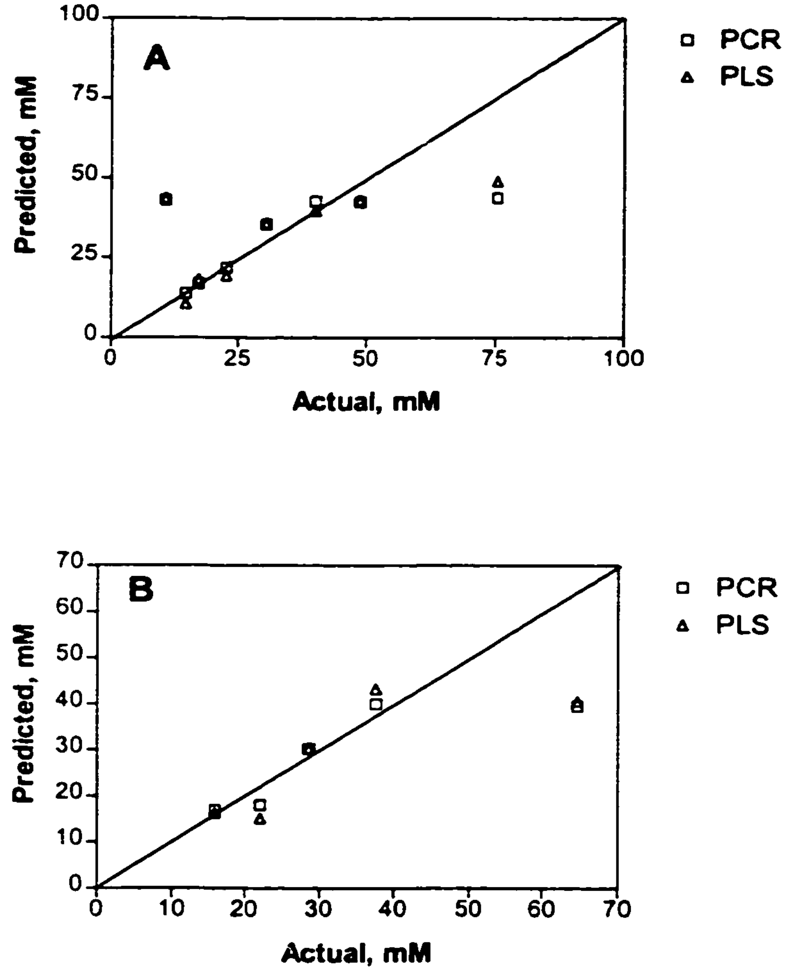
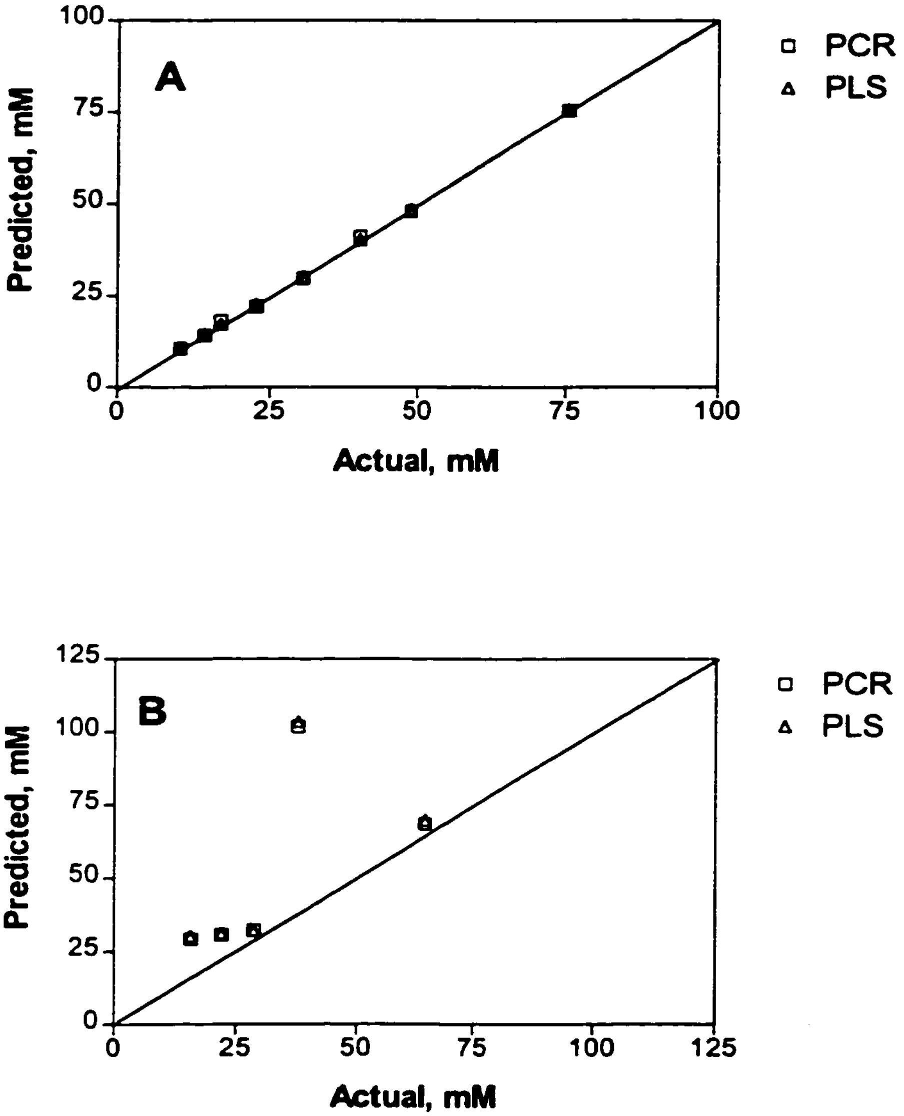


Figure 4.13 Results obtained using PCR and PLS to predict glucose concentrations for calibration (A) and validation (B) data sets. Results are for the entire voltammogram, retaining seven factors.



Overall, the results from these set of experiments have shown that for lactate, ethanol and pyruvate, multivariate calibration can predict, with reasonable accuracy and precision, individual analyte levels in a complex matrix. Results from glucose tend to show poor precision and reproducibility, so further studies need to be done to ascertain the reasons. A possible explanation could be due to non-linearities present in glucose analysis. Since PCR and PLS are both linear techniques, a non-linear method could be applied. In this case, the use of artificial neural networks could be used, as the combination of sigmoidal and linear functions within the hidden layers of an ANN are capable of fitting most types of linear and non-linear data, to return precise and accurate predictions.

Chapter 5 Individual Analyte Quantitation in Ternary Solutions

The combination of square wave voltammetry and PCR or PLS was able to predict individual analytes in PC Brew, with reasonable accuracy (see Chapter 4). A logical extension was to combine pyruvate, ethanol and glucose into samples of PC Brew, and perform simultaneous determination of the individual components.

Initially, a preliminary assessment of the feasibility of simultaneous determination of the three analytes was done. Square wave voltammetry was performed on a series of samples containing 15.0 mL PC Brew and varying amounts of glucose, ethanol and pyruvate added in 20.0 mL total volume. A solution of 0.050 M phosphate (pH=4.57) was used to dilute to volume. The calibration and validation data sets each contained nine data vectors. A calibration study was done, and predicted analyte concentrations were compared to actual values, to determine accuracy of the model. The concentrations shown are the added analyte levels, as discussed previously.

Based on the promising results obtained, a new data set was constructed to cover a wider range of possible values (0 to 0.5 to1.0, if the data were normalized). These varying concentrations of ethanol, pyruvate and glucose were added to PC Brew, as a step towards a structured data set. Varying the concentrations was also done to preclude any synergistic effects combinations of the analytes might have on the resulting signal.

Finally, the effect of the number of factors on prediction of validation results was assessed, by generating predicted concentrations of a validation data set at varying factor number, and comparing the resultant correlations.

5.1.1 Results of the Preliminary Assessment

Tables 5.1 and 5.2 shows actual and predicted sample compositions for calibration and validation data sets obtained by PCR and PLS analysis of square wave voltammograms, under previous optimized conditions, at a Pt working electrode, for ethanol. Tables 5.3 and 5.4 show the predicted results for glucose, and Tables 5.5 and 5.6 are the results for pyruvate. As a single voltammogram was generated for each ternary mixture, rank analysis to determine the optimal number of factors was done using cross-validation, and the resultant optimal number of factors was utilized to predict the analyte levels. The matrix of predictions generated using PCR or PLS was 3 columns, each column corresponding to one of the three analytes.

From the cross-validation analysis, little change was observed in the PRESS over the entire number of factors, indicating that little improvement in relevant information would be obtained by choosing factors. Keeping all the factors for the calibration solution means this is based on a multilinear calibration solution, and also indicates problems with the possible reliability and robustness of the calibration model for prediction of the validation set. As the number of factors is an indication of the important variables in the data set, keeping all factors indicates that there are more than the three components present in the mixture. An examination of the voltammogram (Figure 5.1) shows the signal generated for PC Brew (the blank), and the signal when a representative ternary mixture is scanned by square wave voltammetry. As the blank signal is relatively large in the two peak area of interest, it appears certain that more than the added analytes contribute to the signal. This could have the effect of making any calibration solution matrix-dependant, as reliable predictions could not be made when different beer samples are used.

From the calibration solution generated using all factors, predicted ethanol, glucose and pyruvate concentrations were found for the calibration and the validation data set. Prediction of ethanol percentages, showed good correlation with expected values, as correlation coefficients were greater that 0.96 for both PCR and PLS. Glucose validation results were also good with correlations greater that 0.97. Pyruvate results were poorer than those for glucose and ethanol, with correlation coefficients of 0.921 (PCR) and 0.878 (PLS). Pyruvate predictions tend to deviate significantly at the higher concentrations, however the higher pyruvate concentrations were actually outside the calibration range, so this could just be an indication of poor prediction of the calibration model for outliers.

Despite the relatively high correlations obtained, the predictive ability of the resultant calibration solution was compromised by the poor RRMSEP. Predictive errors from RMSEP tend to be high, resulting in deviations from accurate values from 24% (ethanol), 14% (glucose), and 44% (pyruvate), based on RRMSEP. The validity of the calibration solution to accurately and precisely predict these values was poor. This is illustrated in the Figures of the validation results. Part of the poor predictive ability would be due to retaining all factors in the resultant solution. This has the effect of retaining all the experimental error and seriously affects the resultant solution, as was illustrated in the high RMSEP, and RRMSEP.

Figure 5.2 shows the plots obtained from prediction of the validation data set for ethanol, glucose and pyruvate. As observed from these plots, precision is poor, despite the relatively high correlation coefficients. Significant deviation from linearity is observed, especially for the pyruvate. While the model was able to accurately predict the calibration data set, by retaining all factors, it was expected that the fit would be high as all variables are fit back into the solution. The effect on the validation set was poor accuracy, as the errors in the calibration solution were high. To improve model accuracy, a larger data set which would attempt to cover a wider range of values was required. By expanding the data set for calibration, the predictive ability of the resultant model would improve, as more variables affecting the final predictive capability would be included. Due to the complex nature of the media (as seen from the voltammogram, Figure 5.1), retaining only those factors due to the number of included variables (three in this case), may not be adequate to describe the calibration solution. However, retaining all factors will merely retain the experimental error present, and would adversely affect the final model.

From this preliminary analysis, while prediction of analyte levels of a ternary mixture appears possible, precision will need to improve significantly, and the problem of factor optimization will need to be addressed, especially in terms of matrix-dependancy.

Table 5.1 (A) Calibration set results, using PCR and PLS, for ethanol prediction.

Actual, %	Predicted, PCR	Predicted, PLS
6.410	6.338	6.354
5.540	5.167	5.308
4.970	5.129	5.080
3.860	4.229	4.116
3.302	3.288	3.455
2.741	2.738	2.681
2.180	1.535	1.618
1.061	1.704	1.425
0.500	0.403	0.509

Table 5.1 (B) Error Analysis

Ethanol Calibration	PCR	PLS
Pearson r	0.982	0.991

Table 5.2 (A) Prediction of validation set ethanol values.

Actual, %	Predicted, PCR	Predicted, PLS
6.410	6.611	6.247
6.110	6.674	6.195
4.970	5.424	4.985
4.430	5.606	5.254
3.302	4.383	4.247
2.741	1.855	1.921
1.619	5.412	2.515
1.061	1.894	1.620
0.500	0.666	0.916

Table 5.2 (B) Error Analysis

Ethanol Validation	PCR	PLS
Pearson r	0.962	0.963
RMSE, %	1.453	0.630
RRMSE, %	24.189	11.809

Table 5.3 (A) **Predication of glucose concentration of the calibration data set.**

Actual, M	Predicted- PCR, M	Predicted - PLS, M
0.0106	0.011	0.012
0.056	0.064	0.058
0.073	0.073	0.075
0.137	0.125	0.131
0.152	0.154	0.150
0.181	0.181	0.182
0.214	0.228	0.224
0.238	0.223	0.232
0.275	0.277	0.274

Table 5.3 (B) Error Analysis

Glucose Calibration	PCR	PLS
Pearson r	0.995	0.998

Table 5.4 (A) **Predication of glucose concentration of the vaiidation data set.**

Actual, M	Predicted - PCR, M	Predicted - PLS, M
0.011	-0.010	0.006
0.034	-0.004	0.017
0.082	0.062	0.079
0.108	0.052	0.066
0.149	0.103	0.108
0.182	0.181	0.178
0.228	0.167	0.163
0.241	0.202	0.210
0.262	0.253	0.245

Table 5.4 (B) Error Analysis

Glucose Validation	PCR	PLS
Pearson r	0.975	0.973
RMSE, M	0.038	0.032
RRMSE, %	14.389	13.418

Table 5.5 (A) Prediction of pyruvate concentrations for calibration set.

Actual, M	Predicted - PCR, M	Predicted - PLS, M
2.720e-04	1.654e-03	1.266e-03
1.568e-02	1.264e-02	1.476e-02
2.086e-02	2.249e-02	2.203e-02
4.668e-02	4.698e-02	4.492e-02
5.291e-02	5.108e-02	5.156e-02
3.491e-02	3.543e-02	3.544e-02
2.745e-02	2.959e-02	3.145e-02
6.408e-03	5.942e-03	4.189e-03
2.720e-04	-3.463e-04	-6.506e-05

Table 5.5 (B) Error Analysis

Pyruvate Calibration	PCR	PLS
Pearson r	0.996	0.995

Table 5.6 (A) Prediction of pyruvate concentrations of the validation set.

Actual, M	PCR, M	PLS, M
1.203e-01	4.155e-02	3.502e-02
9.132e-02	6.377e-02	5.501e-02
7.259e-02	5.133e-02	4.516e-02
6.304e-02	5.332e-02	4.806e-02
4.418e-02	5.449e-02	5.212e-02
3.559e-02	4.080e-02	4.104e-02
1.723e-02	2.484e-02	2.608e-02
9.363e-03	1.414e-02	1.197e-02
2.720e-04	-3.306e-03	-1.764e-03

Table 5.6 (B) Error Analysis

Pyruvate Validation	PCR	PLS
Pearson r	0.921	0.878
RMSE, M	0.029	0.033
RRMSE, %	43.690	57.957

Figure 5.1 Square wave voltammogram of ternary mixture containing ethanol, pyruvate and glucose in PC Brew (A), and PC Brew (B) only.

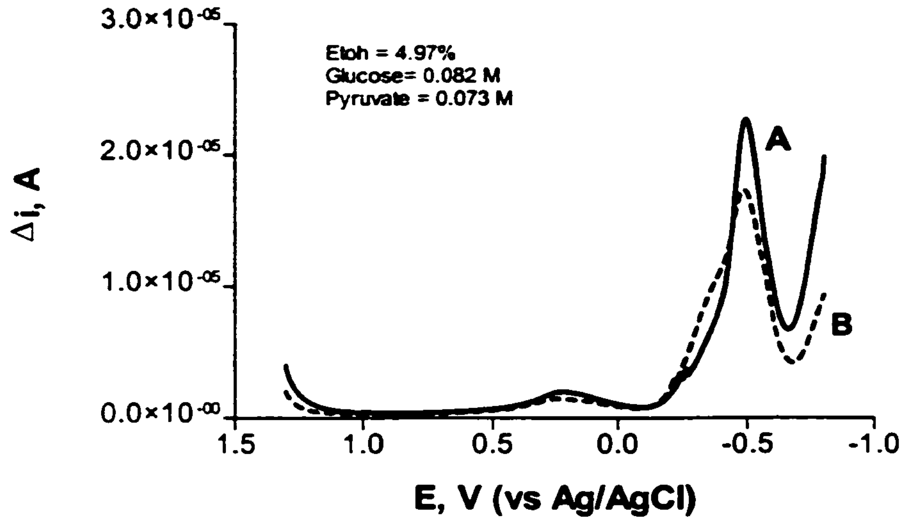
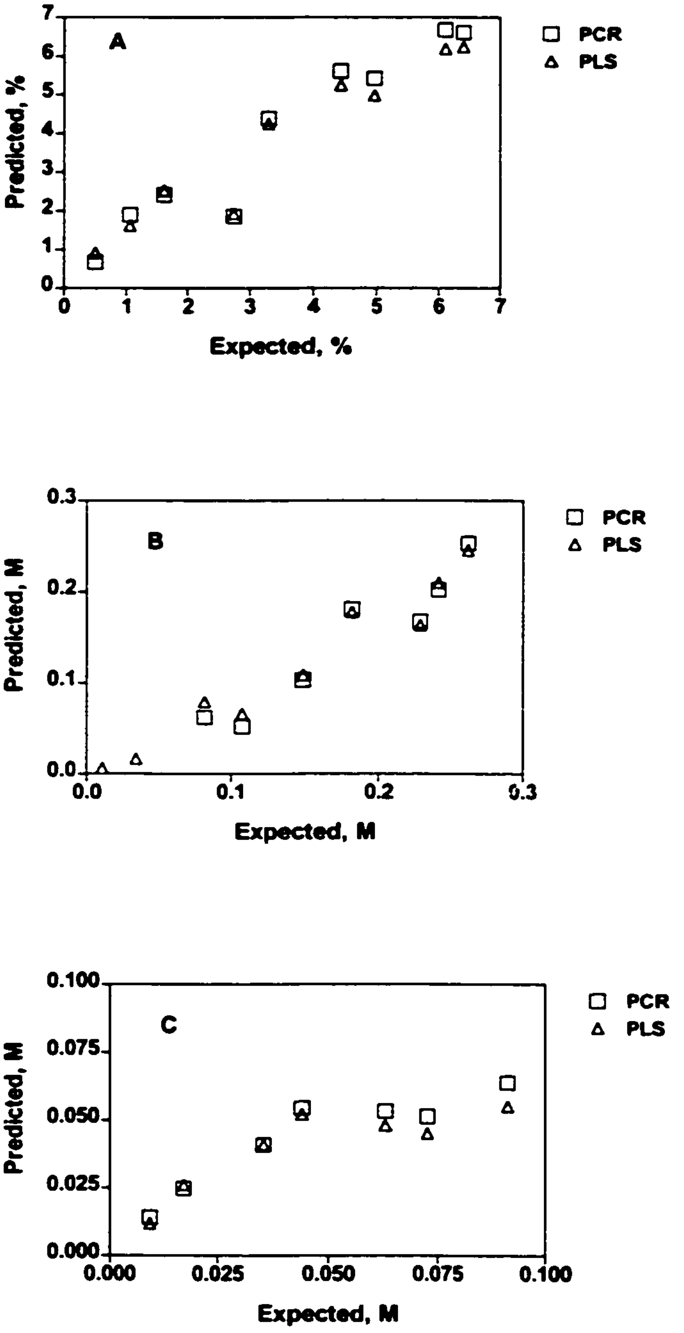


Figure 5.2 Results obtained using PCR and PLS to predict validation data sets for ethanol (A), glucose (B), and pyruvate(C). All seven factors are retained.



5.1.2 Expanded Calibration Data Set

The range of concentrations of the calibration set were expanded to cover a larger data space (42 points). Square wave voltammograms of the ternary mixtures were generated in triplicate, and averaged for further data analysis by PCR and PLS. The validation data set consisted of 10 vectors. After determination of the optimal number of factors (nine) based on the lowest PRESS found from cross-validation of the calibration data set, analyte levels were predicted for the validation data set. These results are shown in Tables 5.7 (ethanol), 5.8 (glucose), and 5.9 (pyruvate). As well, plots of the correlation between the expected and the predicted values for calibration and validation data sets are shown in Figures 5.3 (ethanol), 5.4 (glucose), and 5.5 (pyruvate).

For prediction of the validation set, ethanol prediction gave a good correlation of about 0.98 for both PCR and PLS, and a narrow confidence interval, deviating less than 10% from the expected results; pyruvate predictions gave correlations of 0.966 for PCR and 0.982 for PLS, and confidence intervals that deviate only slightly from expected at 15% (PCR) and less than 8% (PLS). Therefore, the calibration model was adequate to predict ethanol and pyruvate concentrations with good accuracy and precision. The same cannot be said for glucose, because a negative correlation and very large confidence interval is apparent. Prediction of glucose values cannot be done using this method.

In terms of reproducibility, some deviation of the replicates occurs about the expected values obtained from PCR and PLS analysis of the calibration data. Averaged results correlated closely to the expected results. The replicated results were plotted in Figures 5.3

(ethanol), 5.4 (glucose) and 5.5 (pyruvate), and the mean and standard deviation associated with the replicates is shown in Table 5.10. For ethanol and pyruvate in the calibration set (A in the Figures) little deviation occurs, indicating good reproducibility. For glucose, however, there is greater deviation in the replicates, although they do average to the expected results. As glucose prediction has been a problem in all of the previous experiments, this is not surprising. The error analysis for the calibration solution was consistent with the results obtained from correlation coefficients. For pyruvate and ethanol, RMSEP was low for both PCR and PLS. For pyruvate a RMSEP of 0.006 (PCR) compares to 0.007 (PLS), indicating no significant difference between PCR and PLS. Similarly for ethanol, RMSEP was 0.60 (PCR) and 0.63 (PLS). The RRMSEP was at 10% or less indicating relatively good ability of the model to precisely predict accurate results. Predicted results obtained from the calibration solution deviated 10% or less from expected values, making the model reliable for the prediction of ethanol and pyruvate. As mentioned, glucose predicted results had poor correlations, with the model unable to reliably and accurately predict validation results. The RMSEP for glucose was consistent between PCR and PLS (0.074 and 0.080 M, respectively), with the RRMSEP of 141%. This indicates the complete unreliability of the calibration solution to accurately predict glucose. As mentioned, this has been an on-going problem, with unreliable predictions of glucose. A more reliable model which would account for the non-linearities present, or be capable of fitting both linear and non-linear variables would be required.

Designing a larger calibration data set for model development has improved the predictive ability of the resulting calibration solution, especially for ethanol and pyruvate.

Validation results shown Part (B) of the Figures, show that the larger range of values used resulted in increasing linearity of the plot, increased correlations, as determined by the correlation coefficients, and improved error analysis as the error of the resultant model has decreased, as measured by the RMSEP. Pyruvate and ethanol can be predicted with some degree of accuracy and precision, and therefore multivariate calibration methods are appropriate for their quantitation. Glucose will require a different model to more adequately describe this system.

Table 5.7 (A) **Prediction of validation ethanol values from PCR and PLS.**

Actual, %	Predicted -PCR, %	Predicted - PLS, %
8.7500	8.344	8.482
8.000	8.264	8.391
6.500	5.403	5.326
5.000	4.781	4.628
4.250	3.771	3.489
3.500	2.442	2.522
2.750	2.746	2.776
2.000	1.863	2.026
1.250	1.671	1.858
0.500	0.078	0.532

Table 5.7 (B) Error Analysis

Ethanol Validation	PCR	PLS
Pearson r	0.984	0.978
RMSE, M	0.595	0.637
RRMSE, %	7.201	9.612

Table 5.8 (A) **Prediction of glucose concentration of validation data using PCR and** PLS

Actual, M	Predicted - PCR	Predicted - PLS
0.011	0.119	0.141
0.024	0.120	0.151
0.050	0.154	0.152
0.072	0.154	0.149
0.081	0.148	0.145
0.104	0.123	0.112
0.113	0.125	0.111
0.137	0.111	0.095
0.150	0.116	0.097
0.169	0.102	0.102

Table 5.8 (B) Error Analysis

Glucose Validation	PCR	PLS
Pearson r	-0.471	-0.879
RMSE, M	0.074	0.080
RRMSE, %	141.804	140.327

Table 5.9 (A) Prediction of pyruvate concentrations of validation data using PCR and PLS.

Actual, M	PCR	PLS
2.900e-04	5.486e-03	4.317e-03
5.926e-03	7.711e-04	5.957e-03
2.065e-02	2.117e-02	2.903e-02
3.984e-02	4.509e-02	4.905e-02
5.684e-02	6.666e-02	6.843e-02
4.147e-02	4.591e-02	5.060e-02
3.452e-02	2.530e-02	3.088e-02
2.002e-02	1.421e-02	2.084e-02
1.034e-02	1.077e-03	8.724e-03
2.900e-04	-4.226e-04	2.464e-03

Table 5.9 (B) Error Analysis

Pyruvate Validation	PCR	PLS
Pearson r	0.966	0.982
RMSE, M	0.006	0.007
RRMSE, %	9.545	10.471

Table 5.10 Reproducibility of Structured Data Set Results. Predicted results shown are the mean and standard deviation of the replicates.

Expected	PCR	PLS
Ethanol (0%)	0.023 ± 0.23	0.006 ± 0.16
Ethanol (5%)	5.075 ± 0.28	5.035 ± 0.20
Ethanol (10%)	9.915 ± 0.51	9.968 ± 0.31
Glucose (0 M)	0.001 ± 0.01	0.004 ± 0.02
Glucose (0.085 M)	0.086 ± 0.02	0.094 ± 0.03
Glucose (0.160 M)	0.162 ± 0.01	0.155 ± 0.01
Pyruvate (0 M)	0.000 ± 0.00	0.001 ± 0.00
Pyruvate (0.045 M)	0.046 ± 0.00	0.048 ± 0.0 0
Pyruvate (0.099 M)	0.096 ± 0.00	0.093 ± 0.00

Figure 5.3 Results obtained using PCR and PLS to predict ethanol percent in calibration (A) and validation (B) data sets.

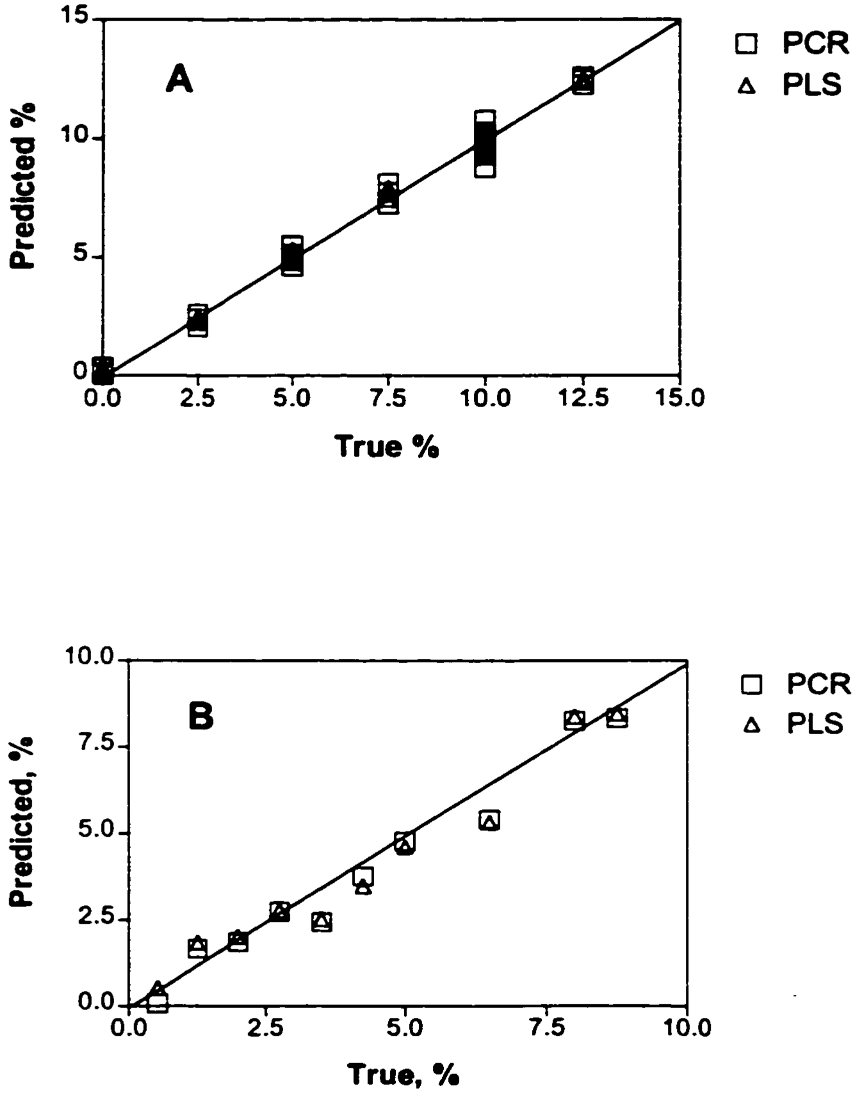


Figure 5.4 Results obtained using PCR and PLS to predict glucose concentration in calibration (A) and validation (B) data sets.

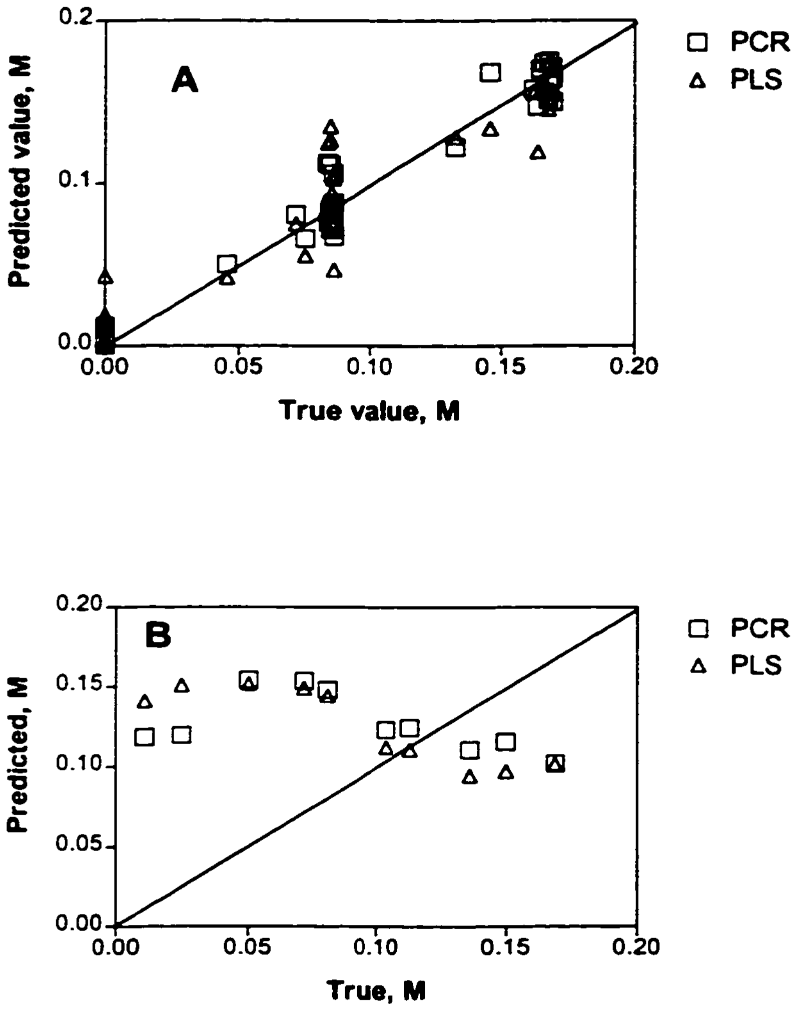
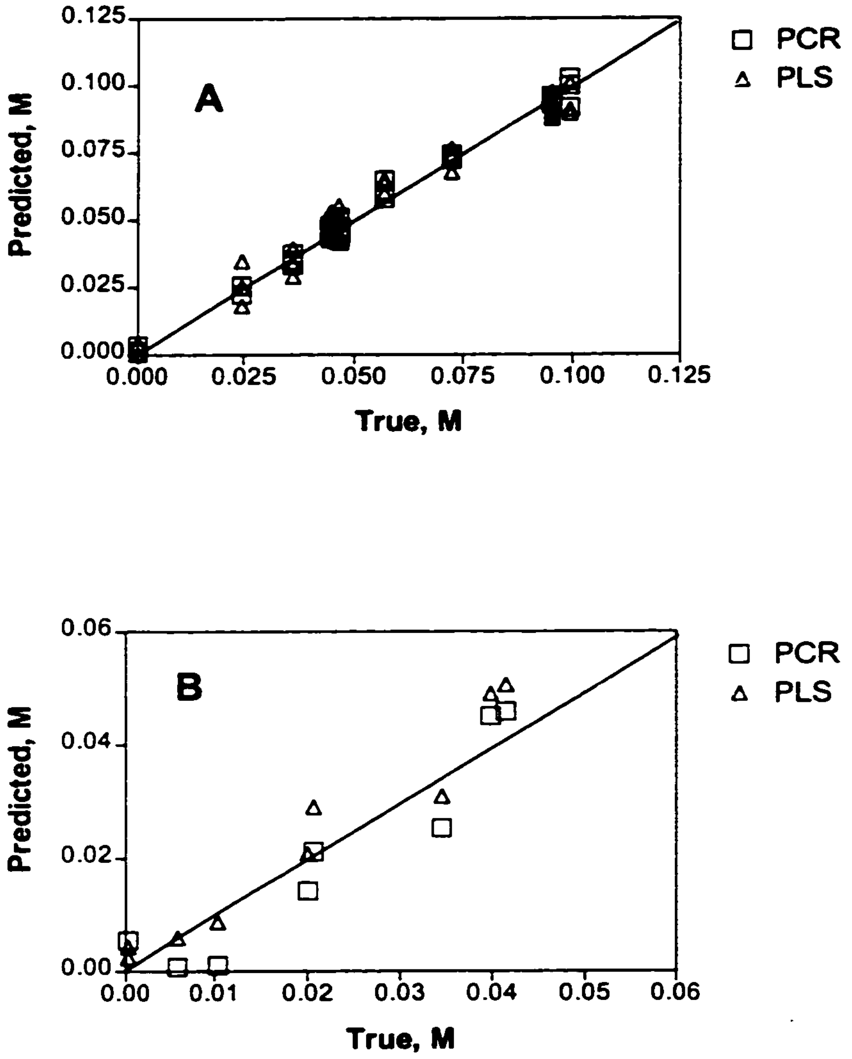


Figure 5.5 Results obtained using PCR and PLS to predict pyruvate concentration in calibration (A) and validation (B) data sets.



5.1.3 Effect of Number of Factors on Prediction

A smaller data set (10 points for the calibration, 5 for the validation) was used to determine the effect the number of factors had on the predictive ability of a calibration model. It is assumed that predictive errors are minimized when the optimal rank is chosen. To determine the optimal number of factors, PRESS at the relevant factor number was computed based on the calibration data set (Table 5.11). From the cross validation results, there is a minimum error of prediction for PCR at 4 factors, and 6 for PLS. Prediction of analyte levels in the calibration and validation sets was done using 2, 4, 6, 9 and all 10 factors. Resulting correlation coefficients are given in Table 5.12 for PCR, and Table 5.13 for PLS.

When all factors are used to predict the calibration data, the correlation coefficients are 1.000, as would be expected, since using all the factors simply regenerates the calibration matrix. Using all factors in the model to predict validation results, however, gives lower correlations, than a model based on smaller number of factors. This is expected, as noise is fit back into the calibration model, and should decrease the accuracy of prediction.

Two factors used in the calibration model, for either PCR or PLS, resulted in poor correlations in the validation set, except for glucose (0.925). However, glucose has shown poor linearity and precision, despite its good correlation, from previous studies. Therefore, glucose results, as indicated previously, are suspect. It is apparent from the lower correlations associated with two factors, that not all relevant information has been included in the model, thus deceasing its predictive capability.

From analysis of the correlation coefficients at the other number of factors, it appears that good correlation between predicted and expected values would be obtained with any of these. For PLS, six factors were recommended from cross-validations studies, and overall, six factors give the best correlations for the validation data. In terms of the PCR results, four factors, which had the minimized PRESS from cross validation, did not give the best overall correlations. However, the correlations were still good, in all cases over 0.95 for the validation data, with only slight improvement in correlation at the higher number of factors. To avoid overfitting of the data as when nine factors were used, more importance should be placed on cross validation results for the determination of the optimal rank.

This study emphasized the importance of correct factor retention, as correlations between predicted and expected values drop sharply when too few factors are retained. This affects the predictive ability of any resulting calibration solution. Retaining too many factors fits the error back into the solution, and while improving the correlation of the calibration data set, has a deleterious effect on prediction. As was presented in previous sections, too many factors retained, resulted in poor predictions of the validation data, and poor validity and reliability of the calibration solution.

Based on the results presented in this chapter, the use of either PCR and PLS for development of an accurate and reliable calibration solution was shown for ethanol and pyruvate. Glucose predicted results were generally unreliable, and a different model, perhaps based on an ANN might improve the accuracy of prediction.

Table 5.11 Square of the error (PRESS) from cross-validation performed on the calibration data set.

Factors	PCR	PLS
1	28.249	3.307
2	14.062	2.332
3	6.886	1.042
4	0.735	0.462
5	0.970	0.248
6	2.689	0.236
7	3.480	0.326
8	3.376	0.389
9	14.491	0.399

Table 5.12 Correlation coefficients for prediction of calibration and validation data sets using PCR, at given number of factors.

Pearson r for:	2	4	6	9	10
EtOH Calibration	-0.520	0.972	0.989	0.998	1.000
EtOH Validation	-0.374	0.962	0.995	0.999	0.997
Glucose Calibration	0.823	0.983	0.990	0.999	1.000
Glucose Validation	0.925	0.988	0.993	0.998	0.998
Pyruvate Calibration	0.510	0.998	0.999	1.000	1.000
Pyruvate Validation	0.672	0.966	0.973	0.980	0.979

Table 5.13 Correlation coefficients for prediction of calibration and validation data sets analysed by PLS, at the given number of factors.

Pearson r for:	2	4	6	9	10
EtOH Calibration	0.845	0.975	0.996	1.000	1.000
EtOH Validation	0.582	0.970	0.999	0.998	0.997
Glucose Calibration	0.748	0.985	0.996	1.000	1.000
Glucose Validation	0.473	0.991	0.997	0.998	0.998
Pyruvate Calibration	0.591	0.997	0.999	1.000	1.000
Pyruvate Validation	0.578	0.967	0.976	0.980	0.979

5.1.4 Summary

In general, the results obtained from these experiments, show that either PCR or PLS are adequate calibration methods for the simultaneous determination of analyte levels from a complex mixture. For ethanol and pyruvate, good correlation between predicted and expected values were obtained, provided the optimal number of factors was chosen. As observed, from the preliminary study which utilized all factors, and from the work done in correlations and factor number, choosing the optimal number of factors will decrease the prediction errors, hence improve the predictive capability of the calibration model.

Glucose, while generally giving good correlation coefficients, showed poor precision, and reproducibility. Replicate experiments with glucose have resulted in correlations which were not reproducible, and from the results presented in this Chapter, some of the problems associated with glucose prediction were shown. Plots of the validation results for glucose tend to show the lack of linearity. Hence further work needs to be done to address the problem of glucose prediction. From the previous Chapter, glucose prediction was also poor, and it was mentioned there may be a non-linear component to glucose measurement. Again, an ANN might help resolve this, as inner functions in a feedforward neural network using backpropagation, contain both sigmoidal and linear functions which are capable of simulating most linear and non-linear problems.

Chapter 6

6.1 Conclusions

This thesis has demonstrated the use of factor-based techniques such as PCR and PLS for both qualitative and quantitative analysis of a variety of complex systems. As well, an artificial neural network for the quantitation of glucose in a beer matrix has been successfully implemented. The input matrices for PCR, PLS and ANN analysis consisted of column vectors of entire voltammograms, thus allowing for whole matrix analysis, and reducing the need for prior separation of specific signals due to the analytes of interest.

Determination of the optimal voltammetric technique to be analysed by PCR and PLS required the prior assessment of normal pulse and square wave voltammetry at different conditions. After scores plots were generated of the resulting matrices, the variance between samples was greatest when a platinum electrode was used with square wave voltammetry. Further, conditions for square wave voltammetry were varied to obtain the best signal at a frequency of 5 Hz, pulse height of 50 mV, and the use of a low pass filter.

Utilizing the same conditions as described, a series of voltammograms were generated of a variety of complex liquids, such as fruit juices, wines, beers, coffees, and milks. Two peak areas observed between 100 to 400 mV, and -400 to -800 mV, were common to voltammograms of all species, due to redox-active species (such as complex sugars, ascorbate) present in all sample matrices, and redox-inactive adsorbing species.

Qualitative analysis was done by calculating the principal components from matrices predicted biomass values were compared to the OD_{600} results on a growth curve. Predicted biomass values from PCR parallelled expected values for all bacterial species.

For the next series of experiments, individual square wave voltammograms were obtained from the addition of glucose, lactate, pyruvate or ethanol in known concentrations into PC beer. Again, voltammograms for all analytes show similar responses as there are two distinct peaks observed in the same regions (0.40 to 0.10 V and -0.40 to -0.60 V). It is presumed that these peaks arise from the reduction of matrix species, specifically sugars present in the beer sample. Addition of the extra sugars or other analytes merely increases the size of these peaks. In a similar study[12], but using dual staircase pulse voltammetry of sugars and ethanol, researchers found peaks centered about -0.3 V due to ethanol, and two peaks at -0.2 and -0.7 V due to sugars. Square wave voltammetry responses do not show a peak at -0.7 V, but due to the large size of the peak in this area, it could be hidden. Correspondingly, peaks due to the ethanol and sugar could be present in the same area as mentioned by Bessant and Saini.

Based on the optimal number of factors, the calibration and validation data set concentrations were predicted using both principal components regression (PCR) and partial least squares (PLS) algorithms. These predicted concentrations were compared to the expected values, and correlations and linearity were assessed. Each individual analyte had its own factor analysis performed in order to determine the optimal number of factors to decrease prediction error. In cases where the optimal number of factors was difficult to assign, prediction values were assessed, and those factors giving the lowest predictive errors were chosen.

Using both PCR and PLS, concentrations were predicted for glucose, lactate, pyruvate and ethanol. The predictive ability of PCR and PLS for calibration data sets is very high with correlations greater than 0.98. Correlations decrease slightly for prediction of validation concentrations, but still within acceptable parameters, except for the glucose results at a correlation of 0.688. Repeated results for glucose prediction showed matrix dependent results, with correlations ranging from a low of 0.3 to a high of 0.9.

Error analysis using RMSEP and RRMSEP was done to assess the predictive ability of the calibration solution. In most cases, when the correlation coefficients were high, and resulting plots of expected and predicted values were linear and well correlated, the RMSEP was low. The resultant calibration solution was considered reliable, as predicted results were both accurate and precise. When deviation from linearity, observed on the plots, occurred, the resultant calibration solution suffered in validity, as the RMSEP of the model increased. The higher predictive errors were especially noted for glucose.

Predicting glucose reliably and repeatedly continued to be a difficulty throughout the set of experiments. The possibility of using a neural network to predict glucose levels was made in order to compensate for any non-linear relationships which exist between the responses generated by a voltammogram and the ensuing concentration.

A mixture of glucose, pyruvate and ethanol in varying concentrations was added to PC brew, and the subsequent voltammograms were used as input vectors for factor analysis. From the voltammogram of the ternary mixture, individual species cannot be resolved as the peaks overlap in the same region. Therefore individual determination of separate species can best be done using factor analysis methods such as PCR and PLS.

6.2 Future Work

It is proposed that in several areas presented in this work, further studies are required. While the generation of scores plots is an acceptable method for sub-population classification, it is in the areas of quantitative analysis that the multivariate solutions discussed show the strongest possibility.

While prediction results of either individual analytes or the ternary mixture components were generally accurate, a more robust calibration solution would improve the reliability and accuracy of the model, and decrease the RMSEP. A structured data set, comprising all possible data space is required to improve the model. This would improve the predictive capability of either PCR or PLS models. Further, while little improvement was observed between PCR and PLS, a model incorporating non-linear functions might improve the correlation, and decrease the RMSEP for prediction of glucose results. The most common model, a feedforward artificial neural network with backpropagation algorithm, incorporates both linear and non-linear functions.

Increased experimentation for the biomass study, incorporating more bacterial species, and a larger data set might result in a sensor capable of detecting either bacterial types or bacterial biomass in a complex broth, without the need for pre-separation. As well, further work in using the combination of voltammetry and multivariate analysis in fermentation media would increase the acceptance of these techniques for accurate and reliable component analysis, without pre-separation or pre-treatment. This would be of use in on-line process control, or as part of a larger multi-sensor array.

6.3 References

1. Wold, S. *Chemom. Intell. Lab. Syst.*, **1995**, *30*, 109-115.

2. Beebe, K.R.; Pell, R.J.; Seasholtz, M.B. *Chemometrics: A Practical Guide* Wiley-Interscience Publication: New York, **1998.**

3. Thomas, E.V. *Anal.* Chem. **1994,** *66*(15), 795A-804A.

4. Pearson, K. *Philos. Mag.* **1901,** Series 6, *2*, 559.

5. Hotelling, H. *J. Educ. Psych.* **1933,** *24*, 417.

6. Malinowski, E.R. *Factor Analysis in Chemistry,* 2nd ed. Wiley-Interscience Publication: New York, **1991.**

7. Kramer, R. *Chemometric Techniques for Quantitative Analysis* Marcel Dekker Inc.: New York, **1998.**

8. Brereton, R.G. *Chemometrics Applications of Mathematics and Statistics to Laboratory* Ellis Horwood: New York, **1990.**

9. Kissinger, P.T.; Ridgway, T.H. In *Laboratory Techniques in Electroanalytical Chemistry,* 2nd ed., P.T. Kissinger, W.R. Heineman, eds. Marcel Dekker, New York: **1996**, 141-163.

10. Bard, A.J.; Faulkner, L.R. *Electrochemical Methods: Fundamentals and Applications* Wiley Interscience, New York: **1980**.

11. Johnson, D.C.; LaCourse, W.R. *Anal. Chem.*, **1990,** *62*(10), 589A-597A.

12. Bessant, C.; Saini, S. *Anal. Chem.* **1999,** *71*, 2806-2813.

13. Perone, S.P.; Kretlow, W.J. *Anal. Chem.* **1966,** *38*(12), 1760-1763.

14. Akkermans, R.P.; Wu, M.; Bain, C.D.; Fidel-Suarez, M.; Compton, R.G. *Electroanalysis*, **1998,** *10*(9), 613-620.

15. Dodd, G.H.; Squirrell, D.J. Symp. Zool. Soc., Lond. **1980,** *45*, 35-56.

16. Gardner, J.W. 8th Int. Congress of European Chemoreception Research Organization, University of Warwick, U.K., July 1987.

72. Glassey, J.; Montague, G.A.; Ward, A.C.; Kara, B.V. *Biotechnol. Bioeng.* **1994,** *44*, 397-405.

73.Seasholtz, M.B. *Chemom. Intell. Lab. Syst.* **1999**, *45*, 55-63.

74.Schonkopf, S. *American Laboratory* **1999**, *31(9)*, 32-34.

75. Eisen, M.B.; Spellman, P.T.; Brown, P.O.; Botstein, D. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 14863-14868.

76. Alon, U.; Barkai, N.; Notterman, D.A.; Gish, K.; Ybarra, S.; Mack, D.; Levine, A.J. (1999), *Proc. Natl. Acad. Sci. USA* **96**, 6745-6750.

77. Alizadeh, A.A.; Eisen, M.B.; Davis, R.E.; Ma, C.; Lossos, I.S.; Rosenwald, A.; Boldrick, J.C.; Sabet, H.; Tran, T.; Yu, X.; Powell, J.I.; Yang, L.; Marti, G.E.; Moore, T.; Hudson, J., Jr.; Lu, L.; Lewis, D.B.; Tibshirani, R.; Sherlock, G.; Chan, W.C.; Greiner, T.C.; Weisenburger, D.D.; Armitage, J.O.; Warnke, R.; Levy, R.; Wilson, W.; Grever, M.R.; Byrd, J.C.; Botstein, D.; Brown, P.O.; Staudt, L.M. (2000) *Nature* **403,** 503-511.

78. Golub, T.R.; Slonim, D.K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J.P.; Coller, H.; Loh, M.L.; Downing, J.R.; Caligiuri, M.A.; Bloomfield, C.D.; Lander, E.S. (1999) *Science* **286**, 531-537.

79. Wang, X.; Zhuang, Z.; Zhu, E.; Yang, C.; Wan, T.; Yu, L. (1995) *Microchem. J.* **51**, 5-14.

80. Zar, J.H. *Biostatistical Analysis*, 2nd ed. Prentice-Hall, New York: **1984**, 306-309.

81. Stewart, G.G.; Russell, I., *MBAA* , **1993**, *30*(4), 159-168.

82. Favier, J.P.,; Bicanic, D.; Helander, P.; van Iersel, M. *J. Biochem. Biophys. Methods*, 1997, *34*, 205-211.

83. Young, T.W. In *Brewing Microbiology*, 2nd ed.; Priest, F.G., Campbell, I. eds., Chapman & Hall: London, **1996,** 13-42.

84. Kramer, R. *Personal Communication*, 12 March, **1999.**