NIR Spectroscopic Classification of Urine, Serum, Plasma and Plasma Anticoagulants using Mahalanobis Distance and Genetic Algorithm Selection of Wavelengths

by

Douglas George Given

A thesis

presented to the University of Waterloo

in fulfilment of the

thesis requirement for the degree of

Master of Applied Science

in

Systems Design Engineering

Waterloo, Ontario, Canada, 1997

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

0-612-44748-0

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Abstract

The need for rapid, non-invasive techniques to screen specimens is becoming more critical as chemical laboratories enter the automation era. Near-infrared spectroscopic instrumentation is capable of meeting the requirements for specimen screening in the automated environment. However, some measurements from near-infrared spectroscopic instruments yield very low signal-to-noise-ratios. Therefore, the data analysis method used to calibrate such instrumentation must optimise the performance and should also provide a parsimonious solution to maintain a rapid measurement.

In this thesis urine, serum, plasma, and plasma anticoagulant spectroscopic data were collected, processed and studied to evaluate the performance of various classification methods, namely, K-Nearest Neighbour and Mahalanobis Distance methods. Wavelengths were transformed into principal component scores, to reduce the number of features. The Mahalanobis Distance method was also optimised using a Genetic algorithm to select the best wavelengths, thus reducing the number of wavelengths required.

The conclusion is that the Mahalanobis Distance method is superior to the K-Nearest Neighbour method in terms of predictability. The Genetic algorithm was able to increase predictability even further, while reducing the number of wavelengths required in the Mahalanobis Distance model.

Acknowledgements

I would like to thank my supervisors, Dr. Andrew K. C. Wong and Dr. Theodore E. Cadell, for their support in the development of this thesis.

I also wish to thank Dr. James Samsoondar for his assistance in the chemical aspects and Dr. Stephen Hughes for providing information and advice on genetic algorithms.

I also wish to thank Aidan Furlong for providing CME Telemetrix's co-operation.

Finally, I wish to thank my wife, Eleanor, and my daughters, Sarah and Elizabeth, for their patience and support throughout this process.

Dedication

To my Father, who models Perseverance

and

To my Mother, who models Sacrifice;

for without

Perseverance and Sacrifice

nothing

Great will be Accomplished.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Clinical laboratories in North America have come under increasing governmental and economic pressures to provide higher quality services at reduced costs. These pressures have necessitated re-engineering and automating the laboratory operations, a process MDS Laboratory Services [1] launched in 1992 and from which it is now starting to reap benefits. Responsibilities that were once the lab technician's now are being performed by a complex network of conveyance devices, robotic arms, bar-code readers, expert system decision makers, and automated analysers. This rapid, automated system can no longer rely on visual inspection by the technician to monitor specimen integrity, that is, specimen quality and type. New instruments are arising and are about to arise which fit into the automated laboratory and fill the gaps in checking specimen integrity. One such instrument is the CME-Automated Specimen Screening (CASS) system [2]. This instrument is capable of measuring the concentration of three potentially interfering conditions or substances: hemolysis, turbidity, and biliverdin. It is important to many analysers to pre-screen these interfering conditions or substances. Incorrect measurements of analytes can occur if the level of one or more of the interfering substances exceeds specified thresholds, appropriate for the specific analyser. The

CASS system is capable of measuring the concentrations of all three interfering substances simultaneously within five seconds. The CASS instrument is based on a Vis-NIR (Visible-Near Infrared) spectroscopic measurement of the specimen.

Another aspect in checking specimen integrity is to verify the specimen type or species. Two major specimen types or species are urine and blood. Serum and plasma are both derived from blood. A plasma specimen contains one of several anticoagulants: citrate, oxalate, iodoacetate, EDTA, or heparin. Presently, no instrument can measure all of these as a pre-screen function, within an automated environment.

The spectroscopic data studied for this thesis were collected from various specimen types using a CASS system. The data were analysed to determine if sufficient signal was present to distinguish the various specimen types or species from each other. Several methods were employed to analyse the data and to optimise the classification rate of unknown specimens. K-Nearest Neighbour (KNN) and Mahalanobis Distance (MD) were used to classify the specimens. A wavelength selection method employing a Genetic Algorithm (GA) was also implemented to form a new method: GA optimised MD (GA-MD) method.

The thesis is divided into seven chapters. Chapter 1 provides background to the problem, the science, the data collection method, and the analysis methodologies. Chapter 2 describes the data transformation methods and outlier detection. Chapter 3 describes the K-Nearest Neighbour classification method. Chapter 4 describes the Mahalanobis Distance classification method. Chapter 5 describes the Genetic Algorithm used to select wavelengths in optimising the Mahalanobis Distance results. Chapter 6 outlines the results of the analyses. Chapter 7 provides conclusions to the study.

1.2 Specimens

Spectroscopic data were collected on a CASS spectrophotometer. Approximately 350 separate specimens were measured once each, one specimen at a time, by the CASS system. Each specimen was either a urine specimen or a blood[1] specimen. The blood

[1] Blood, here, means specimens derived from the whole blood specimens. Generally, whole blood is rarely required for clinical chemistry tests.

specimens were further identified as either serum or plasma. Each plasma specimen contained an anti-coagulant. The possible anti-coagulants tested were citrate, EDTA, oxalate, iodoacetate, and heparin.

A specimen was categorised according to the hierarchy shown in Figure 1.1.

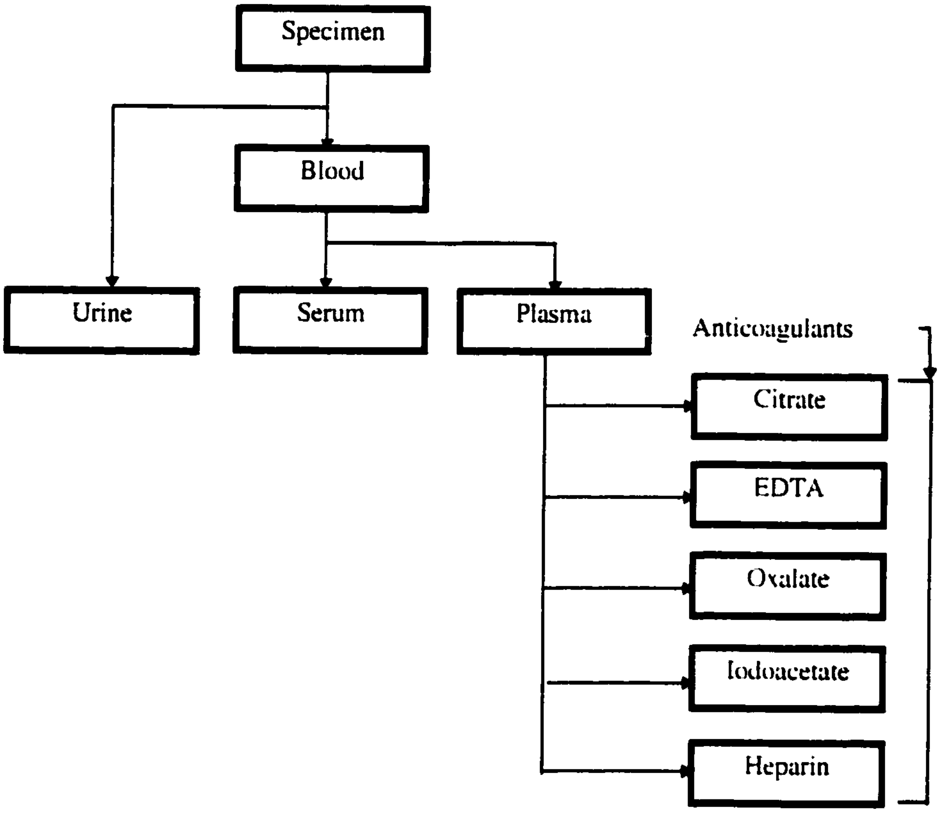


Figure 1.1 Specimen Categorisation

For data analysis purposes the specimens were divided into seven groups or species and identified by a single numeral from 1 to 7 as shown in Table 1.1.

Table 1.1: Specie Identification

Specie	Specie Number
Urine	1
Serum	2
Citrate	3
EDTA	4
Oxalate	5
Iodoacetate	6
Heparin	7

Species 2 to 7 are part of the higher category or specie: blood. Species 3 to 7 are part of the higher category or specie: plasma.

Each spectroscopic measurement is an absorbance measurement of the specimen for 255[2] separate wavelengths ranging from 602 to 1042 nanometers (nm). Therefore, each measurement consists of a vector having 255 elements (or pixels). Each measurement is an average of 32 scans of the instrument spectrometer. There are approximately 50 specimens of each specie.

1.3 Objectives

The ultimate objective is to accurately classify an unknown specimen as one of the seven specie types. An error rate of less than or equal to 5% is the requirement for accurate classification of a given specie. In other words, a given specie must be classified accurately at least 95% of the time. The ultimate objective can be broken down into 3 smaller objectives or steps:

1. The first objective is to accurately classify an unknown specimen according to one of the two major categories: urine or blood.
2. Given that the first objective is satisfied, the second objective is to accurately classify an unknown blood specimen into one of the two sub-groups: serum or plasma.
3. The third objective is to correctly classify an unknown plasma specimen into one of the five anticoagulant categories.

1.4 Spectroscopy

To better understand the spectral data that are collected, a basic understanding of spectroscopy is necessary. This section will provide this understanding and provide a cursory glimpse into what sort of signal can be expected from each specie.

[2] The CASS system actually provides 256 wavelengths but the longest wavelength was "lost" in the data processing due to matrix size limitations of one of the off-the-shelf software programs.

Spectroscopy is the science of the interaction between electromagnetic radiation and matter. Specifically, we are dealing with absorption spectroscopy in the Visible and Near Infrared (Vis-NIR) regions of the electromagnetic spectrum. The wavelength ranges for the Visible and NIR regions are 380 to 780 nanometers and 780 to 2526 nanometers respectively, in a vacuum. However, the discussions in this thesis will focus on the wavelengths from 602 to 1042 nanometers; the measurement range of the CASS system used for data collection.

NIR spectroscopy can be traced back to 1800 but practical applications were not developed until the 1950s, when NIR spectroscopy was used to determine the quality of grain in the agriculture industry. Since then, a number of applications have been developed including those in the pharmaceutical, petroleum, biomedical, and textile industries. Edward Stark, Karen Luchter, and Marvin Margoshes [3] provide more detail on these applications. Figure 1.2 shows the basic configuration for a transmission spectrophotometer.

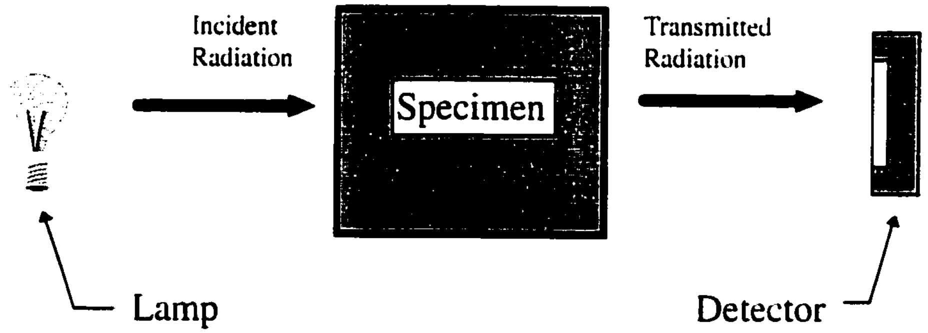


Figure 1.2: Basic Configuration for a Transmission Spectrophotometer

1.4.1 Interaction of Electromagnetic Radiation and Matter

Electromagnetic radiation, or "light"[3] as we shall call it, interacts with matter in several ways: emission, absorption and scatter. Emission is not applicable to this application. These interactions are due to the nature of light, that is, energy carried by electromagnetic waves [4]. The wavelength of a travelling wave in a vacuum is given by

$$\lambda = \frac{c}{v} \times 10^{9} \tag{1.1}$$

where λ is the wavelength in nanometers (nm), c is the speed of light (2.998×10^8 m/s) and v is the frequency of the light in Hertz. The energy in this travelling wave or photon is given by

$$E = hv \tag{1.2}$$

where E is the energy of a single photon (Joules) and h is Planck's constant (6.63×10^{-34} J-sec). Combining Equations 1.1 and 1.2, we obtain:

$$E = hc / \lambda. \tag{1.3}$$

The electrons in atoms, molecules and ions possess discrete (quantized) energies. Therefore, a change in the energy of a molecule must result from a quantum jump. An interaction[4] between a photon and a molecule is given by

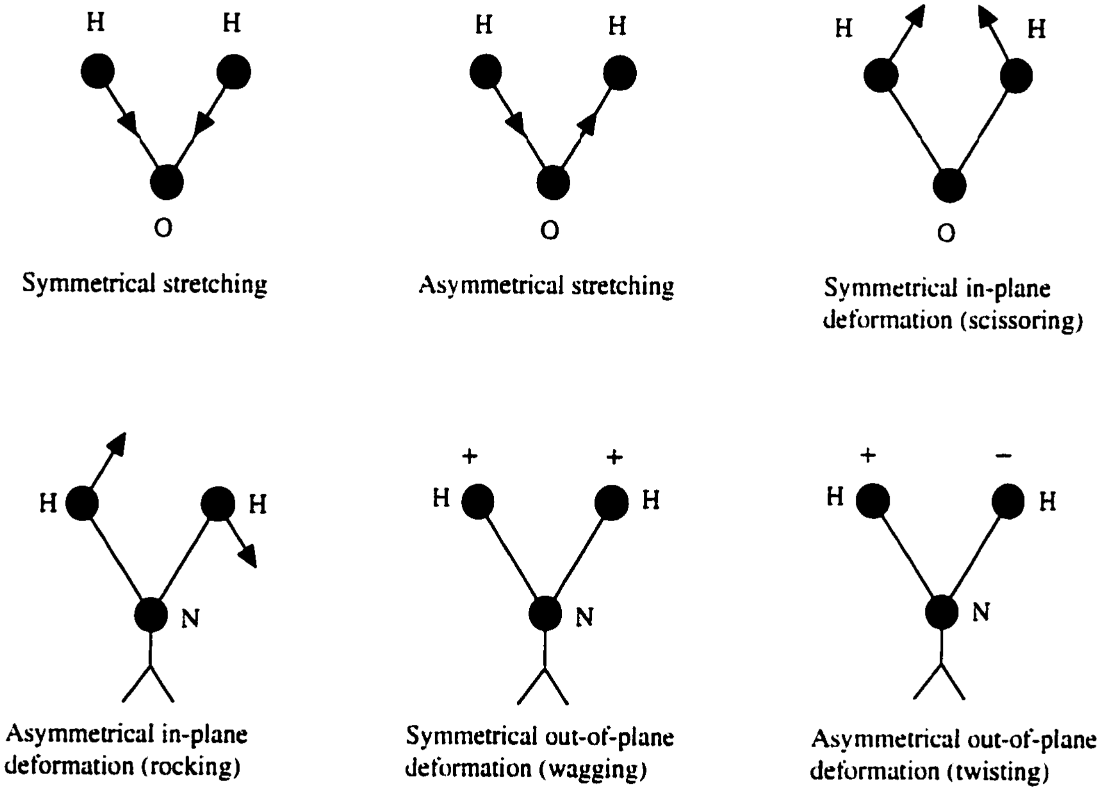
$$E_{photon} = \Delta E = hc / \lambda \tag{1.4}$$

where ΔE is the energy difference between state 1 (initial state) and state 2 (final state). The energy increase of a molecule due to absorption of photons in the infrared region, according to C.N Banwell [5, p. 7], causes the molecule to undergo a "change of configuration". The change in configuration is due to the vibrations of the bond(s) of a

[3] Light normally refers to the part of the electromagnetic spectrum that is visible to the human eye, i.e., wavelength range of approximately 400 to 750 nanometers. Since a portion of the spectrum used is in the visible range, we shall refer to electromagnetic radiation as "light" for the sake of communication and interpretation.

[4] In our case, this will be absorption (or absorption response due to scatter). The photon energy will be absorbed by a molecule, increasing its energy level.

molecule. Vibrations can be classified as either bending (change in bond angle; scissoring) or stretching (change in interatomic distance along the bond axis). The modes of vibration for a triatomic molecule such as water (H_2O) or NH_2 are shown in Figure 1.3. For functional groups such as NH_2 bonded to an organic framework, bending is further classified into three other types: rocking, wagging, or twisting.



Symmetrical stretching

Asymmetrical stretching

Symmetrical in-plane deformation (scissoring)

Asymmetrical in-plane deformation (rocking)

Symmetrical out-of-plane deformation (wagging)

Asymmetrical out-of-plane deformation (twisting)

Figure 1.3: Modes of Vibration for Triatomic Molecule

As an example, the equation for the frequency ν, of a stretch vibration for a diatomic molecule is given by

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{\frac{(m_1 m_2)}{(m_1 + m_2)}}} \qquad (1.5)$$

where π is the constant 3.14159..., k is the force constant of the bond, m_1 is the mass of atom 1 and m_2 is the mass of atom 2. Thus the force constant (related to the strength of the chemical bond) can be determined from measurements of bond frequencies. For polyatomic molecules the procedure is extended through the use of normal co-ordinate calculations. The carbon-hydrogen (C-H) bond stretch is 2900 cm^{-1} (or a wavelength of 3440 nanometers *in vacu*) which occurs in the mid-IR[5] region. This is referred to as the fundamental frequency. Overtone (or harmonic) frequencies also occur. Table 1.2 shows observed NIR absorption bands.

The absorbance response to the concentration of a pure[6] absorber follows the Beer-Lambert law. This relationship is fundamental to spectroscopy and is given by

$$A = \log\frac{I_o}{I_t} = \log\frac{1}{T} = kcl \tag{1.6}$$

where A is the absorbance response, I_o is the intensity of the incident light, I_t is the intensity of the transmitted radiation, T is the transmittance response, k is the extinction coefficient (proportionality constant of molecular absorption), c is the concentration of the absorber and l is the effective pathlength of the specimen.

Table 1.3 shows typical, relative absorbances of different overtone responses. Although there is a larger response (i.e., due to fundamental) in the mid-IR region, NIR has the advantage of being able to measure higher concentration specimens transmissively (i.e., straight through the specimen), using longer pathlengths. Typically, specimen preparation, such as dilution, is not required for NIR, allowing rapid, non-destructive measurement of the specimen.

NIR spectroscopy responds chiefly to the chemical bonds: C-H, O-H, and N-H and their combinations as pointed out by Kirsch and Drennen [6, p. 141] and others.

The absorbance interaction assumes that the radiation path follows a straight-line through the specimen. However, interactions can occur between the matter and the radiation which will change the radiation direction. These interactions are due to reflections off particles in the specimen, or refraction due to changes in index of refraction within the specimen, if it is not homogeneous, or at the optical interfaces to and from the specimen. These reflection and refraction interactions lead to "scatter".

[5] mid-IR covers the range of 2,500 to 50,000 nanometers.

[6] Pure, meaning a substance which does not scatter the light. Scatter will be explained later in this section.

Table 1.2: Chemical Assignments of Some Observed NIR Absorption Bands

Wavelength (nm)	Bond Vibration	Structure
713	C-H str. fourth overtone	benzene
738	O-H str. third overtone	ROH
740	C-H str. fourth overtone	CH_3
746	C-H str. fourth overtone	CH_2
747	O-H str. third overtone	ArOH
760	O-H str. third overtone	H_2O
762	C-H str. fourth overtone	CH_2
779	N-H str. third overtone	RNH_2
790	N-H str. third overtone	$ArNH_2$
806	N-H str. third overtone	RNH_2
808	2 x N-H str. + 2 x N-H def. + 2 x C-N str.	RNHR'
815	N-H str. third overtone	RNHR'
832	2 x N-H str. + 2 x N-H def. + 2 x C-N str.	RNHR'
840	3 x C-H str. + 2 x C-C str.	benzene
874	C-H str. third overtone	benzene
880	C-H str. third overtone	$CHCl_3$
900	C-H str. third overtone	CH_3
910	C-H str. third overtone	protein
913	C-H str. third overtone	CH_2
928	C-H str. third overtone	oil
938	C-H str. third overtone	CH_2
970	O-H str. second overtone	ROH, H_2O
990	O-H str. second overtone	starch
1000	O-H str. second overtone	ArOH
1015	2 x C-H str. + 3 x C-H def.	CH_3
1020	2 x N-H str. + 2 x amide I	protein
1020	N-H str. second overtone	$ArNH_2$
1030	N-H str. second overtone	RNH_2
1037	2 x C-H str. + 2 x C-H def. + $(CH_2)n$	oil
1053	2 x C-H str. + 2 x C-H def. + $(CH_2)n$	CH_2
1060	N-H str. second overtone	RNH_2
1080	2 x C-H str. + 2 x C-C str.	benzene
1097	2 x C-H str. + 2 x C-C str.	cyclopropane

Source: Osborne, B.G., Fearn, T. and Hindle, P.H.; *Practical NIR Spectroscopy with Applications in Food and Beverage Analysis (Second Edition).* Longman Group, Essex, England, 1993. pp. 29-30 Table 2.4.

Table 1.3: Typical Changes in Intensity of Absorbance at Different Overtone Levels

Transition (v_0 to v_n)	Overtone No.	Relative Absorbance for 1-cm Cell
1	Fundamental	100
2	First	9
3	Second	0.3
4	Third	0.01

Source: Williams, P.C. and Norris, K. Editors; *Near-Infrared Technology in the Agricultural and Food Industries*. American Association of Cereal Chemists, Inc., St. Paul, Minnesota, 1987. Chapter 2 p. 19 Table I.

Typically, the reflection component has a larger effect on the measured or "perceived absorbance". One consequence of scattering is that the effective or "perceived pathlength" appears to be longer than for the case where the scatterer was not present. This leads to an increase in the actual and perceived absorbances. What is meant by this, is that the actual absorbance increases due to the mean path being longer for the radiation, and the perceived absorbance increases due to light being scattered outside of the optical collection area, i.e., where the radiation is detected exiting the specimen. This "lost" radiation appears as increased absorbance. The changes to the measured absorbance due to the scatterer has three basic components: a baseline shift due to pure scattering effects, a proportional change due to increased effective pathlength and thus increased absorbance, and a wavelength dependent component due to reflection characteristics of the particles and refraction characteristics of the media. A number of models have been developed for characterising scattering effects. These are discussed briefly in Chapter 2.

1.5 Expected NIR Response of the Specimens

NIR responses of similar substances have been studied by a number of researchers, with some reporting success in quantifying analytes in urine [9] and blood [10] [11].

1.5.1 Urine Versus Blood

Urine is a waste product of the human body excreted by the kidneys. It consists of about 96 per cent water, 2 per cent urea, creatinine, 0.5 per cent uric acid and 1.5 per cent inorganic salts such as sodium, potassium, ammonia, calcium and magnesium [12, p. 701]. The kidney acts as a filter[7], filtering most of the smaller molecules to excretion. Negatively charged molecules are filtered less easily than positively charged molecules [13, pp. 321-322]. The filterability of large molecules such as albumin and proteins is very low and so only a very small percentage of large molecules is filtered through to excretion.

Whole blood is made up of red blood cells, white blood cells, platelets and a yellow liquid called plasma. Plasma consists of approximately 90 per cent water and 10 per cent dissolved matter. The dissolved matter consists of approximately 70 per cent plasma proteins, 20 per cent organic metabolites, urea, and uric acid, and 10 per cent inorganic salts [14, pp. 706-707].

One obvious difference between blood and urine is that blood contains a relatively large concentration of proteins (contained in plasma) compared to urine. Many types of proteins are found in blood but the major ones are: serum albumin, α_1-globulins, α_2-globulins, β-globulins, γ-globulins, fibrinogen, and prothrombin. Protein is synthesised from amino acids. Amino acids contain a basic amino group ($-NH_2$) and an acidic carboxyl group ($-CO_2H$). A larger NIR response to the NH_2 structure is expected for blood than for urine. The concentration of serum albumin, for example, is approximately 3500 mg/dL, a relatively high concentration for producing an NIR response. The NH_2 structure has responses from 779 to 1060 nanometers. Therefore, a distinguishable NIR response should be obtained between urine and blood based on the NH_2 response alone.

[7] However, urine is not an ultra-filtrate of blood because after the filtering process, the kidney has the ability to selectively absorb small molecules as required for maintaining homeostasis.

1.5.2 Serum Versus Plasma

Plasma can be separated from the cells by centrifugation. An anticoagulant (e.g., heparin) is added to the blood to prevent coagulation. Serum is formed when the whole blood is allowed to coagulate. If an anticoagulant is not added, the cells will form a clot which can be separated from the serum by centrifugation. Serum, thus, is the blood plasma minus some clotting proteins used to form the clot; the plasma will also contain the added anticoagulant while serum will not.

Serum and plasma, for the most part, are very similar in composition. The amount of protein used in the clot formation is relatively small compared to the overall amount of protein and the variance in specimens. Therefore, we expect the anticoagulants' NIR response to provide the strongest spectral information to distinguish serum from plasma.

1.5.3 Plasma Anticoagulants

The plasma specimens contain one of five anticoagulants: citrate, oxalate, iodoacetate, ethylenediaminetetraacetic acid (EDTA), and heparin.

Citrate [14, p.442] contains an O-H bond and two CH_2 bonds. Therefore an NIR response is expected.

Sodium oxalate's [15, p.8603] chemical formula is $Na_2C_2O_4$. This contains C-C and CO_2 groups which may have overtone vibrations, although Table 1.2 does list these.

The chemical formula for iodoacetate [14, p.409] is ICH_2COO^-. The CH_2 bond has an NIR response.

EDTA's [15, p.3752] chemical formula is $[(O_2C\ CH_2)\ {}_2N\text{-}CH_2\text{-}CH_2\text{-}N(CH_2CO_2)_2]^{4-}$. Therefore, at least CH_2 and NR_2 groups are contained in EDTA. An NIR response is expected.

Heparin [15, p.4575] is a more complicated structure compared to other anticoagulants. It is a mixture of variably sulphated polysaccharide chains composed of repeating units of D-glucosamine [15, p.4353] and either L-iduronic or D-glucuronic acids.

Glucosamine contains C-H, O-H and N-H bonds. Therefore an NIR response is expected.

Therefore, it is possible for an NIR response to occur for all the anticoagulants except, possibly, for oxalate.

1.6 Measurement Methods and Materials

1.6.1 Instrumentation

The spectrophotometer and associated instrumentation used in this investigation were designed and built by CME Telemetrix Inc., Waterloo, Ontario, Canada. The system is referred to as CASS (CME-Automated Specimen Screening). The specimen is placed inside a cylindrical translucent tube or vial. An unmarked, white, adhesive label covers over half of the circumference of the tube. The specimen interface is remote from the spectrophotometer; radiation is channelled via 3.2 millimetre diameter fibre-optic bundles to and from the specimen. Broadband Visible and NIR radiations are transmitted simultaneously through the label, tube wall, specimen, and again through the tube in that sequence. A representative optical schematic of the system is shown in Figure 1.4 [16, p.6]. The instrument is a double-beam-in-time spectrophotometer, having two paths which are measured in time sequence (i.e., time-division multiplexed). The light source is a quartz-tungsten-halogen lamp. A holographic grating is used to disperse the broadband radiation into its component wavelengths. A linear silicon photodiode array of 256 elements or pixels is used to collect the transmitted radiation of each component wavelength.

The absorbance is calculated based on the ratio of the transmitted radiation of the specimen to transmitted radiation of the reference. Dark measurements and differences

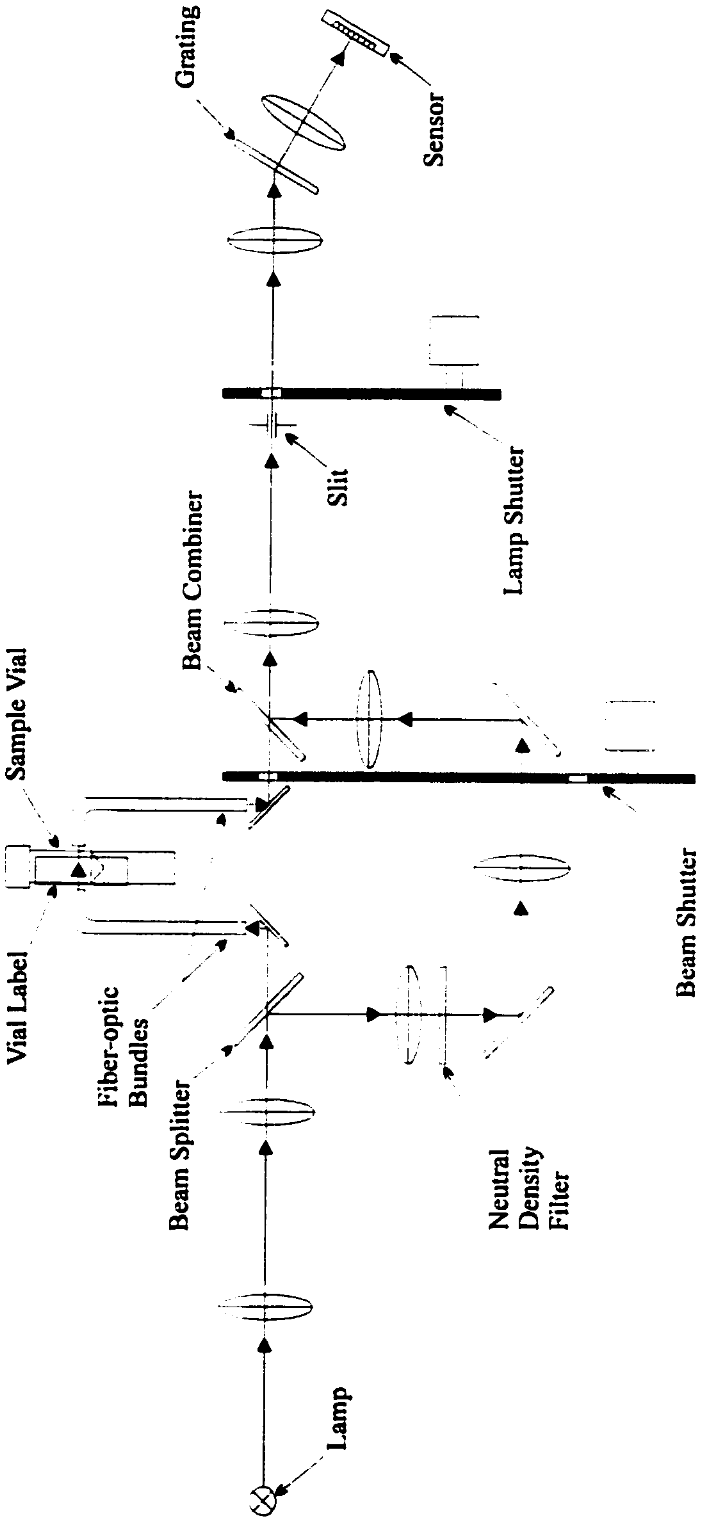


Figure 1.4: Representative Optical Schematic of CASS

in optical gain of the two paths are used to compensate the absorbance measurement[8]. The pixels were calibrated with respect to their wavelength correspondence. The wavelength calibration equation is

$$\lambda(pixel) = \lambda_1 + \lambda_{inc} \times pixel \tag{1.7}$$

where *pixel* is the pixel number from 1 to 255, λ_1 is the wavelength in nanometers for pixel 1, λ_{inc} is the wavelength step in nanometers/pixel between adjacent pixels, and $\lambda(pixel)$ is the wavelength in nanometers for pixel number *pixel*. The wavelength calibration values for this particular investigation are: λ_1 = 602.8259 nanometers and λ_{inc} = 1.729639 nanometers/pixel. The wavelength calibration table is found in Appendix A. The model and serial numbers for the CASS instrument were NIM-DBVT1000 and 9205-4004 respectively.

1.6.2 Measurement Protocol

Specimens were collected from Credit Valley Hospital, Mississauga, Ontario over a one week period. Forty-nine specimens each of urine, serum, and iodoacetate were obtained; fifty specimens each of citrate, EDTA and oxalate were obtained; and fifty-one specimens each of heparin were obtained, providing 348 specimens altogether. The specimens were refrigerated prior to spectrophotometric measurement. The specimens were measured in seven batches of approximately fifty on five different days over a seventeen day period, after the first specimens were obtained. The specimen distribution with respect to specie, for the seven batches, is shown in Table 1.4. The specimens in each batch were measured in random order.

[8] However, the absorbances were not calibrated using standard references and an arbitrary bias offsets the absorbances. The untransformed absorbances should not be compared with standard absorbances. Spectrophotometric measurements are comparative or secondary measurements and as such require calibration against a standard measure of the specimen feature in question. Therefore, it is typically not required to standardise the absorbances as these are intermediate measures of the specimen. In other words, in practice, steps are avoided if nothing is gained by them. Absorbances of high precision are usually more important than accurate absorbances in practical calibration.

Table 1.4: Specimen Distribution by Batch

Batch	Specimen Type	Number of Specimens
1	urine	19
	citrate	20
	oxalate	11
2	heparin	11
	EDTA	10
	serum	9
	oxalate	9
	iodoacetate	11
3	iodoacetate	10
	heparin	10
	EDTA	10
	serum	10
	urine	10
4	citrate	10
	iodoacetate	10
	oxalate	10
	serum	10
	EDTA	10
5	heparin	10
	serum	10
	oxalate	10
	iodoacetate	10
	urine	10
6	heparin	10
	serum	10
	EDTA	10
	iodoacetate	9
	citrate	10
7	urine	10
	EDTA	10
	citrate	10
	heparin	10
	oxalate	10

Quality control measurements were performed before and after each batch. Five measurements each of an empty vial and a distilled-water-filled vial were performed as QC measurements.

Specimens for the batch being measured were removed from the refrigerator one hour prior to measurement to acclimatise. The specimens were transferred from their storage vials to the measurement vials. Each specimen was numbered and specimen information recorded. The Auto-gain feature on the CASS instrument was used for the specimen spectral scan while fixed gain was used for the reference optical scan.

Thirty-two scans were averaged for each measurement. Each specimen was measured once (i.e. one mean of thirty-two scans). Specimen number, gain, filename and any observed abnormalities were recorded in a logbook.

1.6.3 Data Description

Each spectroscopic measurement of a specimen is an absorbance measurement comprised of 255 separate wavelengths from 602 to 1042 nanometers. Therefore, each spectroscopic measurement consists of a vector having 255 elements (or pixels).

1.7 Data Analysis Methodology

1.7.1 General Approach

The approach used here for analysing the data involves four steps:

1. Pre-processing
2. Calibration
3. Validation
4. Optimisation

The Pre-processing phase consists of transforming the data so as to emphasise and "show" the signal. The Pre-processing phase also consists of detecting outliers which may occur due to specimen problems, instrumentation problems, or operator error.

The Calibration phase consists of calibrating or classifying training or known specimen data to produce a model that can be used to predict or classify unknown specimens.

The Validation phase consists of applying the model determined in the Calibration phase to classify unknown specimens.

The Optimisation phase consists of seeking out the most parsimonious solution (i.e., one which balances optimum classification rate with simplicity, according to a criterion).

To facilitate the validation of the classification methods, the specimen data were divided into two parts: training data and prediction data. The training data are used to calibrate the classification method to form a model. The prediction data are used to check the validity or predictability of the model. The prediction data consist of 140 specimens (i.e., 20 per specie) while the training data consist of approximately 200 specimens, depending on the number of outliers removed. The prediction data are created by randomly choosing 20 specimens from each of the species. The rest are used for the training data. Three data sets of training/prediction data are created from the 348 specimens so that the results are not affected by specific set anomalies. All specimens are included in each data set, with the assignment of each specimen, to either training or prediction, performed in a random manner. Therefore, each data set contains the same (i.e., all) specimens but individual specimen assignments (i.e., to training or prediction) may be different. The three data sets are used to evaluate the methods with respect to their ability to correctly classify unknown specimens.

Another approach, if more specimens per specie were available, would be to divide up the data into three data sets but with different specimens in each data set. Using this approach, the training data from one data set could be used to predict on the training and prediction data from another data set. This would force each specimen to be part of both, a training data set and a prediction data set. The present approach does not force a specimen to be part of both. However, statistically, most specimens will be represented in both a training data set and a prediction data set.

Figure 1.5 shows the general approach to calibration while Figure 1.6 shows the general approach to validation. The first step is to transform the absorbance data. The second step involves identifying and removing outliers from the data set. The third step, for calibration, involves creating a model using the training data. The first two steps for validation are the same as for calibration, with the third step being different. The third step, for validation, involves applying the model to the prediction data.

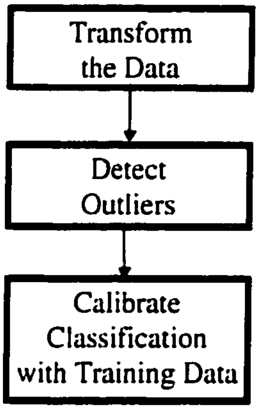


Figure 1.5: General Approach to Calibration

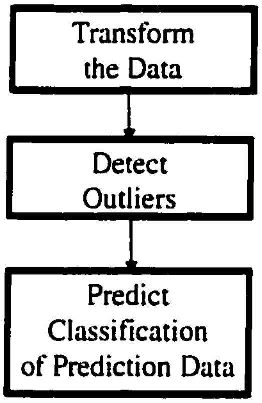


Figure 1.6: General Approach to Validation

The methodologies were developed using either Pirouette™ or Matlab™ software. Pirouette™ is an end-user software package incorporating built-in classification and data transformation functions. Matlab™ allows the user to program the user's own algorithms. KNN was implemented using Pirouette™, while MD and GA-MD were implemented using Matlab™.

1.7.2 Pre-processing of the Data

Pre-processing of the data involves transforming the data and detecting outliers. Data transformations and outlier detection are introduced in the following two sections.

1.7.2.1 Data Transformations

Spectroscopists utilise a number of data transformations including: derivativation (technically, differentiation in the case of discrete wavelengths), smoothing, autoscaling, multiplicative scatter corrections, and principal components derivations. Reasons for transforming spectroscopic data include: reduction of noise, reduction of the effects of scatter (i.e., offsets that may be dependent or independent of wavelength) or increased resolution of overlapping bands (since NIR responses tend to be broader[9] than the instrument adjacent wavelength resolution, and may also overlap with each other), such that the signal is emphasised.

Derivativation involves calculating the derivative (i.e., this can be the 1st, 2nd, or higher order) of each absorbance spectrum. First derivativation will remove additive offsets independent of wavelength while second derivativation will remove additive offsets that change linearly with wavelength. Therefore, derivativation can minimise the effects of scattering. Derivativation can also accentuate sharp spectral features, thus helping to resolve overlapping spectral bands [17, p.335].

Smoothing is often done in conjunction with derivativation as the later tends to amplify high frequency noise. The smoothing and derivativation calculations can be done simultaneously.

Autoscaling involves scaling the absorbance spectra to produce absorbance spectra that have a mean of zero and a variance of one at each wavelength.

Multiplicative scatter correction (MSC), as proposed by Geladi et. al. [18], attempts to correct the scatter of each spectrum to that of an "ideal" spectrum, usually the mean

[9] Although molecular vibrations *in vacu* produce quantified frequency responses, vibrations in the condensed (e.g. liquid) state are much more complex. This complexity is due to the partial electrostatic bondings between atoms of the various molecules, which creates many frequency combinations. An apparent broadening of the NIR response is the result. Also, the optical resolution of the instrument may broaden the response beyond the instrument adjacent wavelength resolution.

spectrum. It involves finding the best fit line for each spectrum by linear regression and then correcting each spectrum such that its best fit line matches those of the "ideal" spectrum. Therefore additive effects, both independent and varying linearly with wavelength, can be minimised. Depending on the data, the effect of applying MSC can be similar to that of a second derivativation.

Principal components analysis involves finding new data dimensions or features that are uncorrelated with each other, and are ordered in descending order with respect to variation [19, p. 76]. Considering that absorbance spectral data often have high correlation with respect to adjacent wavelengths (i.e., multicollinearity), the dimensionality typically can be reduced, i.e., the number of principal components required will be less than the number of wavelengths required. It is possible to transform many wavelengths into a few principal components. Outliers can be detected using principal components analysis; also physical interpretation of the principal components is possible.

The data were analysed using all of the transforms mentioned here except for MSC. Considering that second derivativation (2D) can have a similar effect as MSC, 2D is simpler to implement than MSC, and 2D can be combined with smoothing, MSC was not implemented. While MSC maintains the "appearance" of the spectra, second derivativation enhances the signal to the observer.

1.7.2.2 Outlier Detection

An outlier is a specimen measurement that is considered erroneous or atypical. Erroneous outliers can occur due to operator errors, instrumentation problems or specimen problems. For example, some of the plasma specimens in this study experienced clotting, which led to measurements very different from the typical specimen's. This problem resulted in errors in absorbance spectra. Problems which lead to extreme absorbance effects can be identified and the measurement flagged as erroneous. Corrective measures can then be implemented.

Outliers were identified by studying the transformed spectra and the principal component scores. The principal component scores can be used to automatically detect outliers, as is demonstrated in Section 2.4.

1.7.3 Classification Methods

A number of classification methods have been used in NIR spectroscopy for classifying species, including: Bayesian [20], K-Nearest Neighbour, Mahalanobis Distance, Information Theory [21], and Soft Independent Modelling of Class Analogy (SIMCA). Bayesian and SIMCA are suited to specimen-rich situations, i.e., where there are many specimens per class, while KNN is suited to specimen-poor situations, i.e., where there are few specimens per class. The distance measure for MD is really an extension of that used for KNN, making it suitable to specimen-medium situations. Information Theory methods could be applicable but are optimally suited to qualitative or mixed-mode data. Therefore, KNN and MD were chosen as the classification methods for this study. Specifically, the KNN method provides a baseline for the MD and GA-MD methods.

1.7.3.1 K-Nearest Neighbour Classification Method

K-nearest neighbour (KNN) is a nonparametric similarity distance measure [22, pp. 303-322], calculated between specimens. The distance between a specimen and another in the data set is calculated according to the Euclidean distance given by

$$d_{ab} = \sqrt{\sum_{j=1}^{N} (A_j - B_j)^2}$$

where d_{ab} is the distance measure between specimen a and specimen b, A_j is the absorbance for specimen a at wavelength j, B_j is the absorbance for specimen b at wavelength j, while N is the number of wavelengths.

To classify an unknown specimen, the distances between the unknown specimen and all other specimens in the training set are calculated. An appropriate number of neighbours is chosen to have voting privileges. The maximum number of neighbours was chosen to be 10 for this application, considering there are approximately 20 to 30 specimens in the training set for each specie. For a given number of neighbours, K, the K closest training neighbours to the unknown specimen each votes once for its specie. The specie with the most votes is chosen to be the specie for the unknown specimen. When there is a tie, the shortest accumulated distance is used to break the tie.

To determine calibration error, each specimen in the training set is held-out in turn and classified according to the rest of the training specimens. This is called leave-one-out cross-validation procedure.

Chapter 3 provides more detail on KNN.

1.7.3.2 Mahalanobis Distance Classification Method

This method determines a distance measure which is an extension of the Euclidean distance. The Mahalanobis Distance (MD) [19, p.62-63] is more complex and includes the variances and covariances of the dimensions as given by

$$D^2 = (\mathbf{x}-\mathbf{x}_i)' \mathbf{M}(\mathbf{x}-\mathbf{x}_i)$$

where $\mathbf{x}$ = point in the dimensional space for a particular specimen, $\mathbf{x}_i$ = centre of specie i, $\mathbf{M}$ = inverse covariance matrix for the dimensions, and D = Mahalanobis distance (i.e., MD) from specimen to specie i.

The same basic equation can be used to calculate the distance between the specie centres, i.e., D_{ij}^2 for species i and j.

The MD has several advantages over an equivalent Euclidean distance measure, these being:

- The MD takes into account not only the centre of the specie, but the distribution about the specie centre. Therefore, if there are significant differences in variance between species or between dimensions, these will be accounted for by normalising the distance calculation.
- The MD can be interpreted more universally since it is a normalised distance (i.e., unit distance vector with N dimensions).

  e.g., an MD of 6 between two species indicates the species' centres are separated by 6 standard deviations (i.e., 3 standard deviations each).
- a rule of thumb: in practice a MD of 6 between a specie and all others is required to achieve 95% predictability. Therefore the MD will give you an idea how well an unknown specimen can be predicted.

Chapter 4 describes MD in more detail.

1.7.4 Genetic Algorithm Selection of Wavelengths

Genetic Algorithms (GA's) are classified as "simulated evolution" methods, that is, problem solving methods that simulate the natural evolution process as theorised by Charles Darwin in 1859. The general components contained in the natural evolution theory are mimicked in the GA: an ensemble or population of creatures or members, competition and selection of the members on the basis of some observable fitness quality, reproduction and parenting of the members and modifications or mutations of the members. Although the general principles for genetic algorithms were in place over 100 years ago, GA methodology is still in its infancy, with the first international conference held in 1985 and the first comprehensive textbook published in 1989 [24].

GA's are used as global search methods to find optimum solutions. GA's are starting to be used in chemometrics to select wavelengths as proposed by Lucasius [25], Leardi [26], and others. Lucasius' study of three wavelength selection methods--GA, simulated annealing and stepwise elimination--showed that GA's generally performed the best [25, p.263]. One of the drawbacks of a GA is that it is an interactive technique and thus very computationally intensive. One analysis can literally run for days in optimising the calibration. However, computational speed can be increased in the prediction phase if the number of features required are reduced. Therefore, computational speed may suffer in the calibration phase but will improve in the prediction phase, which typically has greater time constraints thrust upon it.

In this investigation, genetic algorithms were used to optimise the Mahalanobis Distance calibration by selecting the best features or, in this case, wavelengths.

The general flow of a genetic algorithm is shown in Figure 1.7. The steps include: Generate an Initial Population, Evaluate the Population with respect to a Fitness Criterion (Terminate if Conditions Met), Replace the Worst Members with the Best Members, Recombine the Members to Parent a new Population, Mutate the Population Members.

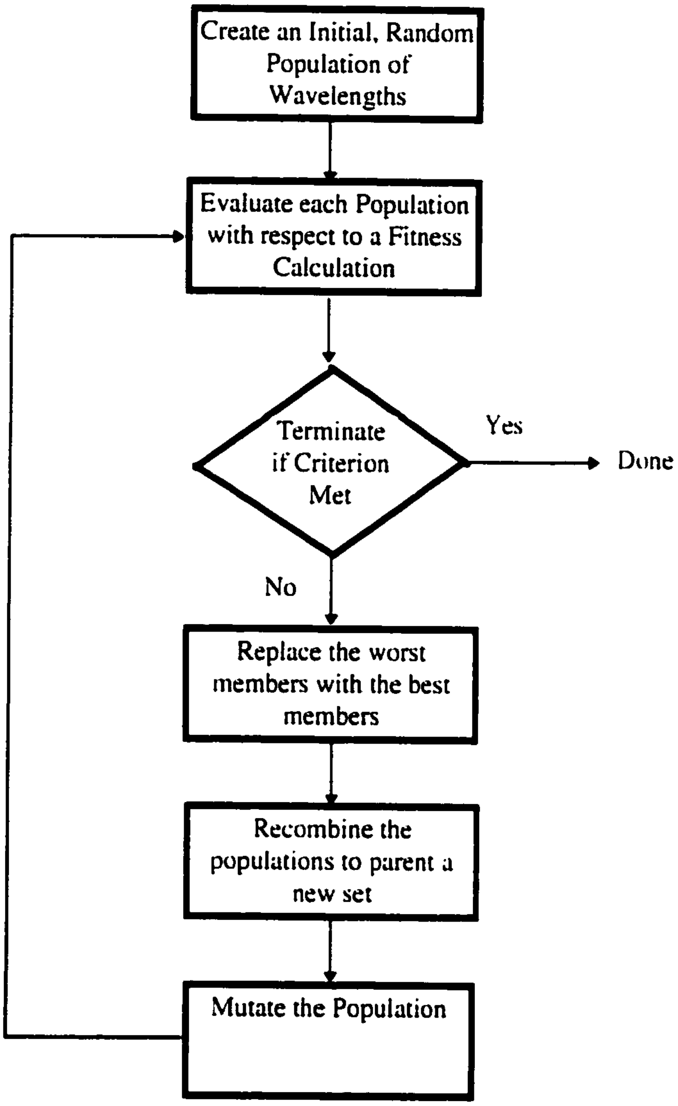

Create an Initial, Random Population of Wavelengths
Evaluate each Population with respect to a Fitness Calculation
Terminate if Criterion Met
Yes
Done
No
Replace the worst members with the best members
Recombine the populations to parent a new set
Mutate the Population

Figure 1.7: Genetic Algorithm to Select Wavelengths

The genetic algorithm steps are described below.

Initial Population: Each member of the population consists of a solution vector (i.e. wavelength selection vector) whose elements select or deselect a particular wavelength. Each element's status is determined randomly for the initial population based on a predefined distribution for selection or deselection. Population size varies according to the number of features or wavelengths. Generally, too small a size will result in quick convergence to a possible non-optimal solution, while too many will result in excessive computation. An enumerative search method would suggest that approximately 2^n subsets be scanned, where n is the number of features (in our case wavelengths). For 255 wavelengths, the amount of computation would be prohibitive. Therefore, a practical implementation for a GA, for general spectroscopic problems, is really a non-local search method that finds an adequate local solution [25, p. 263]. Typical population sizes vary from 30 to 200 [26, p. 77] members with convergence typically occurring in 100 to 500 generations.

Fitness Criterion: A fitness value or quality is calculated for each member of the population. The purpose of the fitness criterion is to direct selection of members from one generation to the next. The fitness criterion includes a factor related to the error rate and a factor related to the number of wavelengths. In this way a parsimonious solution can be achieved. An example of a fitness criterion is shown below:

$$\mathrm{Fit} = 1/\ (\sum_i Error_i)(N)$$

where $Error_i$ is the number of errors associated with predicting specie i , and N is the number of wavelengths selected. The errors are summed over the number of species. This particular criterion would be maximised by the algorithm. Also, the fit criteria are evaluated on their ability to predict unknown specimens. If the "error" factor is too dominant, then the calibration will tend to overfit. If the "wavelength" factor is too dominant, then the calibration will tend to underfit. A balance of these two factors provides the most parsimonious and useful solution.

Replacement: This step allows preservation of the "good" population members and disposition of the "bad" population members. The population members are ordered with respect to the calculated fit value. A certain proportion of the worst members are replaced with the top members.

Recombination: The purpose of this step is to help identify better solutions while maintaining the gains made in the GA selection process. To a certain extent, it is a local exploration; although if there is enough diversity in the population, a large portion of the solution space can be explored by recombination alone. This step involves pairing up members (i.e., parents) randomly, then replacing the parents with offspring derived from the parents. The most common recombination operator is the cross-over. Cross-over involves defining cross-over points randomly. Up to the first cross-over point the string elements are kept the same. The string elements between the second and third cross-over points are swapped between the two parents. The string elements between the third and fourth cross-over points are kept the same. This process continues in like manner to the end of the strings.

Mutation: The purpose of this step is to maintain diversity in the population so that the whole solution space can be explored. It is possible for the status of a particular wavelength to be the same for all population members. An optimal solution may be missed if this wavelength is always included or always excluded. Mutation prevents this loss of diversity by randomly toggling the status of a certain proportion of elements in the population. The mutation rate is typically low, on the order of 0.001 to 0.05.

Chapter 5 provides more detail on the GA implemented.

Chapter 2

Pre-processing of the Data

2.1 Introduction

Transforming or pre-processing data has several purposes: to emphasise the signal component, de-emphasise the noise component, or reduce the dimensionality or feature space. This chapter discusses the data transforms used in this investigation: derivativation, smoothing, autoscaling, and principal components derivation or analysis. The applicability of each transform is based on its ability to enhance or "show" the signal, and ultimately the transform's effect on the ability to predict unknown specimens. The effect on the KNN calibration error rate was also evaluated to select the best data transform (from those studied), as illustrated in Section 2.2.4. The transforms were tested individually and in combination with each other.

2.2 Untransformed Absorbance Spectra

The raw or untransformed data, based on the spectrophotometric measurement, are shown in Appendix B. The untransformed data for urine are shown in Figure 2.1. It is clear from these figures that there is large variability among measurements within a specie and that there appears to be few distinguishable differences between the species.

Therefore, it is desirable to transform the data to minimise the variability within a specie and emphasise distinguishing characteristics between the species. It appears from Appendix B and Figure 2.1 that, at least, a bias correction independent of wavelength would be useful. Therefore, a first or second derivativation is likely required.

The absorbance spectral data are of the form **A**, an $m \times n$ matrix, whose individual elements, a_{ij}, correspond to the absorbance for specimen i and wavelength j (i.e., a row of **A** is the spectrum for a given specimen). The data are arranged such that the first m_1 rows are for specimens belonging to specie *1*, the next m_2 rows are for specie 2 and so on with specie *7* data residing in the last m_7 rows.

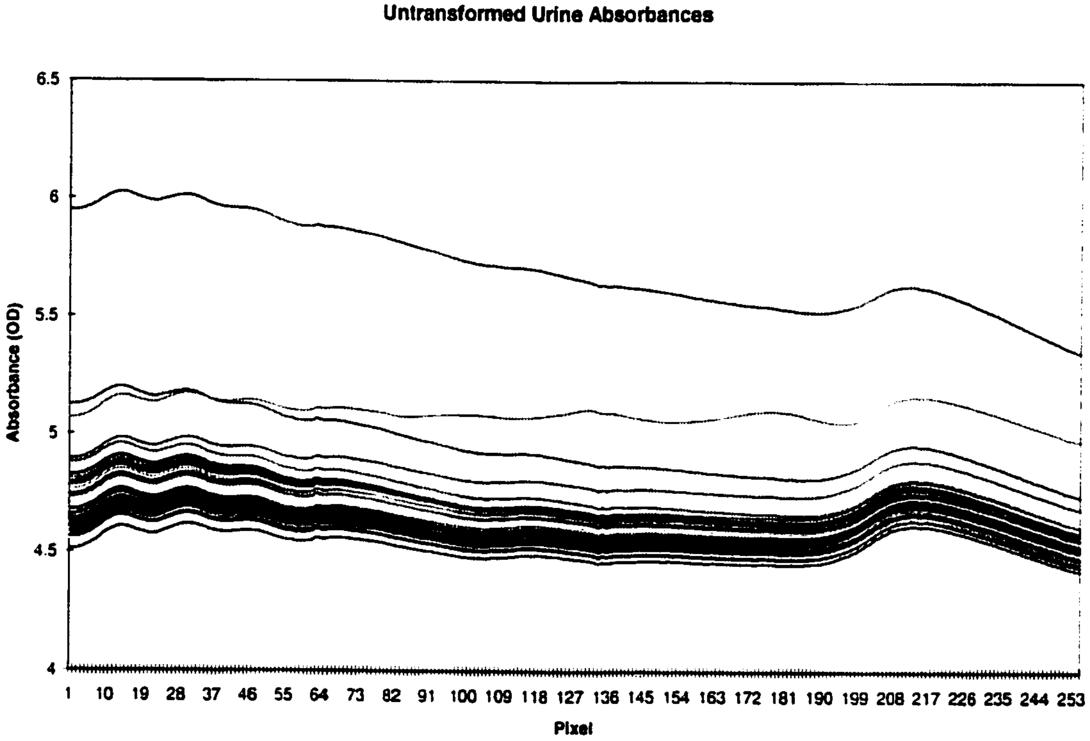


Figure 2.1: Untransformed Absorbance Spectra for Urine

2.3 Data Transformations

2.3.1 Derivativation and Smoothing

Performing derivativation on data will remove offsets (1st derivative) and slopes (2nd derivative). However, derivativation tends to amplify high frequency noise, making smoothing a necessary complementary function. The derivativation and smoothing functions performed here are based on a Savitzy-Golay (S-G) [28] polynomial filter as corrected by Steiner et. al. [29] with end point modifications proposed by Gorry [30]. The S-G filter applies a convolution to a data point by considering the *n* points on either side of the given or centre data point. Savitzky and Golay have shown that the coefficients that they have derived for the convolution window produce exactly the same result as doing a least squares fit to an rth order polynomial using the same, 2*n* + 1, data points, for the centre point. The filtering implemented in this investigation used a second order or quadratic polynomial. For example, the convolution coefficients for an 11 (i.e., *n* =5) point (quadratic) window are shown in Table 2.1.

Table 2.1: Convolution Coefficients for 11 point Smooth and 2nd Derivative

Point	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	*Norm*
Smooth	-36	9	44	69	84	89	84	69	44	9	-36	429
2nd Derivative	15	6	-1	-6	-9	-10	-9	-6	-1	6	15	429

Appendix C shows the spectral data transformed using an 11 point (quadratic) smooth and 11 point 2nd derivativation; this transform will be henceforth referred to as transform SM11-2D11. Figure 2.2 shows the transformed data for urine based on the aforementioned transform. This transform significantly reduces the intra-specie variability. Figure 2.3 shows the median transformed (i.e., SM11-2D11) absorbances for each specie after subtraction with the median QC water absorbance. It is clear from this graph that urine is easily distinguished from blood specimens at wavelengths above 878 nm (i.e., pixel 160) to 1042 nm (pixel 255). Oxalate and iodoacetate show strong differences from the other species at wavelengths between 602 nm (pixel 1) and 635 nm (pixel 20). Oxalate also shows spectral differences from the other species at wavelengths around 912 nm (pixel 180). It is not clear if significant differences exist to separate serum, citrate, EDTA and heparin from each other. Spectrally, in this range, these species look very similar to each other.

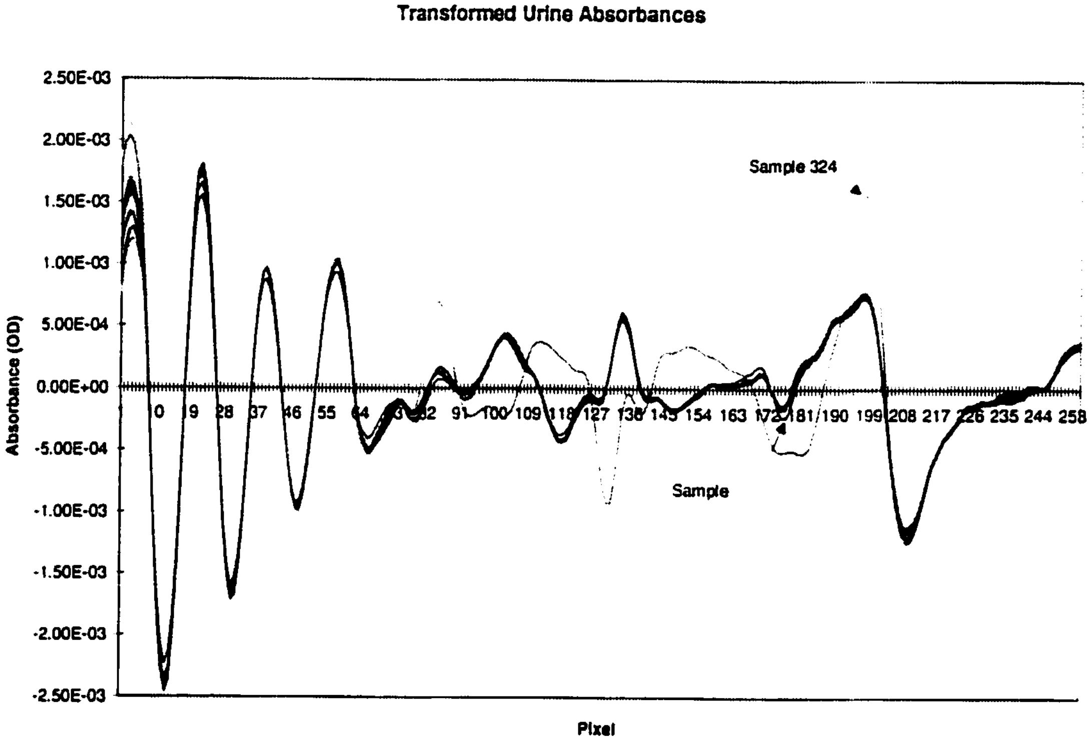


Figure 2.2: Transformed (No Autoscaling) Absorbance Spectra for Urine

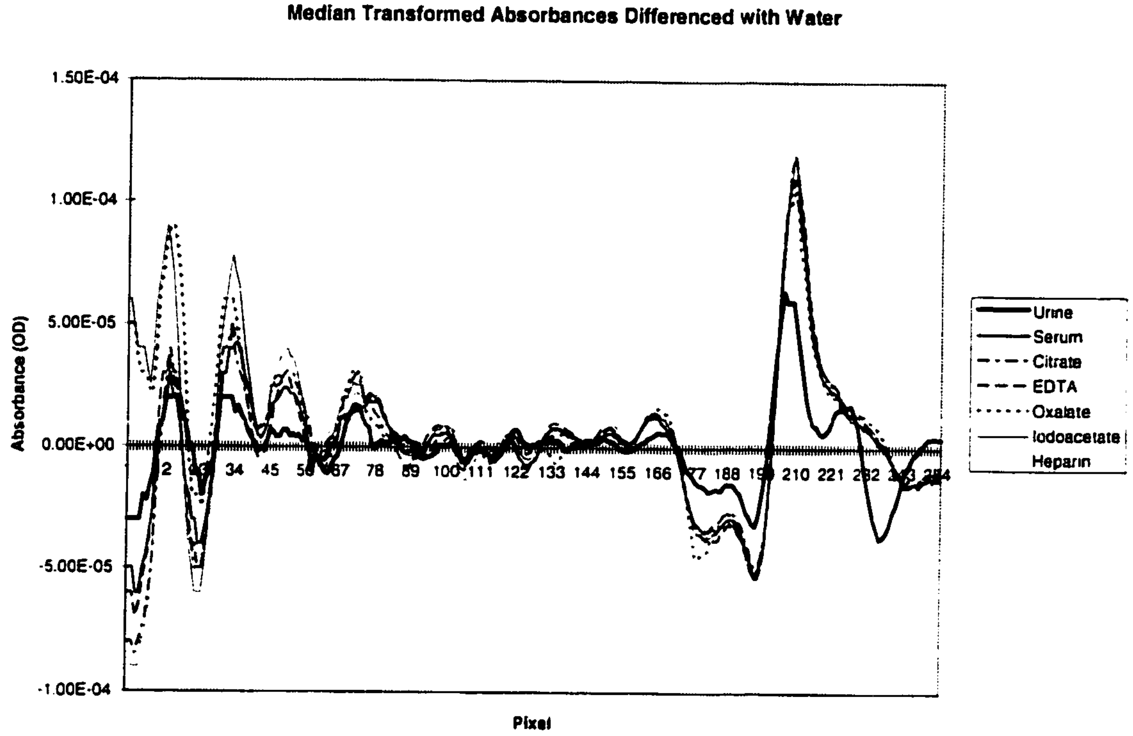


Figure 2.3: Median of Transformed (SM11-2D11) Absorbance Spectra for each Specie differenced with Median of QC Water Absorbances

2.3.2 Autoscaling

Transforming data by autoscaling will normalise the mean of each feature (i.e., wavelength) to zero and the variance to one. Data that exhibit large statistical differences between features may profit from autoscaling by normalising the leverage between the features. Autoscaling is performed in two steps: mean-centring and variance scaling. All the specimens in the training data are used to calculate the mean and variance with respect to wavelength. The mean and variance calculated from the training data are then used to transform both the training and prediction data. Mean-centring transforms the data such that the new mean is zero for all wavelengths. Variance scaling transforms the data such that the new variance is one. The steps are described below.

1. Calculate the mean.

   The mean, $\overline{a_j}$, is calculated using equation 2.1 below:

$$\overline{a_j} = \frac{1}{m}\sum_{i=1}^{m} a_{ij} . \tag{2.1}$$

2. Calculate the variance.

   The variance, s_j^2, is calculated using Equation 2.2 below:

$$s_j^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(a_{ij} - \overline{a_j}\right)^2 . \tag{2.2}$$

3. Apply the autoscale parameters to the data.

   The autoscaled data, $a_{ij(as)}$, are calculated by subtracting the mean from the original data and then dividing by the standard deviation, as shown below in Equation 2.3:

$$a_{ij(as)} = \frac{a_{ij} - \overline{a_j}}{s_j} . \tag{2.3}$$

Figure 2.4 shows transformed (SM11-2D11) and autoscaled spectra for urine. Figure 2.4 shows a significant reduction in the differences between the means of the features (i.e., wavelengths) with possibly a small reduction of variance between the wavelengths, as compared to Figure 2.2. Therefore, autoscaling may help the performance. However, autoscaling is sensitive to outliers (this is why Figure 2.4 is shown with outliers removed) and may not help if statistical differences between features are not significant. In addition, the parameters are based on the statistical characteristics of the data (i.e., mean and variance) which may change over time if instrumental drift and environmental changes exist (and are not compensated for).

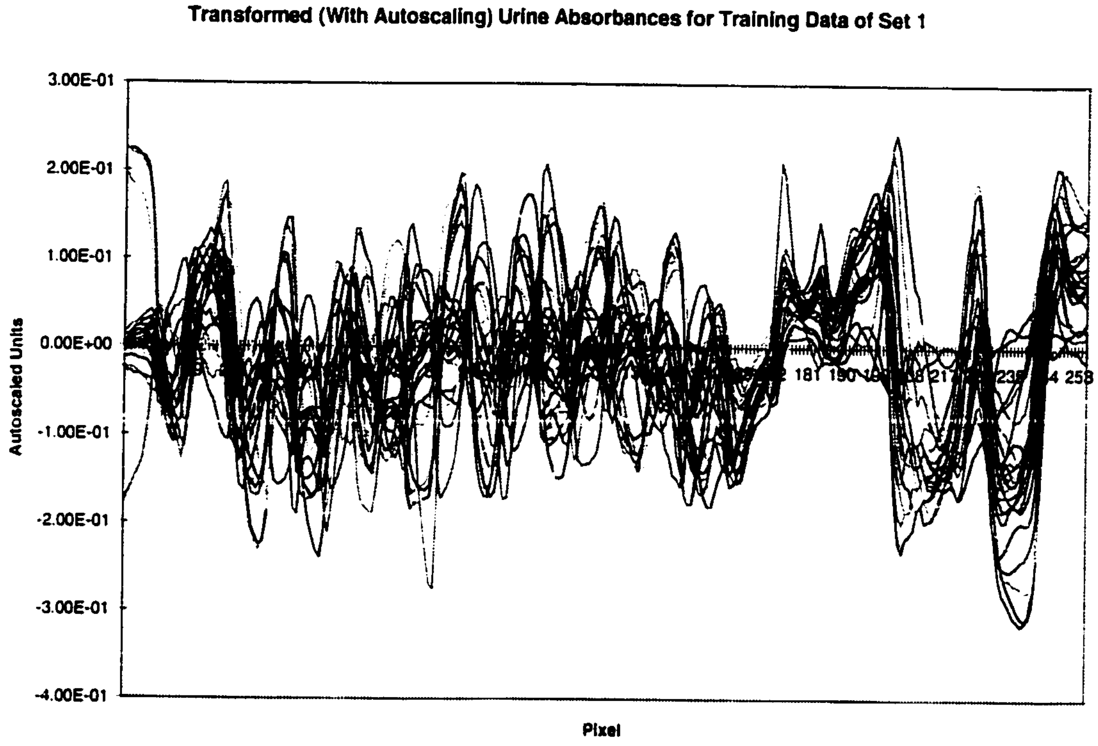


Figure 2.4: Transformed (SM11-2D11) and Autoscaled Absorbance Spectra for Urine with Outliers Removed for Training Data Set 1

2.3.3 Principal Components Analysis

The purpose of transforming data using principal components analysis is to reduce the dimensionality or number of features. A large number of wavelengths may be transformed into a relatively small number of principal components. Principal components analysis (PCA) involves expressing the variables $\mathbf{A} = \{\mathbf{a}_j, j = 1,2,\ldots,n\}$ (in our case absorbance spectra) in terms of a lower number of new, uncorrelated variables $\mathbf{T}_s = \{\mathbf{t}_k, k = 1,2,\ldots,s\}$, that is, the principal components or scores. This is accomplished such that: $s < n$; the $\mathbf{t}_k$'s are orthogonal to each other; $\mathbf{t}_1$ accounts for the largest amount of variation in the data and each subsequent $\mathbf{t}_k$ accounts for less and less variation, with $\mathbf{t}_s$ accounting for the smallest amount of variation in the data. The main advantage of PCA is that a large number of variables can be compressed into a fewer number of variables that explain the data adequately. If the original variables are uncorrelated with each other then PCA will not be an advantage. However, spectral data are highly correlated, making the reduction of dimensionality possible without a significant loss of information.

To determine $\mathbf{T}_r$ from $\mathbf{A}$, the following relationship, as defined by Equation 2.4 is used:

$$\mathbf{A} = \mathbf{T}\mathbf{L}' \tag{2.4}$$

where $\mathbf{T}$ is the uncompressed $m \times n$ matrix for the principal components, $\mathbf{L}$ is the uncompressed weighting or loading matrix such that each principal component is a linear combination of the variables $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$. As an example, the expression for principal component 1 (PC 1) is $\mathbf{t}_1 = l_{11}\mathbf{a}_1 + l_{12}\mathbf{a}_2 + \ldots + l_{1n}\mathbf{a}_n$ with t_{i1} being the PC 1 element calculated for specimen i. Equation 2.4 is re-expressed as Equation 2.5, using the compressed matrices $\mathbf{T}_r$ and $\mathbf{L}_r$ for the principal components and loadings respectively.

$$\mathbf{A} = \mathbf{T}_r\mathbf{L}_r' + \varepsilon \tag{2.5}$$

where ε represents a small error introduced by the compression.

To determine the loading matrix, $\mathbf{L}_r$, three constraints are introduced. The first one is: the sum of the squares for each row totals 1. For example, for row 1: $l_{11}^2 + l_{12}^2 + \ldots + l_{1r}^2 = 1$. The second constraint is: the variance for the first PC is the highest, the variance for the second PC is the second highest, and so on with the last PC's variance being the smallest (i.e., var($\mathbf{t}_1$) > var($\mathbf{t}_2$) > ... > var($\mathbf{t}_r$). The third constraint is: the $\mathbf{t}_k$'s must be orthogonal to each other. It can be shown [31, p.99] that the sample covariance matrix (i.e., covariance across wavelengths not specimens), $\mathbf{C}$, a symmetric $n \times n$ matrix, produces eigenvectors that fit the constraints of the loading matrix, $\mathbf{L}$, except for scaling (i.e., constraint 1). The eigenvalues of $\mathbf{C}$ are used to scale $\mathbf{C}$ and order the principal components in descending order, such that $\mathbf{L}$ is produced. The components that do not contribute significantly to the variance are truncated to produce $\mathbf{L}_r$.

To calculate the principal components matrix, $\mathbf{T}_r$, Equation 2.5 is rewritten as Equation 2.6 given that the expression, $\mathbf{L}_r'\mathbf{L}_r = \mathbf{I}$.

$$\mathbf{T}_r = \mathbf{A}\mathbf{L}_r. \tag{2.6}$$

The loading matrix is calculated for the training data and is used to determine the principal components for both the training and the prediction data, thus maintaining a consistent transformation.

Figure 2.5 shows the first 4 principal components for the data shown in Appendix C, except with autoscaling also. Appendix D shows the first 24 principal components. PC 1 and PC 7 both show strong differences with respect to classifying urine from blood. Several other PC's show weak but detectable signal differences for distinguishing urine from blood. PC 4 and, to a lesser extent, PC 5 show differences capable of classifying oxalate from the other species. PC 12 appears to be the strongest one for classifying iodoacetate from the other species. A number of PC's show strong signals for subsets of species and it may be that other PC's are required to account for interferences which weaken the "signal" for other specimens for a given specie.

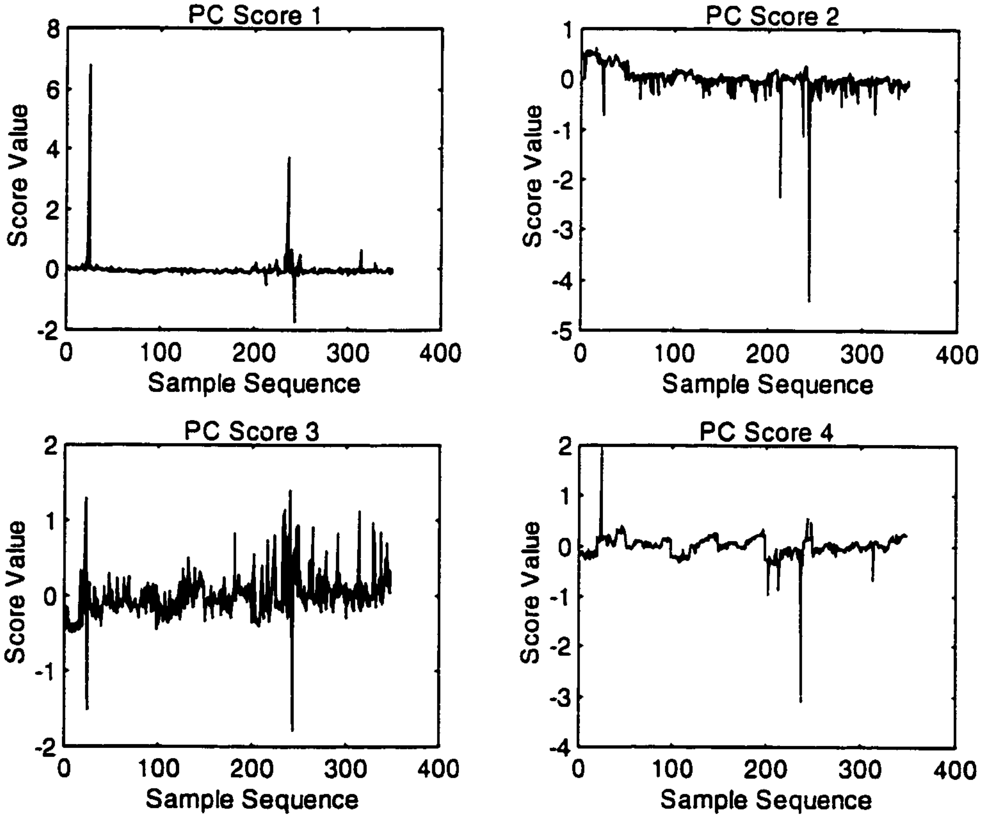


Figure 2.5: First 4 Principal Components for Transformed (SM11-2D11) and Autoscaled Data

2.3.4 KNN Calibration Error Rate Dependence on Data Transform

To select the optimum transform (i.e., SM11-2D11), the whole data set, including outliers, was analysed using KNN, with a variety of transforms and parameters. The calibration error rates of these analyses are shown in Table 2.2. The results of this show that 1st derivativation is better than no transformation and 2nd derivativation is better than the 1st derivativation. The Pirouette™ software does not go to orders higher than the 2nd derivative. Smoothing up to 11 points improves the calibration error. Smoothing beyond 11 points does not make significant improvements. Autoscaling does not appear to make much difference when smoothing is not performed. However, autoscaling did make a significant improvement to the transform SM11-2D11. These results suggest that the best transform is SM11-2D11 with autoscaling, for KNN analysis.

Table 2.2 KNN Calibration Error Rate Versus Data Transform

Derivative	Smoothing	Autoscaling	Calibration Error Rate (%)
No	No	No	66.67
No	No	Yes	66.38
1st	No	No	46.84
1st	No	Yes	47.70
1st, 5 point	Yes, 5 point	Yes	47.13
2nd	No	No	54.60
2nd	No	Yes	53.45
2nd, 5 point	Yes, 5 point	Yes	50.47
2nd, 7 point	Yes, 7 point	Yes	47.87
2nd, 9 point	Yes, 9 point	Yes	41.67
2nd, 11 point	Yes, 11 point	No	42.53
2nd, 11 point	Yes, 11 point	Yes	36.49
2nd, 13 point	Yes, 13 point	Yes	35.92
2nd, 15 point	Yes, 15 point	Yes	35.92

2.4 Outlier Detection

Errors can occur in the measurement of a specimen due to specimen problems, operator errors, and/or instrument problems. It is important to flag or identify a measurement as being erroneous (i.e., an outlier) so that classification errors are minimised. This is especially important during the calibration phase since error rates will undoubtedly be adversely affected if a calibration is based on erroneous measurements. In this study erroneous data were identified by studying the transformed spectra and by studying the principal components. Outlier detection was automated by determining appropriate thresholds for the principal components to exclude spectra. The criteria for determining when a measurement is an outlier will, in practice, be based on the training data. The criteria then are applied to prediction data to flag any outliers. However, in this study the outlier criteria were determined based on all the data, with outliers being removed before the data sets were created. In this way, the results could be compared directly, since all the data sets use the same specimens.

2.4.1 Using Transformed Spectra

Figure 2.2 shows the complete set of transformed spectra for the urine specimens. It is obvious that specimen or sample 324 is significantly different from the other spectra while sample 330 is somewhat different. Appendix C shows the spectra for the other species. Altogether, 8 specimens out of 348 were identified as outliers from visual observations of the transformed spectra. This is a 2.3% rejection rate, which is considered reasonable. It is likely this rejection rate would be smaller in practice, than that experienced in this study, as the specimens would typically be fresher at the time of measurement. The outliers identified included two urine specimens (samples 324, 330), four oxalate specimens (samples 81, 147, 148, 169) and two heparin specimens (samples 285, 305).

2.4.2 Using Principal Component Scores

From the visual observations of the transformed data, appropriate thresholds were determined for the first ten principal components such that the outliers detected visually were identified as outliers automatically by PCA. The thresholds determined are shown in Table 2.3.

Table 2.3: Principal Component Threshold Values for Identifying Outliers

Principal Components	Threshold Values
1 to 5	+/- 1.5
6 to 8	+/- 1.0
9 to 10	+/- 0.75

Spectra which exceeded these values were flagged as outliers and excluded from the data set. Altogether, ten specimens were identified, including all eight as identified by studying the transformed spectra. The additional two, identified by PCA, were both oxalate types (samples 46, 163).

The statistics of the PC scores, calculated after the outliers were removed, were used to determine the probability of accepting a "good" specimen, as shown in Table 2.4. Statistically, PC 1 will flag the most "good" specimens as outliers, at a rate of 1 in 1250 specimens. Therefore, very few normal or typical specimens will be rejected.

Table 2.4: Principal Component Statistics and Specimen Acceptance Rate

PC Score	Mean	Standard Deviation	Acceptance Rate (%)
1	0.1914	0.4468	99.9212
2	-0.0349	0.3832	99.9909
3	-0.2343	0.3218	99.9997
4	0.0429	0.2440	100.0000
5	-0.0266	0.2000	100.0000
6	0.0334	0.1910	100.0000
7	-0.0148	0.1771	100.0000
8	-0.2306	0.1605	100.0000
9	0.0949	0.1423	100.0000
10	-0.1226	0.1183	100.0000

2.4.3 Operator Observations

Table 2.5 shows the Operator's comments on the specimens identified as outliers. All outliers but two have observations recorded that could explain degradation of the specimens, and thus poor results for these specimens. There were other specimens with similar observations, but the interfering component, e.g., a clot, may not have been in the path of the spectrophotometric beam. The locations of the interfering components with respect to the spectrophotometric beam were not recorded.

Table 2.5: Operator Observations

Sample Number	Operator Observation
46	very pink
81	small clot
147	-
148	-
163	50% clotted
169	40% clotted
285	cloudy
305	cloudy
324	cloudy
330	cloudy 2+ lipemic

Chapter 3

K-Nearest Neighbour Classification

3.1 Introduction

KNN is a relatively simple classification method and is readily available in off-the-shelf software. KNN, therefore, is ideal to use for a baseline comparison against MD. KNN is a nonparametric classification method whose similarity measure is based on a multidimensional distance measure, i.e., the Euclidean distance, between the specimens. Since its discrimination measures are not based on estimated statistics of the data, it performs well in extreme situations such as sample-poor/variable-rich (even as low as one specimen per specie) environments and many class environments [22, p. 284]. The error rate is, at most, two times that of the Bayesian error; the Bayesian error being the lowest statistically possible [22, p. 310].

The closest K neighbours to an unknown specimen are used to determine the class of the unknown specimen. Figure 3.1 shows the four closest neighbours to a point of interest, for a two-dimensional case. Most of the information is contained in the first (i.e., K=1) neighbour, however, error rates can be reduced for a larger K (i.e., K=3 to 5) when there is good separability between the species [32, p. 145]. Typically, the optimal K is 1 when there is poor separability among the species.

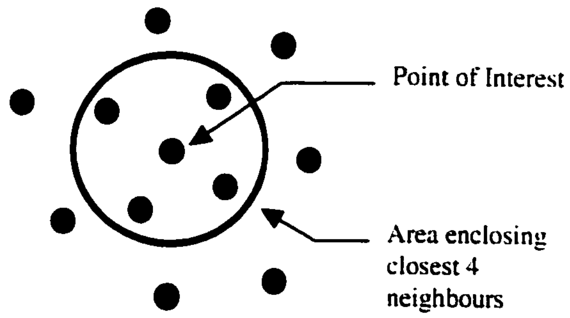


Figure 3.1: Four Closest Neighbours to a Point of Interest

3.2 Calibration using KNN

To be able to classify unknown specimens, training specimens are used to create a KNN model. Calibrating training data to create a KNN model involves 6 steps, as shown in Figure 3.2, which are: Calculate the interspecimen distances, Derive voting matrix, Determine optimal K from the misses vector, Calculate misclassification matrix, and Save the KNN model.

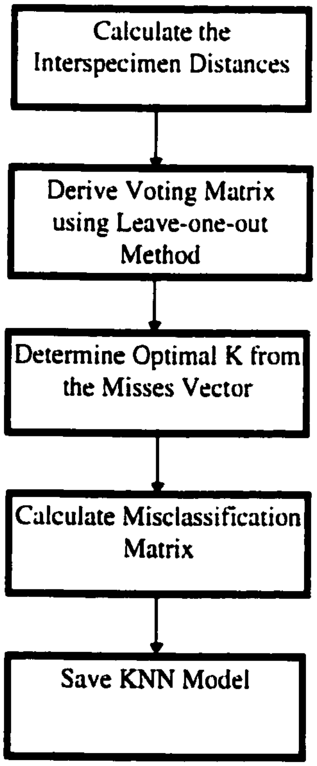

Calculate the Interspecimen Distances
Derive Voting Matrix using Leave-one-out Method
Determine Optimal K from the Misses Vector
Calculate Misclassification Matrix
Save KNN Model

Figure 3.2: Calibration using KNN

3.2.1 Interspecimen Distance Calculation

The interspecimen distances between a specimen of interest and each of the other specimens are used to determine which specimens are the closest neighbours to the specimen of interest. The distance between a specimen and another in the data set is calculated according to the Euclidean distance given by

$$d_{ab} = \sqrt{\sum_{j=1}^{N}(A_j - B_j)^2} \tag{3.1}$$

where d_{ab} is the distance measure between specimen a and specimen b, A_j[10] is the absorbance for specimen a for wavelength j, B_j is the absorbance for specimen b for wavelength j, and N is the number of wavelengths.

3.2.1.1 A Simple Cartesian Example

To help illustrate some of the calculations for KNN, MD and GA-MD, a simple Cartesian example is created. This simple example will only be used to illustrate calculations, and will not be used to compare performances of the various methods.

Suppose that we have seven data points (or specimens), A to G, which are defined in 3-dimensional space. Data points A to C belong to specie "ONE" while D to F belong to specie "TWO". The specie of data point G is unknown. Data points A to F form the training data while G forms the prediction data. Table 3.1 shows the Cartesian co-ordinates of each data point, while Figure 3.3 illustrates these graphically.

[10] If principal components are used, $\mathbf{A}$ is the transformed principal component vector for specimen a, $\mathbf{B}$ is the transformed principal components vector for specimen b, and N is the number of principal components.

Table 3.1: Cartesian Co-ordinates for Simple Example

Data Point	X Co-ordinate	Y Co-ordinate	Z Co-ordinate
A	0.34	2.00	0.51
B	0.23	2.16	0.31
C	0.12	1.85	0.66
D	0.85	1.02	0.94
E	1.13	0.86	0.98
F	1.01	1.15	1.17
G	0.45	1.82	0.71

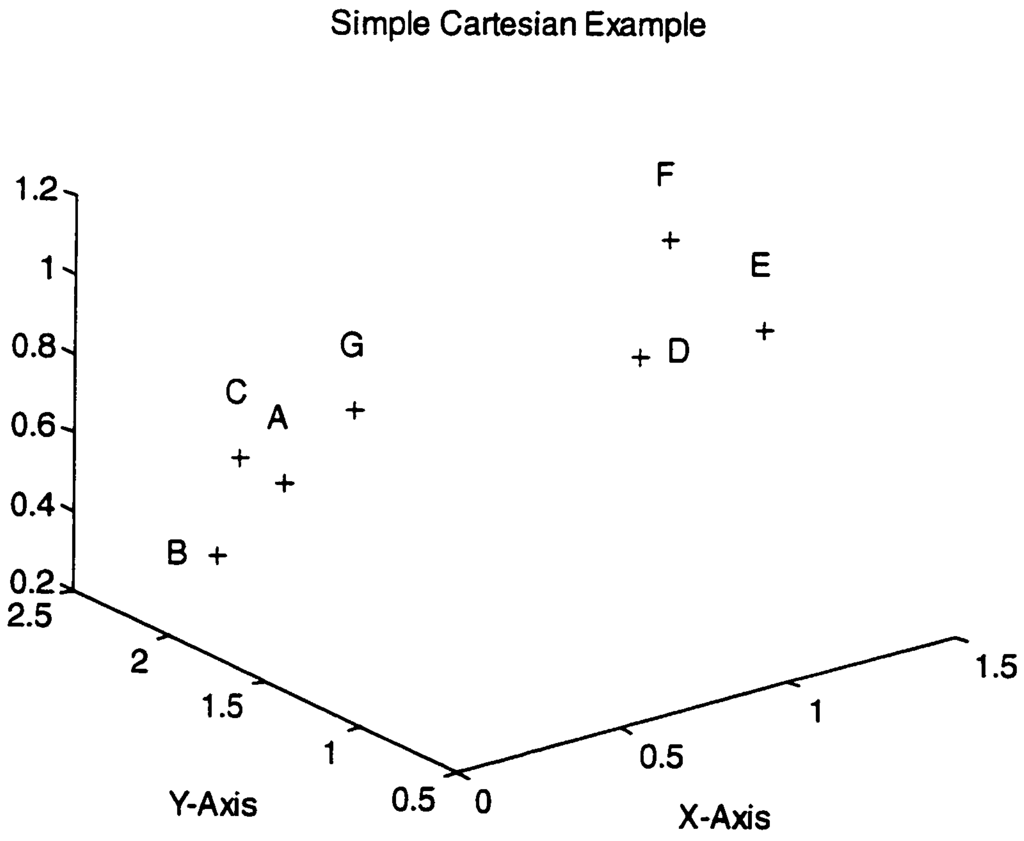


Figure 3.3: Graphical Representation of Data Points for Simple Cartesian Example

The interspecimen distances for the training data points, A to F, are calculated according to Equation 3.1 and are tabulated in Table 3.2.

Table 3.2: Interspecimen Distances for Training Data of Simple Cartesian Example

	A	B	C	D	E	F
A	0	0.279	0.306	1.186	1.464	1.268
B		0	0.480	1.443	1.717	1.539
C			0	1.140	1.450	1.242
D				0	0.325	0.309
E					0	0.367
F						0

3.2.2 Voting Matrix

The purpose of the voting matrix is to identify which specimens or data points are closest to the specimen or point of interest. The interspecimen distances are used to order the specimens with respect to the specimen of interest. The voting matrix is formed by ordering the neighbours for each specimen from closest to furthest away (based on the interspecimen distances) for the first *K* neighbours. The voting matrix for the Simple Cartesian Example is shown in Table 3.3. This matrix is derived from the interspecimen distances in Table 3.2.

Table 3.3: Voting Matrix for Training Data of Simple Cartesian Example

	K1	K2	K3	K4	K5	K6
A	B (ONE)	C (ONE)	D (TWO)	F (TWO)	E (TWO)	-
B	A (ONE)	C (ONE)	D (TWO)	F (TWO)	E (TWO)	-
C	A (ONE)	B (ONE)	D (TWO)	F (TWO)	E (TWO)	-
D	F (TWO)	E (TWO)	C (ONE)	A (ONE)	B (ONE)	-
E	D (TWO)	F (TWO)	C (ONE)	A (ONE)	B (ONE)	-
F	D (TWO)	E (TWO)	C (ONE)	A (ONE)	B (ONE)	-

Table 3.4 shows the voting matrix for fifteen urine specimens. This was based on the transform: SM11-2D11 and autoscaling.

Table 3.4: Voting Matrix for Transformed Urine Specimen Data

	k1	k2	k3	k4	k5	k6	k7	k8	k9	k10
10	1	1	1	1	1	1	1	1	1	1
321	1	1	1	1	1	1	1	1	1	1
14	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1
343	1	1	1	1	1	1	1	1	1	1
336	1	1	1	1	1	1	1	1	1	1
341	1	1	1	1	1	1	1	1	1	1
327	1	1	1	1	1	1	1	1	1	1
11	1	1	1	1	1	1	1	1	1	1
12	1	1	1	1	1	1	1	1	1	1
3	5	5	5	5	5	3	3	3	3	3
7	1	1	1	1	1	1	1	1	1	1

To validate the calibration of the training data, the leave-one-out method is used. Each specimen is left out of the training data and the rest are used to "predict" the specie for the "left-out" specimen. For a given *K*, the predicted class or specie is the one which receives the most votes, where each nearest neighbour, k1 to k*K*, is allowed one vote. In the case of a tie, (i.e., in the case where *K* is even) the summations of the interspecimen distances for the tied species are used to break the tie. The specie with the lowest interspecimen distance sum gets assigned to that specimen.

3.2.3 Optimal *K* based on the Misses Vector

To optimise the results, the number of nearest neighbours which produces the best validation results, for the training data, is used. The optimal number of nearest neighbours minimises the number of misclassifications.

The total number of misclassifications is determined for K=1 to K=K_{max}. An arbitrary high number for K_{max} (i.e., 10) is chosen at the time of calibration. If the number chosen proves to be insufficient a higher number is then chosen and the calibration is re-run. Figure 3.4 shows the misses vector for transformed (i.e., SM11-2D11, autoscaled) training data set 1 (three data sets were created from the data as defined in Section 1.7.1. The optimal *K* for this calibration was 1. This suggests the class separation was not high. However, the optimal *K* for the highly separable, simple Cartesian example was also 1 or 2, but it could have been higher if more specimens were used.

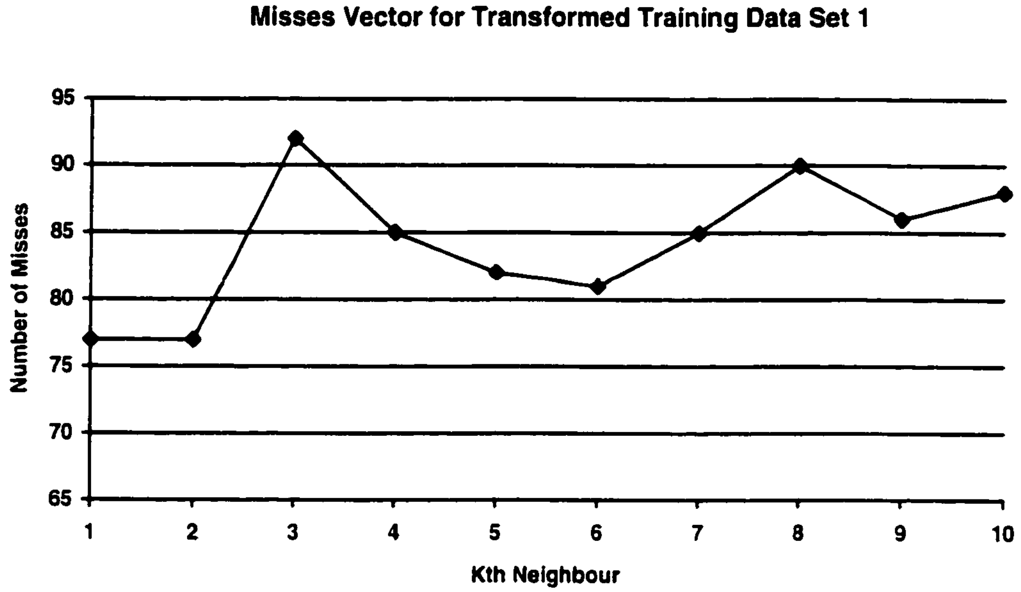


Figure 3.4: Misses Vector for Transformed Training Data Set 1

3.2.4 Misclassification Matrix

The misclassification matrix breaks down the errors by specie. An example, based on the transformed training data set 1, is shown in Table 3.5. The misclassification matrix helps identify which species can be expected to predict accurately. The misclassification rate matrix converts the absolute numbers in the misclassification matrix to percentages. The absolute numbers in Table 3.5 are converted to percentages as shown in Table 3.6. Refer to Table 1.1 for specie to specie number correspondence.

As Table 3.6 shows, only urine (i.e., specie 1) meets the objective of 95% for classification accuracy for this particular training data set.

Table 3.5: KNN Misclassification Matrix for Transformed Training Data Set 1

Actual Specie	Predicted Specie						
	1	2	3	4	5	6	7
1	**26**	0	0	0	1	0	0
2	0	**19**	2	2	0	3	3
3	0	5	**16**	4	1	0	4
4	0	4	2	**16**	0	3	5
5	0	1	5	0	**14**	4	0
6	0	5	0	2	0	**19**	3
7	0	6	5	6	0	1	**11**

Table 3.6: KNN Misclassification Rate Matrix for Transformed Training Data Set 1

Actual Specie	Predicted Specie						
	1	2	3	4	5	6	7
1	**96.3**	0	0	0	3.7	0	0
2	0	**65.5**	6.9	6.9	0	10.3	10.3
3	0	16.7	**53.3**	13.3	3.3	0	13.3
4	0	13.3	6.7	**53.3**	0	10.0	16.7
5	0	4.2	20.8	0	**58.3**	16.7	0
6	0	17.2	0	6.9	0	**65.5**	10.3
7	0	20.7	17.2	20.7	0	34.5	**37.9**

3.2.5 KNN Model

To be able to classify unknown specimens, the critical parameters of the model, determined using the training data are saved.

The model contains two components:

a. Training Data (transformation parameters also).

b. Optimal *K*.

Both of these components are saved to create the KNN model, which will be used to predict the class of unknown specimens. If the training data contain many specimens then the model will be relatively large.

3.3 Prediction using KNN Model

Classifying unknown specimens based on a KNN model involves 3 steps: Calculate Interspecimen Distances, Derive Voting Matrix, and Classify Unknown Specimens according to Voting Matrix and Optimal K. A fourth step is added, only for validation, using "held-out" data, where the "held-out" specimens are known. These steps are shown in Figure 3.5.

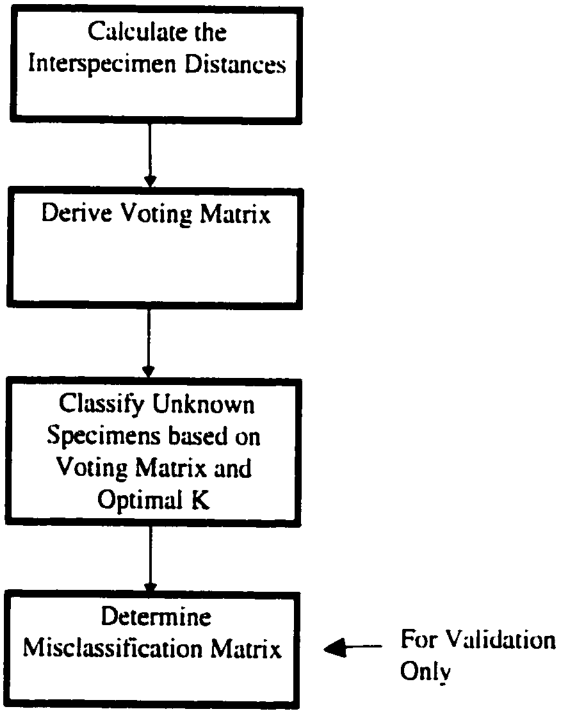

Calculate the Interspecimen Distances
Derive Voting Matrix
Classify Unknown Specimens based on Voting Matrix and Optimal K
Determine Misclassification Matrix
For Validation Only

Figure 3.5: Prediction of Unknown Specimens based on KNN Model

3.3.1 Interspecimen Distances

The interspecimen distances are calculated as described in Section 3.2.1 except that the distances are between each unknown specimens and the training specimens. For the simple Cartesian example, the distances between unknown G and the training data are shown in Table 3.7.

Table 3.7: Interspecimen Distances between Unknown G and Training Data

Training Specimen	A	B	C	D	E	F
Unknown G	0.291	0.569	0.335	0.924	1.207	0.987

3.3.2 Voting Matrix

The derivation of the voting matrix for prediction is similar to that described for calibration in Section 3.2.2, except that all the training data are used to vote for the specie for each unknown specimen, in the case of prediction. For example, the voting matrix (a vector in this case since there is only one unknown data point) is derived for unknown data point G, as shown in Table 3.8. To save computational time, only the matrix elements up to the optimal *K* need to be calculated. For illustration purposes, all matrix elements are shown in Table 3.8.

Table 3.8: Voting Matrix for Unknown Specimen G

	*K*1	*K*2	*K*3	*K*4	*K*5	*K*6
G	A (ONE)	C (ONE)	B (ONE)	F (TWO)	E (TWO)	E (TWO)

3.3.3 Classify Unknown Specimens

This process is similar to that described for calibration in Sections 3.2.2 and 3.2.3, except that the optimal *K* is known. Optimal *K* for the Cartesian example is 1 or 2. Assuming *K*=2, then specie ONE would have two votes while specie TWO would have zero votes for unknown specimen G. Therefore, G is classified as specie ONE.

3.3.4 Misclassification Matrix

The misclassification matrix is determined in similar fashion for the prediction data as for the training data as described in Section 3.2.4. Tables 3.9 and 3.10 show the misclassification matrix and the misclassification rate matrix respectively for prediction data set 1, which is transformed similarly to the training data set 1, as outlined in Section 3.2.4 .

Table 3.9: KNN Misclassification Matrix for Transformed Prediction Data Set 1

Actual Specie	Predicted Specie						
	1	2	3	4	5	6	7
1	**19**	0	0	1	0	0	0
2	0	**11**	2	1	0	2	4
3	0	3	**12**	2	0	2	1
4	0	8	1	**7**	0	1	3
5	0	0	4	2	**10**	1	3
6	0	3	0	1	0	**15**	1
7	0	3	5	0	1	0	**11**

Table 3.10: KNN Misclassification Rate Matrix for Transformed Prediction Data Set 1

Actual Specie	Predicted Specie						
	1	2	3	4	5	6	7
1	**95.0**	0	0	5.0	0	0	0
2	0	**55.0**	10.0	5.0	0	10.0	20.0
3	0	15.0	**60.0**	10.0	0	10.0	5.0
4	0	40.0	5.0	**35.0**	0	5.0	15.0
5	0	0	20.0	10.0	**50.0**	5.0	15.0
6	0	15.0	0	5.0	0	**75.0**	5.0
7	0	15.0	25.0	0	5.0	0	**55.0**

Table 3.11 shows the comparison between the training data and prediction data predictabilities (predictability for the training data is determined by the leave-one-out validation method) rates for data set 1.

Table 3.11: Predictability Rate Comparison between Training and Prediction Data

Specie	1	2	3	4	5	6	7
No. of Specimen Training Set 1	27	29	30	30	24	29	29
No. of Specimen Prediction Set 1	20	20	20	20	20	20	20
Predict. Rate Training Set 1 (%)	96.3	65.5	53.3	53.3	58.3	65.5	37.9
Predict. Rate Prediction Set 1 (%)	95.0	55.0	60.0	35.0	50.0	75.0	55.0

The error rates for training and prediction were 38.9% and 39.3% respectively. Only urine meets the classification accuracy objective of 95% for classifying unknown specimens. The results are presented formally in Chapter 6.

Chapter 4

Mahalanobis Distance Classification

4.1 Introduction

The Mahalanobis[11] Distance (MD) classification method is a more sophisticated method than KNN. The distance measure for MD is an extension of the distance measure for KNN. MD takes into account the distribution (i.e., variance) about a specie's mean location and inter-relationships (i.e., covariances) between the features. Due to the more computationally intensive nature of MD, principal component scores are used for the features instead of the wavelengths, in an effort to reduce the number of features and minimise the computation time. Note: the term factors will be used instead of features to refer to the PC's.

The Mahalanobis distance is defined by Equation 4.1:

$$D^2 = (\mathbf{x}-\mathbf{x}_i)' \mathbf{M}(\mathbf{x}-\mathbf{x}_i) \tag{4.1}$$

where $\mathbf{x}$ = point in the dimensional space for a particular specimen, $\mathbf{x}_i$ = centre of specie i , $\mathbf{M}$ = inverse covariance matrix for the dimensions, and D = Mahalanobis

[11] Named after P.C. Mahalanobis.

distance (i.e., MD) from specimen to specie i . Equation 4.1 is used as the basis for interspecie distances and for specie to specimen distances.

This method was implemented using the principal components of the data as the dimensional space. Wavelengths were not implemented as the direct input features to the MD analysis.

4.2 Calibration using Mahalanobis Distance

As in the case for KNN, calibration using the training data is necessary to create a MD model that can be used to classify unknown specimens. Calibrating, using training data, to create a Mahalanobis Distance model involves 7 steps, as shown in Figure 4.1: Calculate Specie Statistics on Transformed Data, Calculate Interspecie Mahalanobis Distances, Calculate Mahalanobis Distances between Species and Specimens, Classify Specimens by Minimum Mahalanobis Distance for each Specimen, Derive Misclassification Matrix, Determine Optimal Number of Factors, Save MD Model.

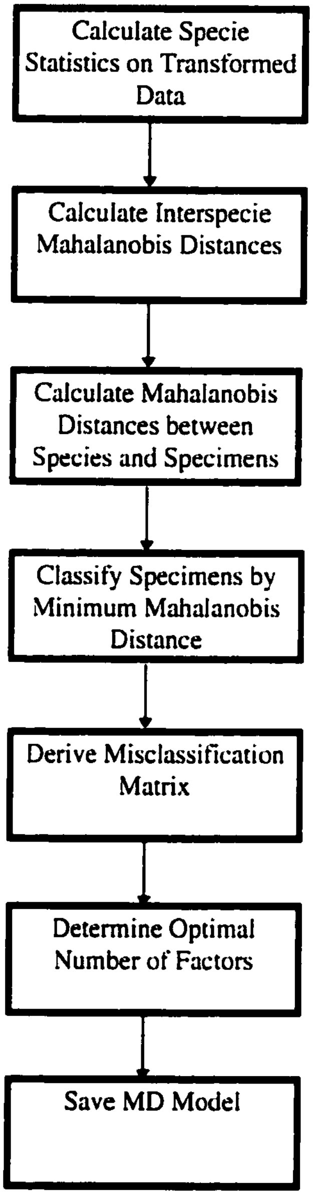

Calculate Specie Statistics on Transformed Data
Calculate Interspecie Mahalanobis Distances
Calculate Mahalanobis Distances between Species and Specimens
Classify Specimens by Minimum Mahalanobis Distance
Derive Misclassification Matrix
Determine Optimal Number of Factors
Save MD Model

Figure 4.1: Calibration using Mahalanobis Distance

4.2.1 Specie Statistics

Statistics--i.e., mean, variance, and covariances--are calculated for each specie based on the features, in this case the principal component scores. The results of these calculations are a mean vector and a covariance matrix. The specie statistics form the basis from which the MD (Equation 4.1) can be calculated.

The mean vector, $\bar{\mathbf{x}}_s$, and specimen variance-covariance matrix, $\mathbf{C}_s$, are calculated for each specie s. The elements of the mean vector are calculated according to Equation 4.2 as given by

$$\bar{x}_{sj} = \sum_{i=1}^{m_s} \frac{x_{sij}}{m_s} \tag{4.2}$$

where x_{sij} is the $(i, j)th$ element of the $m_s \times p$ matrix, $\mathbf{X}_s$, m_s is the number of specimens for specie s, and p is the number of principal components. The matrix $\mathbf{X}_s$ is a subset of the transformed data matrix, $\mathbf{X}$, where $\mathbf{X} = \mathbf{T}_p$, the principal component matrix as defined in Equation 2.4.

The elements, c_{sjk}, of the sample variance-covariance matrix, $\mathbf{C}_s$, are given by Equation 4.3:

$$c_{sjk} = \sum_{i=1}^{p} \left(x_{sij} - \bar{x}_{sj}\right)\left(x_{sik} - \bar{x}_{sk}\right) / (p-1) \tag{4.3}$$

where x_{sij} and x_{sik} are elements of the matrix $\mathbf{X}_s$, $\bar{x}_{sj}$ and $\bar{x}_{sk}$ are elements of the mean vector $\bar{\mathbf{x}}_s$, for specie s, and p is the number of principal components.

For $j = k$, Equation 4.3 reduces to:

$$c_{sjj} = s_{sj}^2 = \sum_{i=1}^{p} \left(x_{sij} - \overline{x_{sj}}\right)^2 / (p-1) \tag{4.4}$$

where s_{sj}^2 is the sample variance for variable (i.e., principal component) j and specie s.

The individual sample variance-covariance matrices are combined together, as suggested by Mark and Tunnel [33], to form the pooled covariance matrix. The advantage of this is that fewer specimens are required for calibration. It is expected that the differences between species with respect to the covariances will be subtle; therefore, pooling the covariances is a reasonable step to reduce computation time. The pooled covariance matrix, $\mathbf{C}$, is calculated as the mean of the individual covariance matrices as shown in Equation 4.5[12]:

$$c_{ij} = \frac{1}{S}\sum_{s=1}^{S} c_{sij} \tag{4.5}$$

where c_{ij} is an element of $\mathbf{C}$ and S is the number of species.

4.2.1.1 A Sample Calculation of the Specie Statistics

To illustrate the calculations of the specie statistics, the simple Cartesian example from Chapter 3 is used. The groupmeans (i.e., specie means) for the training data are calculated and shown in Table 4.1. The variance-covariance matrices are shown in Figure 4.2.

Table 4.1: Groupmeans for Simple Cartesian Example

	X co-ordinate	Y co-ordinate	Z co-ordinate
Specie "ONE"	0.23	2.00	0.49
Specie "TWO"	1.00	1.01	1.03

$$\mathbf{C}_{ONE} = \begin{bmatrix} 0.0121 & 0.0082 & -0.0083 \\ 0.0082 & 0.0240 & -0.0272 \\ -0.0083 & -0.0272 & 0.0308 \end{bmatrix}$$

$$\mathbf{C}_{TWO} = \begin{bmatrix} 0.0197 & -0.0098 & 0.0042 \\ -0.0098 & 0.0211 & 0.0131 \\ 0.0042 & 0.0131 & 0.0151 \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 0.0159 & -0.0008 & -0.0020 \\ -0.0008 & 0.0226 & -0.0070 \\ -0.0020 & -0.0070 & 0.0230 \end{bmatrix}$$

Figure 4.2: Variance-Covariance Matrices for Simple Cartesian Example

[12] A weighted mean could be used if the number of specimens were different on a specie-by-specie basis. Also, the variance-covariance matrix could be calculated directly using the whole data set instead of doing of calculating on a specie-by-specie basis and then pooling.

The Z-plane projections of the groupmeans and the three standard deviation borders (as calculated from the individual covariance matrices) are plotted in Figure 4.3. Good separability exists if there is no overlapping between species of their three standard deviation zones, about their respective groupmeans. Therefore, the simple example shows good separability in the Z-plane projection.

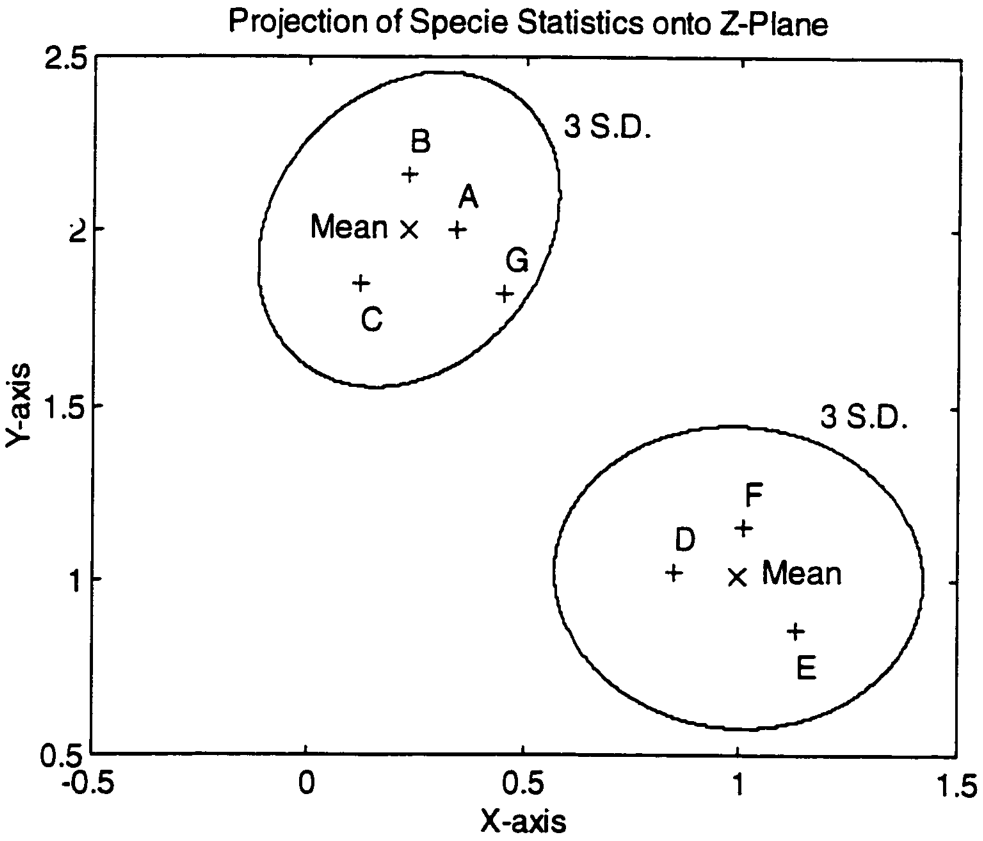


Figure 4.3: Projection of Specie Statistics onto Z-Plane for Simple Cartesian Example

4.2.2 Interspecie Mahalanobis Distances

The interspecie MD's (as derived from Equation 4.1) are calculated from the specie statistics. The interspecie MD's give an indication how separable the species are from each other. An interspecie MD greater than 6 between two species indicates that the two species are highly separable, while an interspecie MD less than 6 between two species indicates that the species are not highly separable, compared to classifying unknown specimens at a 95% accuracy rate.

Equation 4.1 is re-expressed to calculate the distance between the specie centres, i.e., D_{ij}^2 for species i and j, as shown in Equation 4. 6:

$$D_{ij}^{2} = \left(\mathbf{x}_i - \mathbf{x}_j\right)' \mathbf{M}\left(\mathbf{x}_i - \mathbf{x}_j\right) \tag{4.6}$$

where $\mathbf{M} = \mathbf{C}^{-1}$, the inverse sample variance-covariance matrix.

4.2.2.1 A Sample Calculation of the Interspecie MD

Using the simple Cartesian example, the Mahalanobis Distance, $D_{ONE.TWO}$, is calculated using the species statistics calculated in Section 4.2.1.1 and using Equation 4.6. The calculation is shown below:

$$D_{ONE.TWO}^{2} = \left[\begin{bmatrix}0.23 & 2.00 & 0.49\end{bmatrix} - \begin{bmatrix}1.00 & 1.01 & 1.03\end{bmatrix}\right] \begin{bmatrix} 0.0159 & -0.0008 & -0.0020 \\ -0.0008 & 0.0226 & -0.0070 \\ -0.0020 & -.0070 & 0.0230 \end{bmatrix}^{-1} \left[\begin{bmatrix}0.23\\ 2.00\\ 0.49\end{bmatrix} - \begin{bmatrix}1.00\\ 1.01\\ 1.03\end{bmatrix}\right].$$

Performing the matrix operations, $D_{ONE.TWO}^{2}$= 82.87 , or $D_{ONE.TWO}$ = 9.10. Since the Mahalanobis distance between the means of the two species is greater than six standard deviations, good separability exists between them.

The Mahalanobis distances between all species of the biological data (i.e., urine, serum, etc.), with data transformation SM11-2D11, autoscaling, 255 wavelengths, and 19 PC's; are shown in Table 4.2. For example, the element in Row 2 and Column 3 is the MD between species 2 and 3.

From Table 4.2, it is observed that specie 1, or urine, is highly separable from all the other species with a minimum interspecie MD of 10.48. This is significantly higher than the 6 standard deviations required for good separability. Species 5 (oxalate) and 6 (iodoacetate) are reasonably separable with minimum interspecie MD's of 4.57 and 4.19 respectively. Species 2, 3, 4, and 7 (i.e., serum, citrate, EDTA, and heparin) are not very separable from each other with minimum interspecie MD's of 1.64, 1.96, 1.67 and 1.64 respectively.

Table 4.2: Mahalanobis Distances between Species

	1	2	3	4	5	6	7
1	0	10.48	10.71	10.67	11.35	12.32	11.22
2		0	2.60	1.67	4.78	4.19	1.64
3			0	2.67	4.57	4.94	1.96
4				0	5.11	4.66	2.02
5					0	5.37	4.67
6						0	4.44
7							0

4.2.3 Mahalanobis Distances between Species and Specimens

The MD between a specimen and a particular specie is used to determine if the specimen belongs to the particular specie. The minimum MD between a specimen and the species is used to classify the specimen as specie *S*, where the MD between specie *S* and the particular specimen is a minimum.

To calculate the MD between a specimen and a specie, Equation 4.1 is applied.

$$D^2 = (\mathbf{x} - \mathbf{x}_i)' \mathbf{M} (\mathbf{x} - \mathbf{x}_i) \tag{4.1}$$

where $\mathbf{x}$ = point in the dimensional space for a particular specimen, $\mathbf{x}_i$ = centre of specie i , $\mathbf{M}$ = inverse covariance matrix for the dimensions, and D = Mahalanobis distance (i.e., MD) from specimen to specie i .

4.2.3.1 A Sample Calculation of the MD between a Specimen and a Specie

Using the simple Cartesian example, the Mahalanobis distance, $D_{A.ONE}$, between data point A and specie ONE is calculated using the vectors and matrices as shown below:

$$D^2_{A.ONE} = \left[\begin{bmatrix}0.34 & 2.00 & 0.51\end{bmatrix} - \begin{bmatrix}0.23 & 2.00 & 0.49\end{bmatrix}\right] \begin{bmatrix} 0.0159 & -0.0008 & -0.0020 \\ -0.0008 & 0.0226 & -0.0070 \\ -0.0020 & -.0070 & 0.0230 \end{bmatrix}^{-1} \left[\begin{bmatrix}0.34 \\ 2.00 \\ 0.49\end{bmatrix} - \begin{bmatrix}0.23 \\ 2.00 \\ 0.49\end{bmatrix}\right].$$

After performing the matrix operations: $D^2_{A.ONE} = 0.8083$, or $D_{A.ONE} = 0.8991$. Table 4.3 shows the specimen to specie MD's for the training data.

Table 4.3: Specimen to Specie Mahalanobis Distances

Training Data Point	Specie ONE	Specie TWO
A	0.8891	8.4994
B	1.4014	10.1227
C	1.5463	8.8921
D	8.1808	1.3855
E	10.3771	1.5271
F	8.9912	1.6078

4.2.4 Classification of Specimen by Minimum Mahalanobis Distance

The minimum specie to specimen Mahalanobis distance is used to classify a specimen. For example, the minimum MD for specimen A is determined by $\min\{D_{A,s}\}$ for all $s = 1$ to S, where S is the number of species. For the simple Cartesian example, $D_{A.ONE} < D_{A.TWO}$ (i.e., 0.8891 < 8.4994 from table 4.3); therefore specimen A is classified as specie ONE. From table 4.3, specimens A, B, C are classified as ONE's while specimens D, E, F are classified as TWO's. These classifications are correct.

4.2.5 Misclassification Matrix

The misclassification matrix breaks down the errors by specie, as described in Section 3.2.4. The misclassification matrix is used to identify which species are expected to have high classification rates and which will not. The misclassification rate matrix converts the absolute errors to percentages. For example, the misclassification matrix for the transformed training data set 1 is shown in Table 4.4.

Table 4.4: MD Misclassification Matrix for Transformed Training Data Set 1

Actual Specie	Predicted Specie						
	1	2	3	4	5	6	7
1	**27**	0	0	0	0	0	0
2	0	**21**	0	5	0	0	3
3	0	1	**20**	1	1	0	7
4	0	5	0	**20**	0	0	5
5	0	0	2	0	**22**	0	0
6	0	2	0	0	0	**26**	1
7	0	8	4	2	0	0	**15**

Table 4.5: MD Misclassification Rate Matrix for Transformed Training Data Set 1

Actual Specie	Predicted Specie						
	1	2	3	4	5	6	7
1	**100**	0	0	0	0	0	0
2	0	**72.4**	0	17.2	0	0	10.3
3	0	3.3	**66.7**	3.3	3.3	0	23.3
4	0	16.7	0	**66.7**	0	0	16.7
5	0	0	8.3	0	**91.7**	0	0
6	0	6.9	0	0	0	**89.7**	3.4
7	0	27.6	13.8	6.9	0	0	**51.7**

4.2.6 Optimal Number of Factors

As in the case for KNN, the optimal number of factors (i.e., principal components for MD) are determined such that the calibration error rate (i.e., error rate of training data set) is minimised. The determination of the optimal number of principal components is accomplished according to the following algorithm:

```
FOR $PC$ = 1 TO $PC_{max}$ STEP =1
    1. Calculate Misclassification Matrix
    2. Calculate Aggregate Error
END FOR
FOR $PC$ = 1 TO $PC_{max}$ - 1 STEP= 1
    3. Calculate Test Statistic
END FOR
```

where PC is the number of principal components and PC_{max} is an arbitrarily high number of principal components (e.g., 40). The calculation of the misclassification matrix is described in Section 4.2.5. The aggregate error is simply the summation of the errors in the misclassification matrix. For example, the aggregate error for the data in Table 4.4 would be 48 (i.e., 1+5+3+1+1+1+7+5+5+2+2+1+8+4+2). The aggregate error rates for these data are plotted as shown in Figure 4.4.

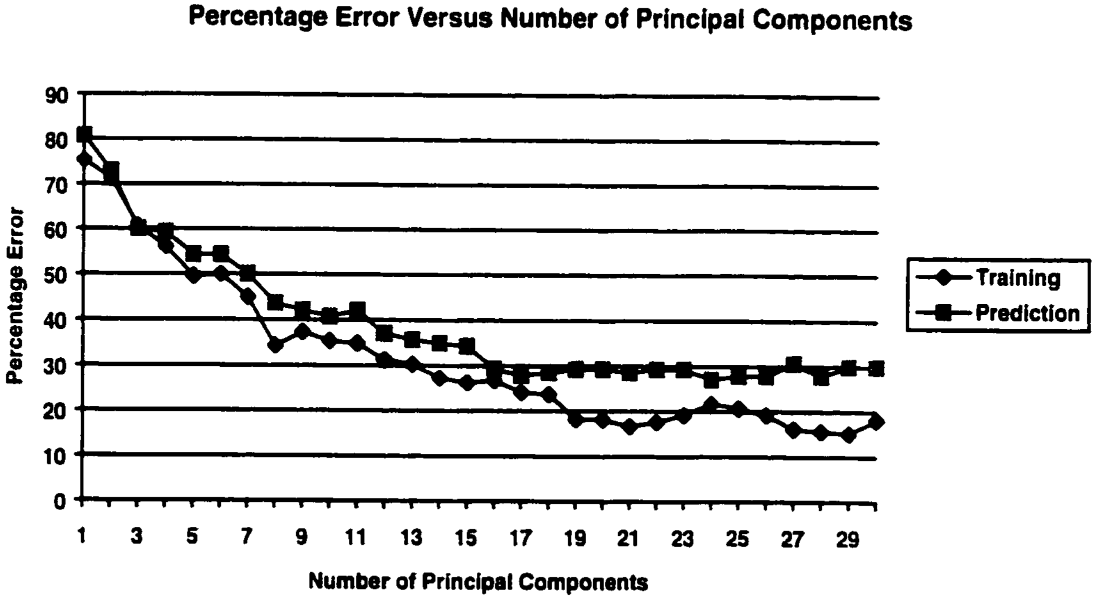


Figure 4.4: Aggregate Error Rates Versus Number of Principal Components

4.2.6.1 Test Statistic

To determine the optimal number of principal components, each principal component is tested to determine if its addition statistically reduces the error rate. A hypothesis test is set-up to determine this. In quantitative problems, the statistic used is the F-statistic. It compares the variance (i.e., the error rate) between two cases. In classification problems the F-statistic cannot be used since the error rate determined is not a measure of its variance, but rather an estimate of its mean. This is due to the fact that errors in a classification problem are positive only. Therefore, a hypothesis test concerning two means is used. In this situation, the Z-statistic is chosen over the Student t-statistic since a relatively large number of specimens (i.e., 198 in training data) are available to estimate the mean and variance. The Z-statistic is given by:

$$z = \frac{\left(\overline{x_{PC}} - \overline{x_{PC+1}}\right) - \delta}{\sigma^2_{\overline{x_{PC}} - \overline{x_{PC+1}}}} \qquad (4.7)$$

where z is the Z-statistic, $\overline{x_{PC}}$ and $\overline{x_{PC+1}}$ are the estimates of the mean aggregate errors for the number of principal components PC and $PC+1$ respectively, $\sigma_{\overline{x_{PC}} - \overline{x_{PC+1}}}$ is the standard deviation of the sampling distribution of the difference between the sample means, and δ is a specified constant (of difference between the means). The variance, $\sigma^2_{\overline{x_{PC}} - \overline{x_{PC+1}}}$, is estimated according to Equation 4.8 [34, p. 216]:

$$\sigma^2_{\overline{x_{PC}} - \overline{x_{PC+1}}} = \frac{\sigma^2_{PC}}{m_{PC}} + \frac{\sigma^2_{PC+1}}{m_{PC+1}}. \qquad (4.8)$$

Combining Equations 4.7 and 4.8 yields

$$z = \frac{\left(\overline{x_{PC}} - \overline{x_{PC+1}}\right) - \delta}{\sqrt{\frac{\sigma^2_{PC}}{m_{PC}} + \frac{\sigma^2_{PC+1}}{m_{PC+1}}}}. \qquad (4.9)$$

To determine if $\mu_{PC} > \mu_{PC+1}$ (i.e., if error rate is decreasing), δ is set to 0; the null hypothesis, $\mu_{PC} - \mu_{PC+1} = 0$, is tested against the alternate hypothesis, $\mu_{PC} - \mu_{PC+1} > 0$. The null hypothesis is rejected if $z > z_\alpha$, where α is the significance level.

To estimate the means and variances, the curve shown in Figure 4.4 is smoothed using a five point moving average. The smoothed curve is used to estimate the means. The smoothed curve is differenced with the original curve to determine residuals. Standard deviations are estimated using five points (i.e., two points on either side of the centre point). The estimated standard deviation curve is then smoothed using a five point moving average.

Figure 4.5 shows the calculated Z-statistic for the data shown in Figure 4.4. Figure 4.6 is the same as Figure 4.5 except the vertical axis is expanded for Figure 4.6, to show the crossings more clearly. Table 4.6 shows the number of principal components determined for several values of α.

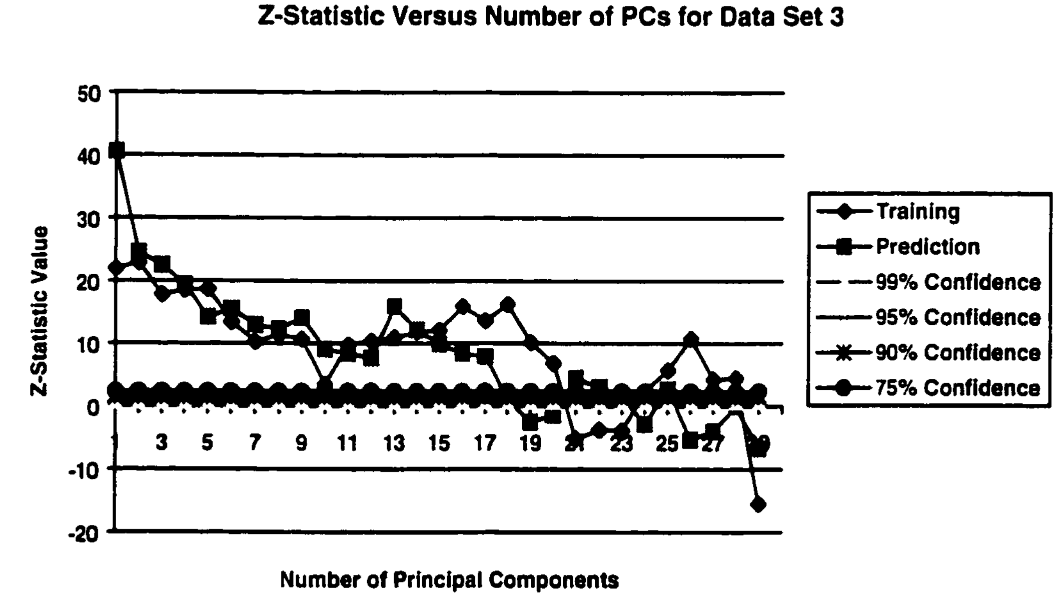


Figure 4.5: Z-Statistic Versus Number of Principal Components

Table 4.6: Optimal Number of Principal Components

α	z_α	*PC*
0.25	0.6745	23
0.10	1.2816	16
0.05	1.6449	16
0.01	2.3263	16

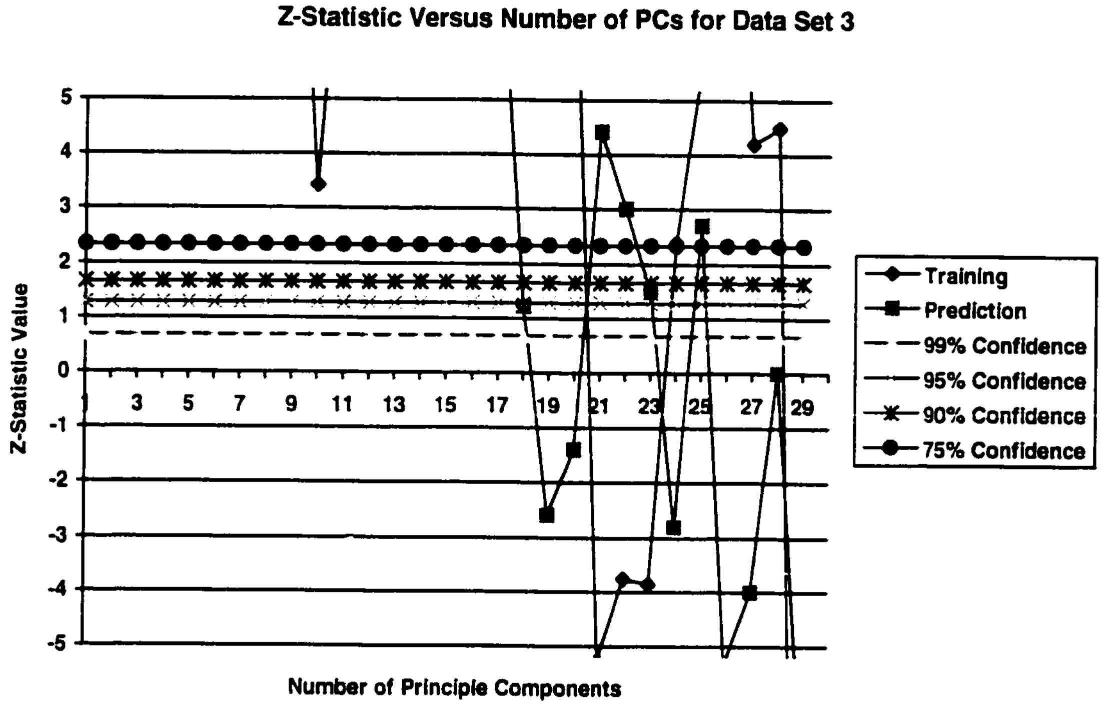


Figure 4.6: Z-Statistic Versus Number of Principal Components (Expanded Vertical Scale)

An α of 0.25 was chosen considering that the Z-statistic is a relatively gross estimate.

4.2.7 Save Mahalanobis Distance Model

To be able to classify unknown specimens, the necessary parameters, i.e., specie statistics and optimal number of factors (# of PC's), are stored.

The parameters that form the MD Model are:

- the optimal number of principal components, PC.
- the species' mean vectors, $\bar{\mathbf{x}}_s$, for $s = 1,2,\ldots,S$, where $\bar{\mathbf{x}}_s$ has PC number of elements.
- the pooled variance-covariance matrix, $\mathbf{C}$, a $PC \times PC$ matrix.

4.3 Prediction using Mahalanobis Distance Model

Predicting or classifying unknown specimens involves two steps: Calculate Mahalanobis Distances from the Specimen to Each Specie and Classify the Specimen based on the Minimum Mahalanobis Distance. A third step, Derive Misclassification Matrix, is used to validate the method using held-out specimens. These steps are shown in Figure 4.7.

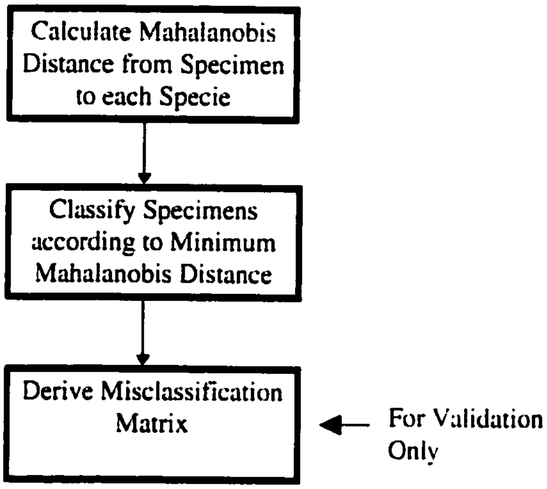

Calculate Mahalanobis Distance from Specimen to each Specie
Classify Specimens according to Minimum Mahalanobis Distance
Derive Misclassification Matrix
For Validation Only

Figure 4.7: Prediction using Mahalanobis Distance Model

4.3.1 Mahalanobis Distance between Species and Specimen

The MD between a specimen and a particular specie is used to determine if the specimen belongs to that specie. The minimum MD between a specimen and the species is used to classify the specimen as Specie *S*, where MD between Specie *S* and the particular specimen is a minimum. This calculation is the same as described in Section 4.2.3, using Equation 4.1, except that prediction specimens are used instead of training specimens. Table 4.7 shows the MD's for the "unknown" data point G of the simple Cartesian example.

Table 4.7: Specimen to Specie Mahalanobis Distances

	Specie ONE	Specie TWO
G	2.4757	6.8553

4.3.2 Classification of Specimen by Minimum Mahalanobis Distance

The minimum specie to specimen Mahalanobis distance is used to classify a specimen. This step is the same as described in Section 4.2.4, except that prediction specimens are used instead of training specimens. From Table 4.7, data point G is classified as specie ONE since the MD between G and specie ONE is the smallest.

4.3.3 Misclassification Matrix

The misclassification matrix breaks down the errors by specie, as described in Section 3.2.4. The misclassification matrix is determined as described in Section 4.2.5, except that prediction specimens are used instead of training specimens. In actual practice, the misclassification matrix for prediction cannot be determined since the actual classification of a prediction specimen is unknown. Therefore, this step will only be necessary when known specimens are held-out (i.e., left-out) from the training data to validate the method. Tables 4.8 and 4.9 show the misclassification and misclassification rate matrices for prediction data set 1.

Table 4.8: MD Misclassification Matrix for Transformed Prediction Data Set 1

Actual Specie	Predicted Specie						
	1	2	3	4	5	6	7
1	**19**	1	0	0	0	0	0
2	0	**12**	1	1	0	1	5
3	0	4	**12**	0	0	1	3
4	0	12	0	**6**	0	0	2
5	0	0	0	0	**20**	0	0
6	0	3	0	0	0	**16**	1
7	0	5	2	3	0	0	**10**

Table 4.9: MD Misclassification Rate Matrix for Transformed Prediction Data Set 1

Actual Specie	Predicted Specie						
	1	2	3	4	5	6	7
1	**95**	5	0	0	0	0	0
2	0	**60.0**	5.0	5.0	0	5.0	25.0
3	0	20.0	**60.0**	0	0	5.0	15.0
4	0	60.0	0	**30.0**	0	0	10.0
5	0	0	0	0	**100.**	0	0
6	0	15.0	0	0	0	**80.0**	5.0
7	0	25.0	10.0	15.0	0	0	**50.0**

Table 4.10 shows the comparison between the training and prediction predictability rates for data set 1.

Table 4.10: Predictability Rate Comparison between Training and Prediction

Specie	1	2	3	4	5	6	7
No. of Specimen Training Set 1	27	29	30	30	24	29	29
No. of Specimen Prediction Set 1	20	20	20	20	20	20	20
Predict Rate Training Set 1 (%)	100.0	72.4	66.7	66.7	91.7	89.7	51.7
Predict Rate Prediction Set 1 (%)	95.0	60.0	60.0	30.0	100.0	80.0	50.0

The error rates for training and prediction were 23.7% and 32.1% respectively. The results are presented formally in Chapter 6.

Chapter 5

Genetic Algorithm Selection of Wavelengths

5.1 Introduction

A Genetic Algorithm (GA) is used to search the solution space for an optimum solution. In this application, a GA is used to optimise the MD classification by identifying the features or wavelengths that are most important for an optimum MD model. A GA searches the solution space iteratively via a combination of exploration (i.e., random searching) and incremental improvement. The GA terminates when a predefined criterion is achieved; the termination criterion may be a given number of iterations or generations, or the termination criterion may be exceeding a specified threshold for the fitness.

To date, investigations involving feature selection in chemometrics, using a genetic algorithm, have almost exclusively focused on quantitative problems, i.e., those which attempt to minimise the error due to fitting to a regression equation (e.g., Multiple Linear Regression, Partial Least Squares, or Principal Component Regression); e.g., [25], [26], [35] and [36]. Little work has been done in applying GA feature selection to classification problems; in fact only a few papers even address the quantitative chemometric situation.

The approach taken here is to minimise the error rate of a Mahalanobis distance calibration, using the principal components. The principal components transformation is performed for each member during each generation. The principal components transformation is performed using the selected wavelengths only, as defined by the particular member. The other data transformations (i.e., SM11-2D11, and autoscaling) occur only once, at the beginning.

Approximately thirty GA analyses were run to determine reasonable values for the population size (i.e., 100), the mutation rate (i.e., 0.05) and the replacement rate (i.e., 0.20). Results are not particularly sensitive to these parameters but the number of generations, and thus the computation time, is sensitive to these GA parameters. Another seventy GA analyses were run to provide the results as presented in Chapter 6. Each GA analysis took twelve to forty-eight hours to run on a 486 DX2 personal computer operating at 66 MHz.

The next section will outline in detail the GA algorithm used in this investigation.

5.2 Selection of Wavelengths and Calibration

To understand how a GA works, it is important to first understand the concept of a population. A population consists of a set of members whose elements determine whether a given feature is selected or not. Table 5.1 shows an example of a population. This population shown in Table 5.1 has five members and five features. In our application, the features are wavelengths. The elements of the population matrix are select status bits. A select status value of "1" means that feature is selected for the particular member. A select status value of "0" means that feature is deselected, or not selected, for that member. Selected features are used in calculations involving the member while deselected features are not used.

A complete MD calibration is performed for each member of a population, using the training data. Based on the calibration error rate and the number of features used, a fitness value is determined for each member, as well as a mean fitness value for the population. A new population is then created for the next generation based on the current members' fitness values. Some members are eliminated, while others are cloned. Parenting and mutating of these members adds diversity to the new population. The steps are formalised as shown in Figure 5.1. The steps include: Generate an Initial Population, Evaluate the Population with respect to a Fitness Criterion (Terminate if Conditions Met), Replace the Worst Members with the Best Members, Recombine Members to Parent a new Members, Mutate Members, Create Wavelength Selection Vector, and Calibrate. These steps are described in detail in Sections 5.2.1 to 5.2.8.

Table 5.1: Example of a Population

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
Member 1	1	0	1	0	1
Member 2	0	1	1	0	0
Member 3	1	0	0	1	1
Member 4	0	0	1	1	0
Member 5	1	1	0	0	1

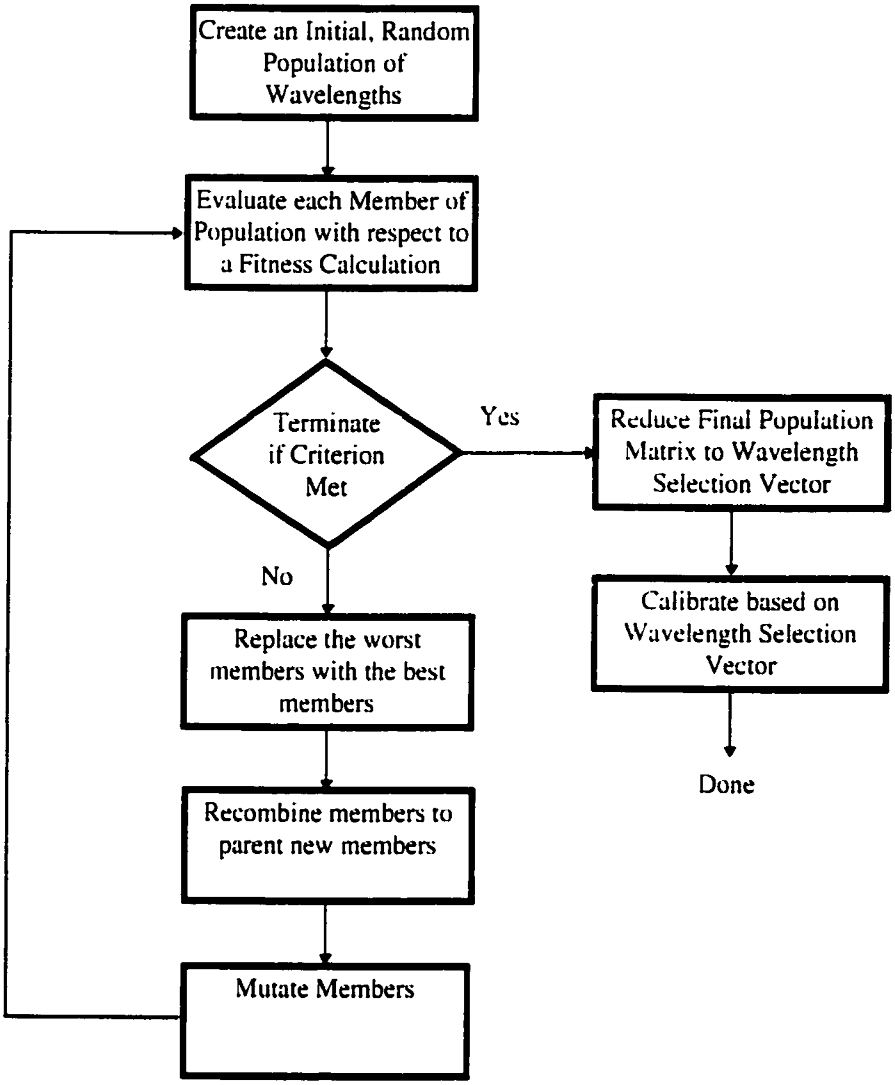

Create an Initial, Random Population of Wavelengths
Evaluate each Member of Population with respect to a Fitness Calculation
Terminate if Criterion Met
Yes
Reduce Final Population Matrix to Wavelength Selection Vector
Calibrate based on Wavelength Selection Vector
Done
No
Replace the worst members with the best members
Recombine members to parent new members
Mutate Members

Figure 5.1: Genetic Algorithm to Select Wavelengths and Calibrate

5.2.1 Initial Population

To start off the GA process, an initial population matrix is defined, whose elements are chosen randomly. This is done only once, at the beginning of the algorithm. The initial population, obviously, is not based on known good features, thus the early results (i.e., mean error rates of the early generations) are typically worse than if all the features are used. Each member is used to define a wavelength selection vector, which is used to "filter" the spectral data, thereby defining new spectral data composed of the selected features or wavelengths only. A calibration, along with calibration error, is determined for each member.

Each member of the population consists of a 255 element vector of binary values. A "0" deselects the corresponding wavelength, while a "1" selects the corresponding wavelength. The population size (i.e., number of members) was chosen to be 100. An element of the population matrix, $\mathbf{W}$, is expressed as w_{ij}, for member i and wavelength j. A member is expressed by the vector, $\mathbf{w}_i$. Each element has a 50% chance of being selected or deselected.

For example, a four member, initial population for the simple Cartesian example is defined below:

$$\mathbf{W} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Therefore, member 3 would be expressed as $\mathbf{w}_3 = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}$.

5.2.2 Evaluation

An evaluation is performed for each member, based on the member's selected features. Each evaluation results in a calibration of the training data. The calibration error and the number of features are used to calculate the fitness value for each member.

The evaluation process is shown in more detail below:

```
FOR I = 1 TO Number of Members in Population (i.e., 100)

        1.    Create New Transformed Absorbance Matrix using only Selected
              Wavelengths.

        2.    Determine Principal Components of New Transformed
              Absorbance Matrix.

        3.    Perform Mahalanobis Distance Calibration
        4.      Calculate Fitness Value.
    END FOR
```

5.2.2.1 Absorbance Matrix with only Selected Wavelengths

The first step in the evaluation process is to create new spectral or transformed absorbance data from the training data, based on the members' selected wavelengths. The derivativation, smoothing, and autoscaling transforms have been performed prior, using all of the wavelengths. Only the PCA transformation is performed with the reduced feature set. The new spectral data are used to determine calibration and to evaluate fitness.

The transformed absorbance matrix $\mathbf{A}$, an $m \times n$ matrix, is compressed to form $\mathbf{B}$, an $m \times L$ matrix, where $n > L$, such that $\mathbf{B}$ is the $\mathbf{A}$ matrix with the deselected columns (i.e., wavelengths) removed, as defined by the selection vector, $\mathbf{w}_k$, for population member k.

The algorithm for compressing $\mathbf{A}$ to form $\mathbf{B}$ is shown below:

```
$l = 1;$
FOR j = 1 to N
    IF $w_{kj} \neq 0$
        FOR I = 1 to M
            $b_{il} = w_{kj} \times a_{ij};$
        END FOR
    $l = l + 1;$
    END IF
END FOR
```

where b_{il} is the (i,l)th element of matrix $\mathbf{B}$, w_{kj} is the (k,j)th element of the matrix $\mathbf{W}$ and a_{ij} is the (i,j)th element of matrix $\mathbf{A}$.

For example, the Cartesian matrix shown in Table 3.1 would be compressed by member

$\mathbf{w}_3$, as defined in Section 5.2.1, as shown below:

$$\begin{bmatrix} 0.34 & 2.00 \\ 0.23 & 2.16 \\ 0.12 & 1.85 \\ 0.85 & 1.02 \\ 1.13 & 0.86 \\ 1.01 & 1.15 \\ 0.45 & 1.82 \end{bmatrix}.$$

5.2.2.2 Principal Components Analysis

As previously mentioned, the PCA is performed on the new spectral data with the reduced or selected feature set, for a particular member. The principal component scores that are produced, are used in the MD calibration.

The principal component matrix, $\mathbf{T}$, of the compressed and transformed absorbance matrix, $\mathbf{B}$, is determined such that:

$$\mathbf{B} = \mathbf{TL}'$$

as defined by Equation 2.4 in Section 2.3.3. The truncated principal component matrix, $\mathbf{T}_p$, where p is the number of principal components, is determined according to Equations 2.5 and 2.6. The optimal number of PC's determined in the Mahalanobis distance calibration, for a given data set, is used for the GA optimisation process for the same data set.

5.2.2.3 Mahalanobis Distance Calibration

The PC scores produced from the PCA analysis, on the reduced feature set for a particular member, are used as the data input for the MD calibration. As outlined in detail in Section 4.2, the MD calibration requires 7 steps: Calculate Specie Statistics on Transformed Spectra, Calculate Interspecie MD's, Calculate MD's between Species and Specimens, Classify Specimen by Minimum MD, Derive Misclassification Matrix, Determine Optimal Number of Factors (i.e., PC's) and Save MD Model. The calculations are identical to those outlined in Section 4.2, even though there are now fewer wavelengths, since the MD input feature spaces consist of PC scores, in both cases.

5.2.2.4 Fitness Criteria

The fitness value for each member is calculated based on a fitness criterion calculation. The fitness criterion is then used to "order" the members, as explained in Section 5.2.4. The fitness criterion includes a factor related to the error rate and a factor related to the number of wavelengths. The correct "balance" between the "error factor" and the "wavelength factor" provides the most parsimonious solution. A number of criteria were tested, each with a different balance between the two factors. The criteria evaluated, for classifying all species simultaneously, are shown below:

FTEALLN1: $$\text{Fit} = 1/\left(\left(\sum_{s=1}^{S} Error_s\right)(N)\right)$$

FTEALLN2: $$\text{Fit} = 1/\left(\left(\sum_{s=1}^{S} Error_s\right)\left(\sqrt[2]{N}\right)\right)$$

FTEALLN3: $$\text{Fit} = 1/\left(\left(\sum_{s=1}^{S} Error_s\right)\left(\sqrt[3]{N}\right)\right)$$

FTEALLN4: $$\text{Fit} = 1/\left(\left(\sum_{s=1}^{S} Error_s\right)\left(\sqrt[4]{N}\right)\right)$$

where $Error_s$ is the number of errors associated with predicting actual specimens of specie s, and N is the number of wavelengths selected. The errors are summed over the number of species. These criteria are maximised in the GA process.

A number of criteria were tested that classified only one specie (i.e., urine), from the rest (i.e., blood). This could be used for a hierarchical classification as the top-level classification. That is, a model for classifying the various blood species could be determined, once it was known that a particular specimen was categorised as blood.

The "urine versus blood" fitness criteria equations are:

FTE1N8: $\text{Fit} = 1/\left(\left(Error_U + Error_B + 1\right)\left(\sqrt[8]{N}\right)\right)$

FTE1N4: $\text{Fit} = 1/\left(\left(Error_U + Error_B + 1\right)\left(\sqrt[4]{N}\right)\right)$

FTE1N2: $\text{Fit} = 1/\left(\left(Error_U + Error_B + 1\right)\left(\sqrt[2]{N}\right)\right)$

FTE1N1: $\text{Fit} = 1/\left(\left(Error_U + Error_B + 1\right)(N)\right)$

FTE1NSQR: $\text{Fit} = 1/\left(\left(Error_U + Error_B + 1\right)\left(N^2\right)\right)$

FTE1NSR4: $\text{Fit} = 1/\left(\left(Error_U + Error_B + 1\right)\left(N^4\right)\right)$

where $Error_U$ is the number of errors associated with predicting actual urine specimens (i.e., specie 1) and $Error_B$ is the number of errors associated with predicting actual blood specimens (i.e., species 2 to 7). In this case, unity is added to $Error_U + Error_B$, since $Error_U + Error_B$ can become zero, so that the denominator of the fit value does not go to zero. This was not required for the summation of the errors of all the species, since this did not go to zero, in any situation.

The fitness vector, $\mathbf{f}$, comprised of K elements, contains the individual fitness values, f_k, where K is the number of members in the population. For the example population described in Section 5.2.1, using the fitness criterion, FTEALLN1 (modified by adding 1 to the error summation as in FTE1N1 to prevent the denominator going to zero), yields the fit values, as shown in Table 5.2.

Table 5.2: Fit values for Example Population for Simple Cartesian Example

Member, k	$\sum_{s=1}^{2} Error_s + 1$	N	Fit, f_k
1	0+1=1	2	0.5
2	0+1=1	1	1.0
3	0+1=1	2	0.5
4	0+1=1	1	1.0

5.2.3 Termination Criterion

The termination criterion is used to stop the GA iteration process, and thus finalise the wavelength set and the calibration model. The termination criterion can be defined in several different ways including:

- the fitness value exceeds a predefined threshold
- the fitness value change from generation to generation is less than a predefined threshold
- the population reaches a certain level of homogeneity
- a predefined number of generations is exceeded.

The termination criterion used here, is that when a given number of generations is exceeded, the GA process is instructed to stop. This appears to be a reasonable criterion based on the data, as the following discussion indicates.

The minimum number of generations produced for most of the analyses was 500. This number was determined empirically by running a number of analyses and studying the fitness trends, the populations' homogeneity, and the error trends. For example, the fitness trend (for maximum, mean, and minimum values) for data set 1 is shown in Figure 5.2.

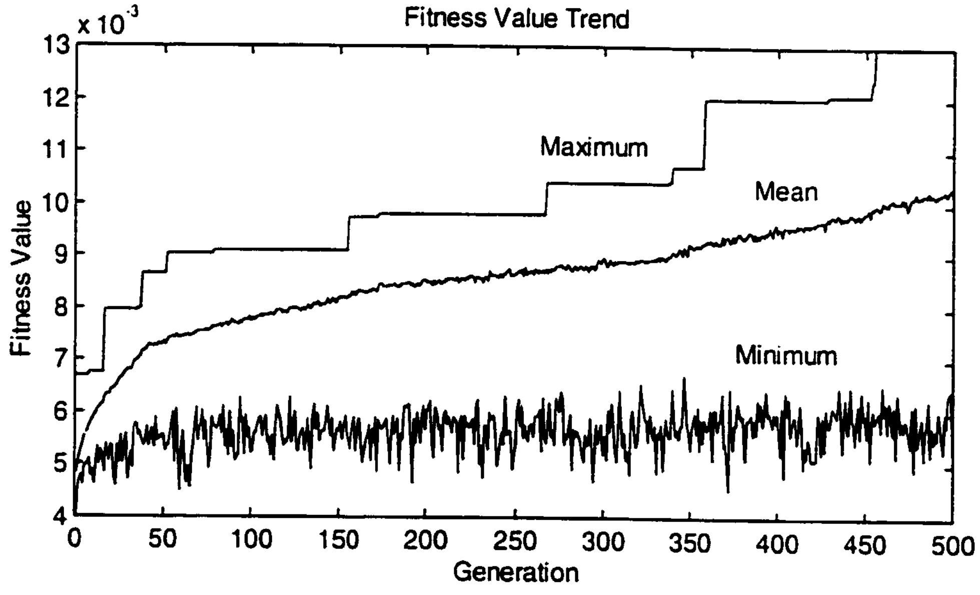


Figure 5.2: Fitness Trend for Training Data Set 1

Figure 5.3 shows the wavelength selection histogram for generation 400. Figure 5.4 shows the wavelength selection histogram for generation 500.

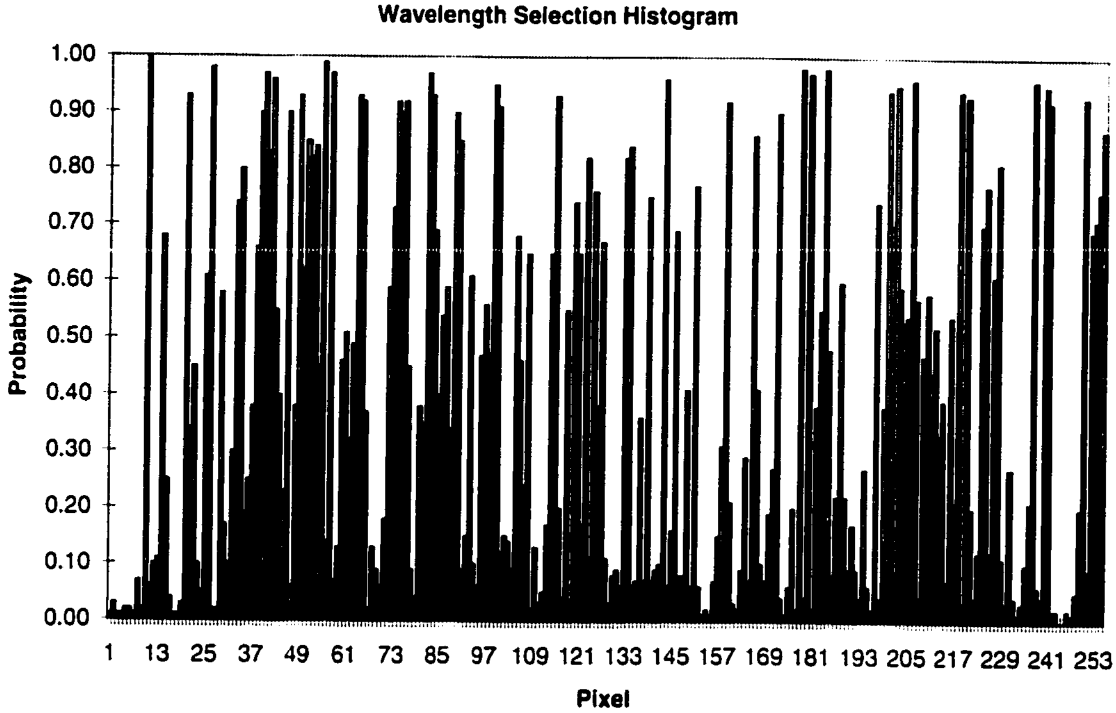


Figure 5.3: Wavelength Selection Histogram for Generation 400

The fitness trend in Figure 5.2 shows that the most significant improvements occur in the first 50 generations, for these data. After generation 50, the fitness keeps increasing but at a slower rate. The wavelength selection histograms for generation 400 and 500 are very similar but exhibit some differences. It is difficult from the fitness trend and the histogram homogeneity to state definitely that there is a significant amount of settling. Therefore, a study of the error trend is necessary. Figure 5.5 shows the error trend for the fitness criterion that yielded the best results for training data set 1. Figure 5.5 shows that the error rate has significantly settled after generation 200. Note: the error rates for generation 1 and generation 50 were estimated based on the known fitness value as the error rates were not stored for generation 1 and 50. Convergence was not expected to be attained for generations below 50; therefore, error rates were not stored for generations below 50.

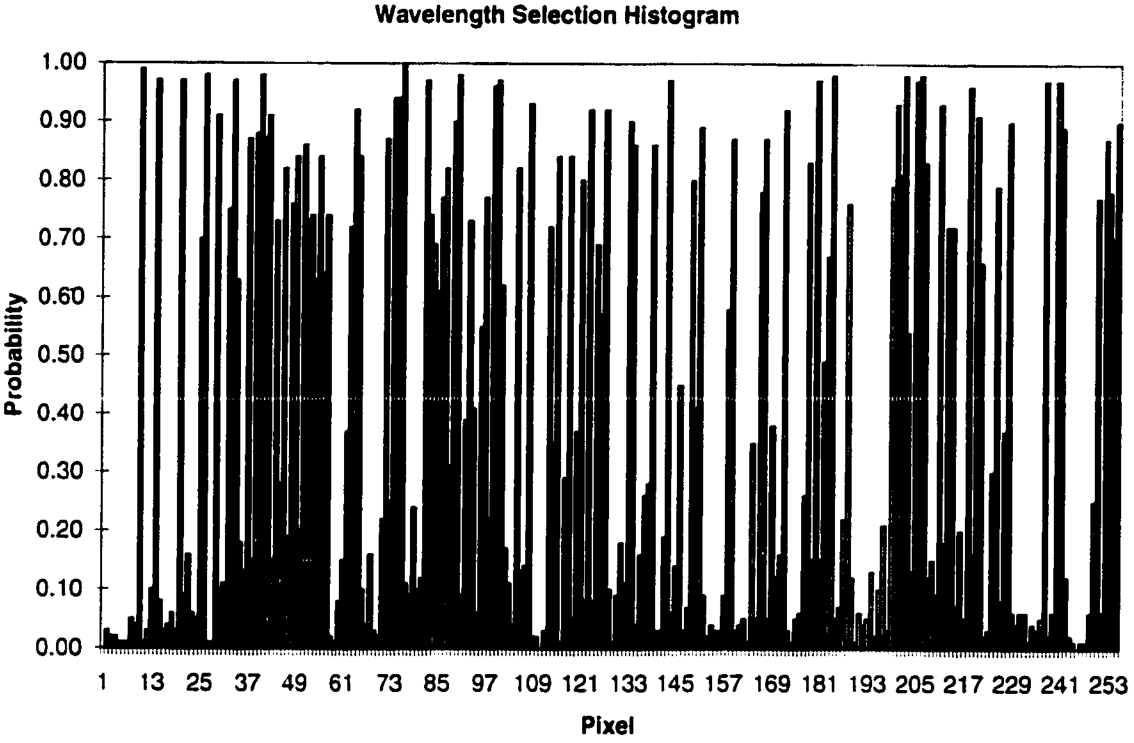


Figure 5.4: Wavelength Selection Histogram for Generation 500

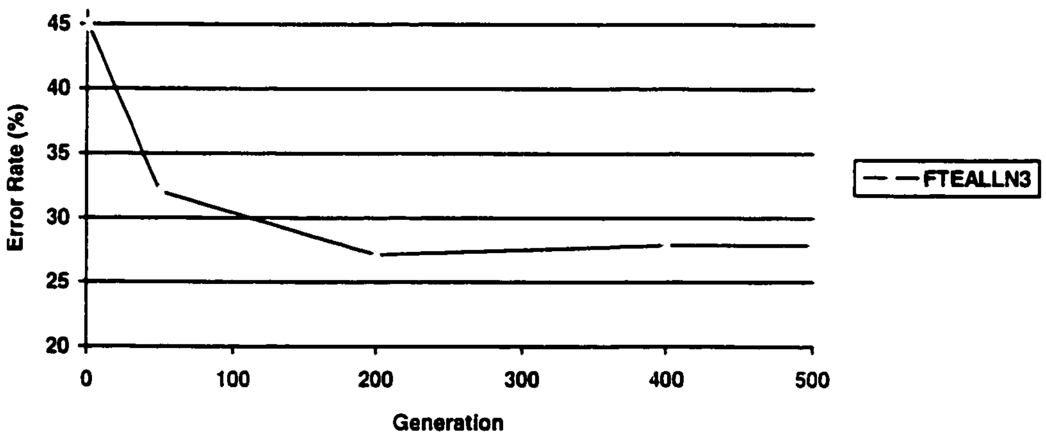


Figure 5.5: Error Trend for Best Fitness Criterion

5.2.4 Member Replacement

To make incremental improvements in the selection of features, the members with the worst fitness values are replaced by the members with the best fitness values. The member replacement is a straight substitution. The bottom twenty percent of the members are directly replaced by the top twenty percent of the members. The details of member replacement are described below.

Member replacement is performed in two steps:

1. Order the members according to the fitness calculated.
2. Replace the worst members with the best members.

5.2.4.1 Order the Members

The first step in ordering the members is to determine their relative order based on the fitness vector. The order vector, $\mathbf{o}$, specifies in ascending order, the relative position of each member k such that o_K identifies the member with the maximum fitness value and o_1 identifies the member with the minimum fitness value. That is, o_q identifies the member k which has the q^{th} lowest fitness value such that $f_{o_q} > f_{o_r}$ where $q > r$, and $f_{o_q} < f_{o_t}$ where $q < t$. The order vector is determined by comparing the f_k values with each other.

The second step is to create an ordered population matrix, $\mathbf{W}_A$, from the unordered population matrix, $\mathbf{W}$, such that $f_{A,q} > f_{A,r}$ where $q > r$. This is accomplished according to the algorithm below:

```
FOR $q = 1$ TO $K$
    $\mathbf{w}_{A,q} = \mathbf{w}_{o_q}$
    $f_{A,q} = f_{o_q}$
END FOR
```

For example, the ordered population matrix, $\mathbf{W}_A$, for the simple Cartesian example would be:

$$\mathbf{W}_A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

For this case, there was a tie between some members. The algorithm assumes, in the case of a tie, the first member found, i.e., lower k, is placed ahead of the second member found.

5.2.4.2 Replace Worst Members with Best Members

This step involves replacing the members with the lowest fitness values with those having the highest fitness values. In other words, the new population, $\mathbf{W}_B$, is created from the ordered population, $\mathbf{W}_A$, such that the members with the lowest N_R fitness values are replaced directly by the members having the highest N_R fitness values. The members with the highest $(K - N_R)$ fitness values remain unchanged. The algorithm for creating $\mathbf{W}_B$ is shown below:

```
FOR $q = 1$ TO $N_R$
    $\mathbf{w}_{B.q} = \mathbf{w}_{A.K-q-N_R}$
END FOR
FOR $q = N_R + 1$ TO $K$
    $\mathbf{w}_{B.q} = \mathbf{w}_{A.q}$
END FOR
```

A 20% replacement rate was used for our analyses. In other words, N_R was 20 for a population size $K = 100$.

Using the simple Cartesian example and $N_R = 1$, the new population matrix, $\mathbf{W}_B$, is created as shown below:

$$\mathbf{W}_B = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

5.2.5 Recombination

To this point, the new population matrix that has been created, has only had incremental improvements made to it. To add diversity and explore the solution space a little more, the members that were replaced, now called parents or the replacement population, are recombined to form new members, that is, children. Recombining only the members that were replaced, leaving the other members as they were, eliminates the need to recalculate the fitness value for the members that were not recombined. This reduces the computation time while still allowing adequate exploration. The population matrices before and after recombination have the same wavelength selection histogram, but the information is mixed up among the members that were replaced. The number of members to replace must be an even number for recombination.

Recombination involves creating a new population matrix, $\mathbf{W}_C$, from the population matrix, $\mathbf{W}_B$, which was created during the replacement stage. The new population matrix is created by forming new members, i.e., children, by recombining paired members, i.e., parents, that were replaced in the member replacement step.

The recombination method used here is Holland's classical one-point crossover, **1X**. This method swaps bitsegments between two specified members of the population after a common breakpoint or crossover point. Each pairing, i.e., a set of parents, produces a second pairing, i.e., two children. An example is shown in Figure 5.6.

The recombination steps are:

1. Pair up the members that were replaced, in a random fashion.
2. For each pair determine randomly a crossover point according to a uniform distribution.

3. Recombine each pair, i.e., parents, to form a new pair, i.e., children.

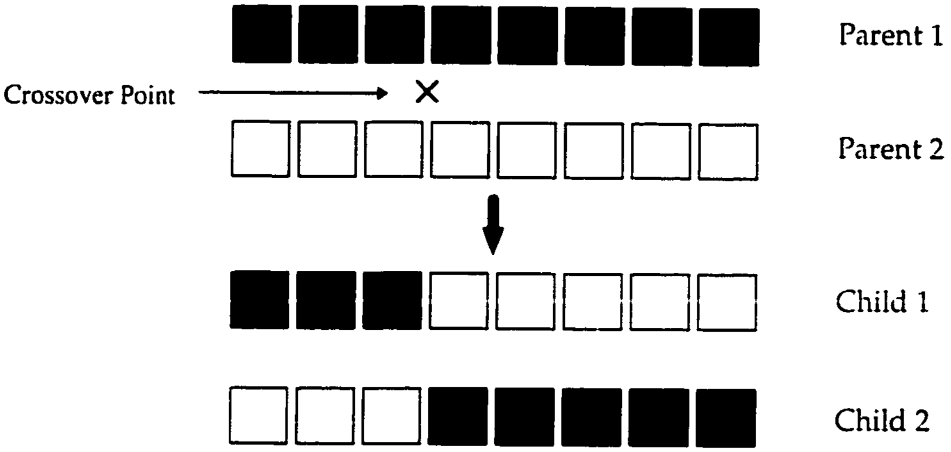


Figure 5.6 Example of 1X Recombination

5.2.5.1 A Recombination Example

Using the $\mathbf{W}_B$ determined from the simple Cartesian example, as defined in Section 5.2.4.2, suppose that the population members are paired randomly to form the following pairings:

$$\text{Pair 1:} \quad (\mathbf{w}_{B.1}, \mathbf{w}_{B.4})$$

$$\text{Pair 2:} \quad (\mathbf{w}_{B.2}, \mathbf{w}_{B.3}).$$

Suppose then, that the crossover point for Pair 1 is after feature 1, and for Pair 2 is after feature 2. The resultant children are shown below:

$$\begin{array}{cccc}
 & \text{Parents} & & \text{Children} \\
 & \mathbf{w}_{B.1} = [0 \quad 0 \quad 1] & & \mathbf{w}_{C.1} = [0 \quad 0 \quad 1] \\
\text{Pair 1} & \times & \Rightarrow & \\
 & \mathbf{w}_{B.4} = [0 \quad 0 \quad 1] & & \mathbf{w}_{C.2} = [0 \quad 0 \quad 1] \\
 & \mathbf{w}_{B.2} = [1 \quad 0 \quad 1] & & \mathbf{w}_{C.3} = [1 \quad 0 \quad 0] \\
\text{Pair 2} & \times & \Rightarrow & \\
 & \mathbf{w}_{B.3} = [0 \quad 1 \quad 0] & & \mathbf{w}_{C.4} = [0 \quad 1 \quad 1].
\end{array}$$

Therefore, the new population, $\mathbf{W}_C$, is produced as shown below:

$$\mathbf{W}_C = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}.$$

5.2.6 Mutation

Mutation, like recombination, is a step in the exploration search of the solution space. Mutation involves randomly toggling or inverting the select status for a small percentage (i.e., in our case five percent) of the population matrix elements (for the same reasons given in recombination, only the replacement population is mutated). The purposes of mutation are to create greater diversity, and to prevent features or wavelengths from disappearing from the solution space. The details of mutation are discussed below.

The result of the mutation step is a new population matrix, $\mathbf{W}_D$, created from the population matrix, $\mathbf{W}_C$. To maintain diversity in the population, a certain proportion of the replacement members' elements of $\mathbf{W}_C$ are toggled in a random fashion. If a given element is chosen to be toggled, and its present status is a "0" (or "1") then it will be changed to a "1" (or "0").

Suppose the population matrix, $\mathbf{W}_C$, is defined as shown below:

$$\mathbf{W}_C = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The second dimension in this $\mathbf{W}_C$ is deselected for all members. Therefore, without the mutation operator, this dimension is lost to future generations. However, the mutation operator can, statistically, bring this dimension back into a future population, so that it is not lost forever from the solution search space.

The mutation rate used for our data was 0.05; or 5% of the replacement members' elements were toggled.

5.2.6.1 A Mutation Example

Using the simple Cartesian example for $\mathbf{W}_C$ from Section 5.2.5, suppose that our mutation rate is 0.10. Considering that our population matrix has twelve elements, on average one element will be chosen to be toggled. Suppose that $w_{C,12}$ is chosen to be toggled, therefore $\mathbf{W}_D$ is formed as shown below:

$$\mathbf{W}_D = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}.$$

Note that $w_{C,12}$ has now become a "1" from a "0".

5.2.7 Wavelength Selection Vector

The wavelength selection vector is the final result of the GA process. It specifies which wavelengths are to be used in the MD model. It is determined from the histogram of the final population (i.e., the population at termination). The wavelength selection vector is saved so that the MD calibration can be determined using the wavelengths identified in the wavelength selection vector.

The wavelength selection vector, $\mathbf{w}_S$, is created by reducing the population matrix, $\mathbf{W}$, of the final generation. This is done in two steps as shown below:

1. Determine the wavelength selection histogram of the population matrix.
2. Discretise the wavelength selection histogram.

5.2.7.1 Histogram of Population Matrix

The histogram of $\mathbf{W}$ is created by calculating the relative proportion of selections of a particular wavelength across the members, for each wavelength. The elements of the wavelength selection histogram vector, $\mathbf{w}_H$, are expressed mathematically in

Equation 5.1:

$$w_{H,j} = \frac{1}{N}\sum_{i=1}^{N} w_{D,ij} \tag{5.1}$$

where $w_{H,j}$ is an element of the wavelength selection histogram vector, $\mathbf{w}_H$, for wavelength j, and $w_{D,ij}$ is an element of the population matrix, $\mathbf{W}_D$, after mutation, for member i and wavelength j. Using $\mathbf{W}_D$ from the simple Cartesian example in Section 5.2.6, $\mathbf{w}_H$ is:

$$\mathbf{w}_H = [0.25 \quad 0.50 \quad 0.75].$$

5.2.7.2 Discretise the Histogram

The wavelength selection vector, $\mathbf{w}_S$, is created from the wavelength selection histogram vector, $\mathbf{w}_H$, by discretising the elements of $\mathbf{w}_H$ according to the algorithm below:

```
IF $w_{H,j} \geq 0.5$
    THEN $w_{S,j} = 1$
    ELSE $w_{S,j} = 0$
END IF.
```

An element of $\mathbf{w}_H$ must equal or exceed the defined threshold of 0.5 for that wavelength to be selected.

Using $\mathbf{w}_H$ from the simple Cartesian example, $\mathbf{w}_S$ becomes:

$$\mathbf{w}_S = [0 \quad 1 \quad 1].$$

5.2.7.3 An Example using the Biological Data

The histogram produced from training data set 1 is shown in Figure 5.7 while the corresponding wavelength selection vector is shown in Figure 5.8.

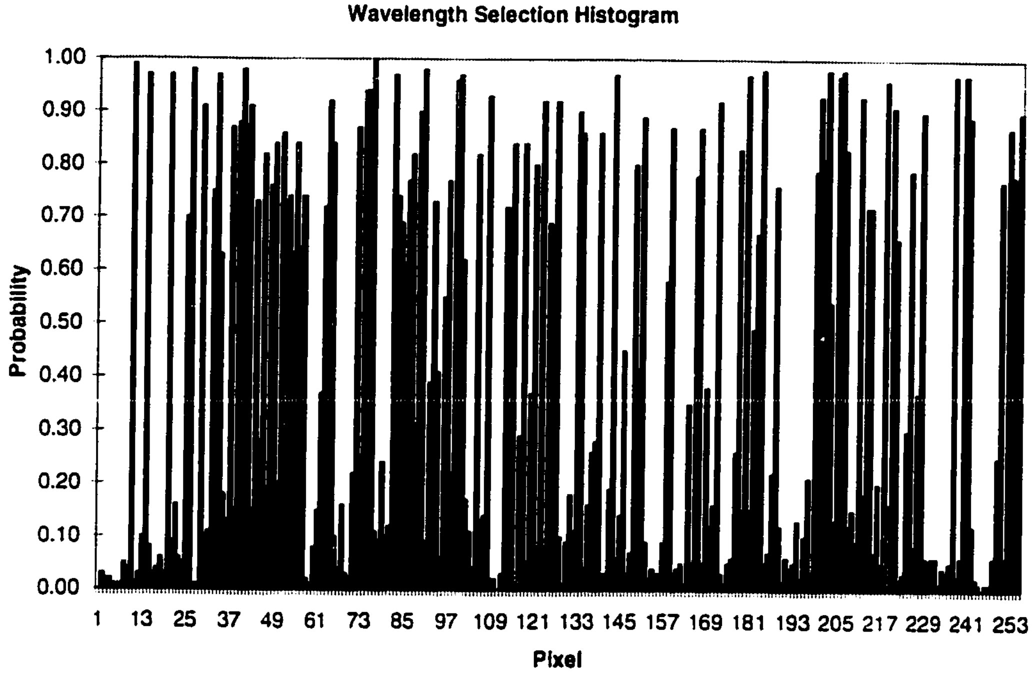


Figure 5.7: Wavelength Selection Histogram of Population Matrix

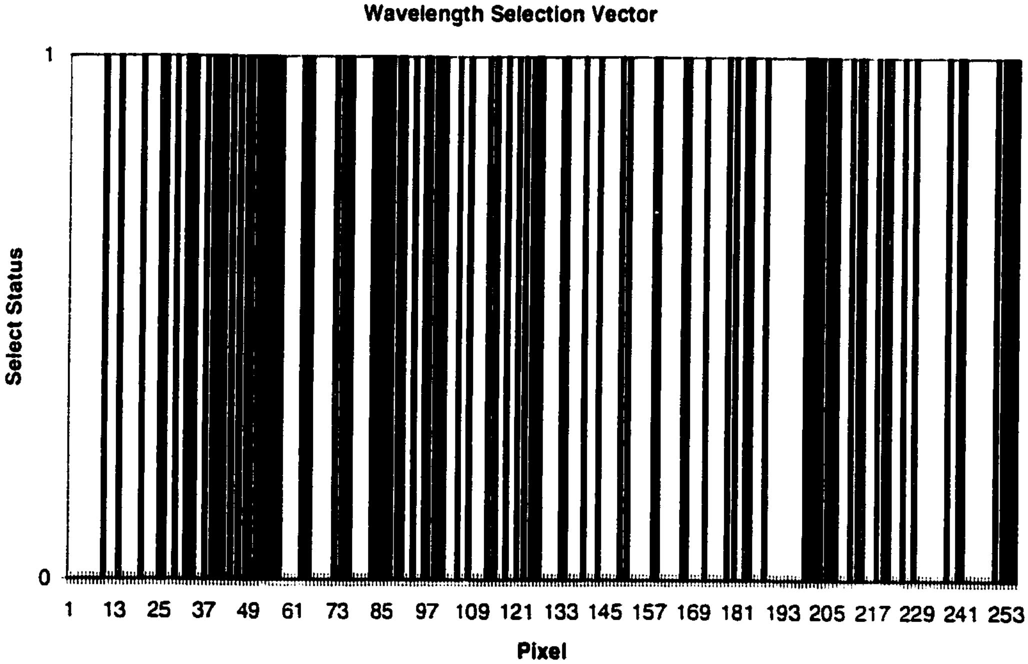


Figure 5.8: Wavelength Selection Vector

5.2.8 Calibrate Based on Wavelength Selection Vector

The final step in the GA process is to create and save the MD calibration model, from which classification can be performed on unknown specimens. This step is essentially one iteration of the evaluation stage as described in Section 5.2.2, except without the fitness calculation. Also, the only member to be evaluated is the wavelength selection vector. The steps are:

1. Create New Transformed Absorbance Matrix using only Selected Wavelengths.
2. Determine Principal Components of New Transformed Absorbance Matrix.
3. Perform Mahalanobis Distance Calibration.

These steps are described in detail in Section 5.2.2.

5.3 Prediction Algorithm Modification Using Selected Wavelengths

To classify an unknown specimen using the reduced wavelength set, the MD prediction algorithm described in Section 4.3 requires modification. The new algorithm is essentially identical to the previous one, except for the data transformation step. The data transformation modifications are as described in Section 5.2.2, i.e., all transforms are performed excluding PCA, then the new reduced transformed absorbance matrix is created as specified by the wavelength selection vector, and then PCA is performed. The complete algorithm is shown in Figure 5.9. All steps have previously been described either in Chapter 4 or 5.

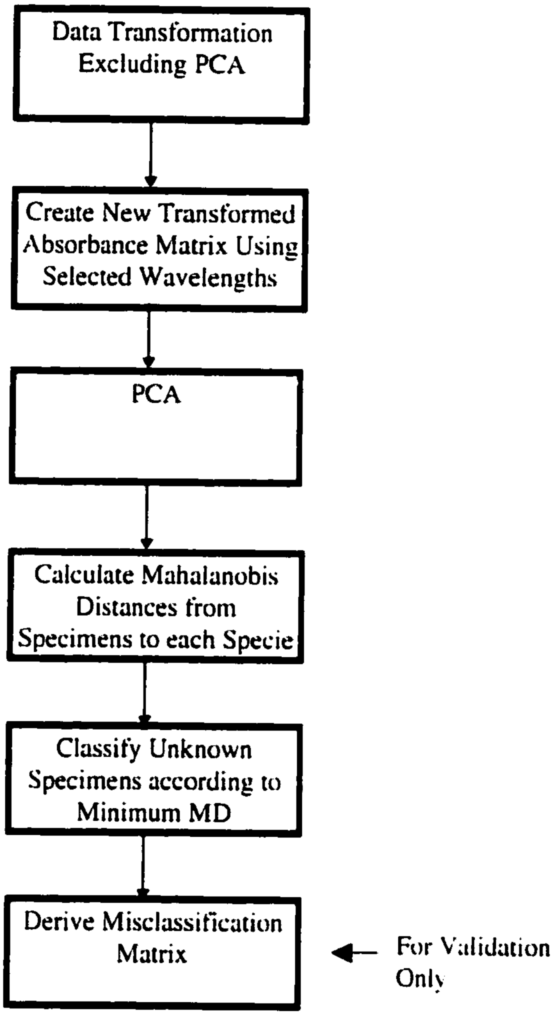

Data Transformation Excluding PCA
Create New Transformed Absorbance Matrix Using Selected Wavelengths
PCA
Calculate Mahalanobis Distances from Specimens to each Specie
Classify Unknown Specimens according to Minimum MD
Derive Misclassification Matrix
For Validation Only

Figure 5.9: Modified Mahalanobis Distance Prediction Algorithm

Chapter 6

Results of Analyses

6.1 Introduction

Analyses were performed to classify all seven species simultaneously, as well as the case of classifying urine from blood. The analyses discussions in this chapter are divided into two sections: one (i.e., Section 6.2) that deals with the analyses that attempt to classify all seven species, and one (i.e., Section 6.3) that deals with the analyses classifying urine from blood. KNN, MD and GA-MD methods were used to classify all species while MD and GA-MD were used to classify urine from blood. All analyses used data that were transformed using SM11-2D11. Some of the analyses used autoscaling while others did not. All MD and GA-MD analyses used principal component scores for the final features while KNN analyses used either wavelengths or PC scores. As defined in Section 1.7.1, three separate data sets were created, consisting of training and prediction data. For final comparisons of the methods, analyses were performed with each data set, and the results averaged over the three data sets, for a given case. Examples of results of analyses are provided in Appendix E.

6.2 Classification of All Species

The data were first analysed using the KNN method. The results were then compared against those obtained using the MD method. Section 6.2.1 covers this comparison. The MD method was then optimised using GA selected wavelengths. The results of this optimisation are described in Section 6.2.2.

6.2.1 KNN Versus Mahalanobis Distance Classification

The data were analysed using the data transformation SM11-2D11 and autoscaling. Both the wavelength and the principal component dimensional spaces were used for the KNN analysis. Only the principal components were used for the MD analysis. The results are shown in Table 6.1.

Table 6.1: Error Rates for KNN and Mahalanobis Distance with Autoscaling

	KNN							Mahalanobis Distance		
	Using Wavelengths			Using Principal Components				Using Principal Components		
Data Set	Train'g Error Rate (%)	Pred'n Error Rate (%)	# of NN's	Train'g Error Rate (%)	Pred'n Error Rate (%)	# of NN's	# of PC's	Train'g Error Rate (%)	Pred'n Error Rate (%)	# of PC's
1	38.89	39.29	1	38.38	40.00	1	26	12.63	30.71	26
2	39.90	37.86	1	40.40	36.43	1	31	12.63	30.71	31
3	41.41	41.43	1	41.41	38.57	1	29	15.15	30.00	29
Mean	40.07	39.53	1	40.06	38.33	1	28.7	13.47	30.47	28.7
S.D.	1.27	1.80	-	1.54	1.80	-	2.5	1.45	0.41	2.5

Note: Train'g Error Rate is the Error Rate calculated by applying the calibration model to the training data. Pred'n Error Rate is the Error Rate calculated by applying the calibration model to the prediction data. Therefore Predictability = 100% - Pred'n Error Rate (%).

The prediction error rates for the KNN method are not significantly different whether using wavelengths or whether using PC scores for the features. The training and prediction error rates for KNN are essentially the same, with the prediction error rate slightly better than the training error rate. The prediction error rate for the MD method is significantly lower than for the KNN method--38.33% and 30.47% for KNN and MD respectively. Therefore, under these conditions, the MD prediction error rate is 20.5% lower than that for KNN. The improvement is expected as the MD method is a more sophisticated method than the KNN method.

The data were also analysed without autoscaling. The results of these are shown in Table 6.2. The spread between the prediction error rates for the MD and KNN methods is even larger for this condition, i.e., prediction error rates being 46.19% and 25.72% for KNN and MD respectively, giving a 44.3% improvement. The autoscale transform improves the predictability for the KNN method but degrades the predictability for the MD method.

Table 6.2: Error Rates for KNN and Mahalanobis Distance with No Autoscaling

	KNN			Mahalanobis Distance		
	Using Wavelengths			Using Principal Components		
Data Set	Training Error Rate (%)	Pred'n Error Rate (%)	# of NN's	Training Error Rate (%)	Pred'n Error Rate (%)	# of PC's
1	37.37	50.71	5	11.62	24.29	29
2	40.91	42.86	3	12.63	26.43	23
3	43.94	45.00	5	13.64	26.43	23
Mean	40.74	46.19	4.3	12.63	25.72	25.0
S.D.	3.29	4.06	1.15	1.01	1.24	3.46

Table 6.3 shows the specie by specie, mean predictability for the best individual conditions (i.e., best transforms), for each of the KNN and MD methods--autoscale for KNN and no autoscale for MD.

Table 6.3: Comparison of Individual Mean Specie Predictability for Best KNN and Best MD Conditions

		Specie	1	2	3	4	5	6	7	Error Rate (%)
KNN	Training	Mean	97.5	54.0	56.7	57.8	54.2	63.2	37.9	40.06
(with		S.D.	2.14	10.5	3.35	1.91	4.15	1.96	0	1.54
AS	Pred'n	Mean	93.3	60.0	61.7	40.0	56.7	62.5	51.7	38.33
& PC's)		S.D.	2.89	5.0	12.6	8.66	2.89	8.22	2.89	1.80
MD	Training	Mean	100	80.5	81.1	85.6	94.4	96.6	75.9	12.63
(with		S.D.	0	5.25	3.81	1.96	2.37	3.45	3.45	1.01
PC's but	Pred'n	Mean	96.7	58.3	71.7	66.7	85.0	81.7	60.0	25.72
no AS)		S.D.	2.89	5.77	5.77	2.89	13.2	2.89	8.66	1.24

Note: AS = autoscaling

Comparing the results using the best individual conditions for KNN and MD yields prediction error rates of 38.33% and 25.72% for KNN and MD respectively. Therefore, under these conditions, MD produced a 32.9% improvement over KNN in prediction error rate.

Urine (specie 1) is the only specie to achieve 95% predictability, and that using the MD method. Using KNN, urine predictability is only slightly lower, at 93.3%. Oxalate (specie 5) and iodoacetate (specie 6) are close to the 95% rate, achieving predictability rates of 85.0% and 81.7% respectively, using the MD method.

Table 6.4 shows the interspecie MD's for data set 1, without autoscaling. The mean interspecie MD's are also calculated. The means are calculated two ways: including and excluding the interspecie MD with respect to Urine. Since the interspecie MD's with respect to Urine are all relatively high, interpretation is enhanced by excluding it. This is indicated by the significant drop in the standard deviations.

Table 6.5 shows the relationship, i.e., correlation, between the mean interspecie MD's and the predictability of that specie. Even without plotting the numbers, the correlation is evident. As the MD approaches 6 and greater, the predictability approaches the 95% level as expected.

Table 6.4: Interspecie Mahalanobis Distances for Data Set 1 for Best MD Conditions

Specie Y	Specie X							Including Specie 1		Excluding Specie 1	
	1	2	3	4	5	6	7	Mean	S.D.	Mean	S.D.
1	0	11.62	11.58	11.50	12.80	13.18	12.19	12.15	0.71	-	-
2	-	0	3.21	2.53	5.37	4.54	2.24	4.92	3.49	3.58	1.34
3	-	-	0	2.89	4.90	5.39	2.18	5.03	3.44	3.71	1.37
4	-	-	-	0	5.66	5.22	2.48	5.05	3.45	3.76	1.55
5	-	-	-	-	0	5.92	4.91	6.59	3.07	5.35	0.45
6	-	-	-	-	-	0	4.78	6.51	3.31	5.17	0.54
7	-	-	-	-	-	-	0	4.80	3.83	3.32	1.40

Table 6.5: Correlation Between Predictability and Interspecie MD

Specie	1	2	3	4	5	6	7
Predictability (%) (Mean for all Data Sets)	96.7	58.3	71.7	66.7	85.0	81.7	60.0
Mean Interspecie MD for Data Set 1 (Excluding Specie 1 MD for 2-7)	12.15	3.58	3.71	3.76	5.35	5.17	3.32

6.2.1.1 Summary of KNN Versus MD

The Mahalanobis Distance method outperformed the KNN method in terms of classification accuracy for unknown specimens. The results are summarised in Table 6.6. The autoscaling transform improved the performance of the KNN method but degraded the performance of the MD method. This may be due to the fact that MD inherently performs its own scaling, which may be more optimised than the autoscale transform. This redundancy may be the cause for the degradation using autoscaling for MD. This suggests that caution should be used when applying the autoscale transform. There is good correlation between the mean interspecie MD and the predictability of a specie.

Note: The test statistic for determining the number of PC's was not monotonic for a few of the analyses; therefore some judgement was used in determining the number of PC's for these cases. Although the smoothing of the test statistic helped significantly, the nature of the data (i.e., discrete number of errors) precluded the test statistic from working perfectly for every case.

Table 6.6: Error Rate Summary

Method	Prediction Data Error Rate (%)
KNN (with autoscale)	38.33
KNN (no autoscale)	46.19
MD (with autoscale)	30.47
MD (no autoscale)	25.72

6.2.2 MD Versus GA Optimised MD (GA-MD) Classification

An overlay to applying the GA as defined in Chapter 5, is the process of determining the best fitness criterion. The most important variable in the GA is the fitness criterion and its identity is crucial. Section 6.2.2.1 describes how this is done.

6.2.2.1 Identification of the Best Fitness Criterion

To identify the best fitness criterion, data set 1 with autoscaling was analysed using the GA-MD method using several fitness criteria. The selection of the best fitness criterion was done using the autoscaling transform as the autoscaling transform had produced the best results for KNN. The no autoscaling analyses were not done until the late stages of the study; the no autoscaling analyses used the best fitness criterion as determined by the autoscaling case. The fitness equations are described in Section 5.2.2.4. Figure 6.1 shows the prediction error rate versus number of generations for four of the fitness criteria. FTEALLN1 is the most aggressive in reducing the number of wavelengths, while FTEALLN4 is the least aggressive. The trend for both FTEALLN1 and FTEALLN2 is that the error rate increases from generation 200 to 500. This suggests that underfitting is occurring as the algorithm was attempting to reduce the number of wavelengths quite aggressively. The trend for FTEALLN4 is that the error rate decreases from generation 200 to 500 but its end value is higher than for FTEALLN3. This suggests that overfitting is occurring as FTEALLN4 is not reducing the number of wavelengths aggressively enough. The trend for FTEALLN3, an intermediately aggressive calculation with respect to reducing wavelengths, is that the error rate has a decreasing trend up to generation 200, then the error rate is reasonably constant from generation 200 to 500. Also, the error rate that FTEALLN3 settles to is lower than those for the other three fitness criteria. This suggests that an appropriate level of fitting is occurring for the FTEALLN3 fitness criterion.

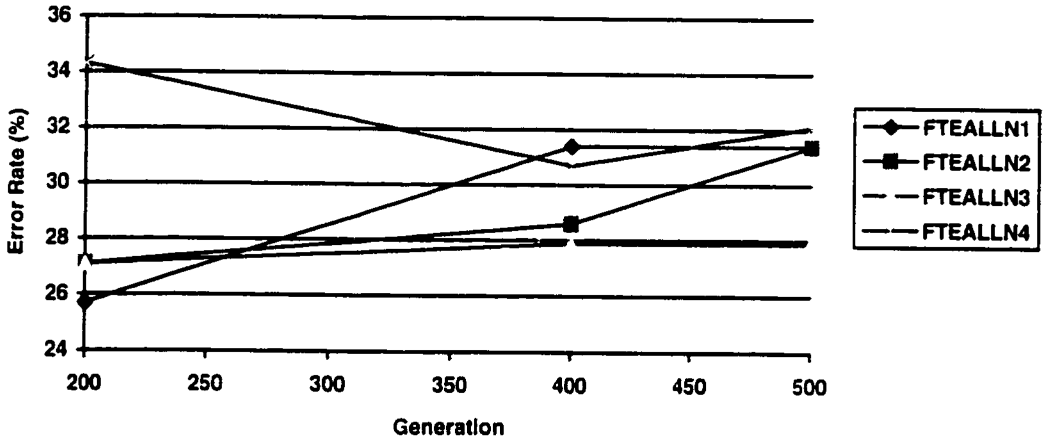


Figure 6.1: Prediction Error Rate Trend for Various Fitness Criteria

Another way to look at the error rate is to plot it versus the number of wavelengths. This is shown in Figure 6.2.

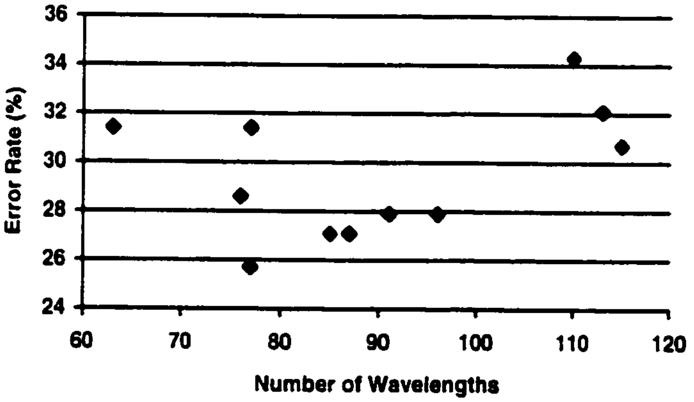


Figure 6.2: Prediction Error Rate Versus Number of Wavelengths

Figure 6.2 shows that the error rate is minimised when the number of wavelengths is between 77 and 96. Error rates are larger outside this range. This is independent of the fitness criterion. However, the criterion FTEALLN3 tended to be maximised when the number of wavelengths was in the range of 87 to 96, which is within the optimum range for number of wavelengths according to Figure 6.2. Based on these results, FTEALLN3 was chosen as the best fitness criterion for these data, i.e., transform SM11-2D11 and autoscaling.

6.2.2.2 Results of Analyses using the Best Fitness Criterion

Table 6.7 shows results of analysing the data using fitness criterion FTEALLN3, with and without autoscaling.

Table 6.7: GA-MD Results Using FTEALLN3

	GA - MD with Autoscaling				GA - MD with No Autoscaling			
Data Set	Training Error Rate (%)	Pred'n Error Rate (%)	# of λ's	# of PC's	Training Error Rate (%)	Pred'n Error Rate (%)	# of λ's	# of PC's
1	10.61	27.86	96	20	8.59	33.57	107	22
2	11.11	27.14	98	20	8.08	30.71	104	22
3	10.10	30.00	99	20	10.61	25.71	97	23
Mean	10.61	28.33	97.7	20	9.09	30.00	102.7	22.3
S.D.	0.51	1.49	1.5	0	1.34	3.98	5.1	0.58

The number of wavelengths (# of λ's) ranged between 96 and 99 for autoscaling and between 97 to 107 for no autoscaling. The mean prediction error rates are 28.33% and 30.00% for autoscaling and no autoscaling respectively. However, the mean training error rates are 10.61% and 9.09% for autoscaling and no autoscaling respectively. This suggests that some overfitting is occurring for the no autoscaling case. The fitness criterion, i.e., FTEALLN3, was determined by optimising the autoscaling case, and is possibly not optimum for the no autoscaling case. A fitness criterion which more aggressively reduces the number of wavelengths is likely required for the no autoscaling situation. For example, FTEALLN2 may work better for the no autoscaling case, or perhaps a calculation somewhere between FTEALLN3 and FTEALLN2 may perform better. These results suggest that it is important to determine the best criterion for a particular set of data, with a particular transformation, to optimise performance. These results also suggest that an algorithm which determines the optimum number of wavelengths, for a particular set of data, would also be appropriate.

Table 6.8 shows the comparison of the classification performances of MD versus GA-MD for both the autoscaling and no autoscaling cases.

Table 6.8: Comparison of MD Versus GA-MD Results

		Mahalanobis Distance				GA - MD			
Data Transform	Stat.	Training Error Rate (%)	Pred'n Error Rate (%)	# of λ's	# of PC's	Training Error Rate (%)	Pred'n Error Rate (%)	# of λ's	# of PC's
Autoscale	Mean	13.47	30.47	255	28.7	10.61	28.33	97.7	20
	S.D.	1.45	0.41	0	2.5	0.51	1.49	1.5	0
No Autoscale	Mean	12.63	25.72	255	25.0	9.09	30.00	102.7	22.3
	S.D.	1.01	1.24	0	3.46	1.34	3.98	5.1	0.58

GA-MD was able to reduce the error rate to 28.33% from 30.47% for MD, for the autoscaling case. This 7.0% reduction in error rate is statistically significant as the mean error rates for MD and GA-MD, for the autoscaling case, are more than 2 standard deviations apart. The percentage improvement was even larger for the four species with the worst predictabilities (i.e., serum, citrate, EDTA, and heparin) and that improvement was 8.0%. However, for the no autoscaling case, GA-MD increased the error rate to 30.00% from 25.72% for MD. There are two reasons for this.

The first reason for the increase in error for the no autoscaling case is that the fitness criterion was optimised for the autoscaling case and not for the no autoscaling case. Table 6.7 and Figure 6.2 show that when the number of wavelengths exceeds 98, the error rate increases significantly. Two of the data sets for the no autoscaling case exceeded 98 wavelengths, thus producing a larger error rate. If only the one data set (i.e., data set 3) is considered for the no autoscaling case, when the number of wavelengths was 97, the error rate was reduced to 25.71% from 26.43%, when GA optimisation was used. Therefore, if an optimised fitness criterion had been used for the no autoscaling scale, the data suggest that an improvement would be achieved.

The second reason for the increase in error rate for the no autoscaling case is contained in Table 6.9. Table 6.9 is derived from the same data analysis as that used for Table 6.8, except that the number of PC's selected in determining the results shown in Table 6.9, optimised the error rate for the prediction data rather than for the training data. Comparing the number of PC's shown in Tables 6.8 and 6.9, the no autoscaling case for GA-MD shows a greater difference in the number of PC's between those required for optimising the training error rate and those required for optimising the prediction error rate, than for the other 3 situations (i.e., MD with no autoscaling, MD with autoscaling, and GA-MD with autoscaling). The PC difference, between optimising for prediction and training, for the GA-MD no autoscaling case, is 3.7 while it is less than 2.4 for the other 3 cases. This suggests that using cross-validation or a separate validation data set within the training algorithm may help in choosing a more optimum number of PC's, thus reducing the differences. In any case, the results as tabulated in Tables 6.7 and 6.9 provide strong evidence that GA-MD does produce better results than MD, and does it with fewer wavelengths, and possibly with fewer PC's, given that proper optimisation is done.

Table 6.9: Comparison of MD Vs. GA-MD Results with Optimising # of PC's for Best Prediction Results

		Mahalanobis Distance				GA - MD			
Data Transform	Stat.	Training Error Rate (%)	Pred'n Error Rate (%)	# of λ's	# of PC's	Training Error Rate (%)	Pred'n Error Rate (%)	# of λ's	# of PC's
Autoscale	Mean	17.01	26.67	255	26.3	11.79	26.43	97.7	21.7
	S.D.	4.21	1.49	0	2.5	1.27	1.24	1.5	1.2
No Autoscale	Mean	14.14	23.57	255	26.3	11.28	25.24	102.7	26.0
	S.D.	2.20	1.24	0	3.8	0.29	1.09	5.1	2.6

6.2.2.3 Discussion on the Wavelengths Selected

A question that may be asked is, "Are the same wavelengths selected each time (i.e., for the same type of data and from run to run) and if not does this matter? ". Figures 6.3 and 6.4 show the wavelengths selected for data sets 1 and 3 respectively, for the autoscaling case. There are a number of similarities between these two wavelength sets those being: most of the wavelengths in the pixel 35 to 60 range are selected, there are wavelengths selected across the whole spectrum, and there are no spectral gaps wider than 10 or 12 pixels. But there are also significant differences: data set 3 has few wavelengths selected below pixel 25, and approximately only half of the wavelengths selected across the spectrum are common between the two data sets. Despite these significant differences, Table 6.10 shows there is no loss in mean predictability (if anything a slight improvement) when wavelengths selected from the GA analysis of one data set are used to predict on the other data set. The slight improvement when using wavelengths from another data set is likely due to the fact that some of the specimens in the prediction data of one data set are used in the training data of the other data set. Therefore, the apparent significant differences in wavelength selection are really not that significant in terms of predictability. This is likely due to the fact that there is high correlation between adjacent pixels (especially at the wavelength resolution used) and this makes predictability not particularly sensitive to whether pixel *p* is selected or whether pixel *p±1* is selected. In other words, the solution space with respect to wavelengths, is relatively shallow. Even though different wavelengths may be chosen from one analysis to another, each wavelength set has utility.

For future work: to reduce the apparent differences in wavelength sets, the resolution could be reduced by excluding wavelengths or by grouping the wavelengths into subsets of *w* wavelengths, where *w* could be on the order of 10. The analysis would then be based on the reduced or grouped set.

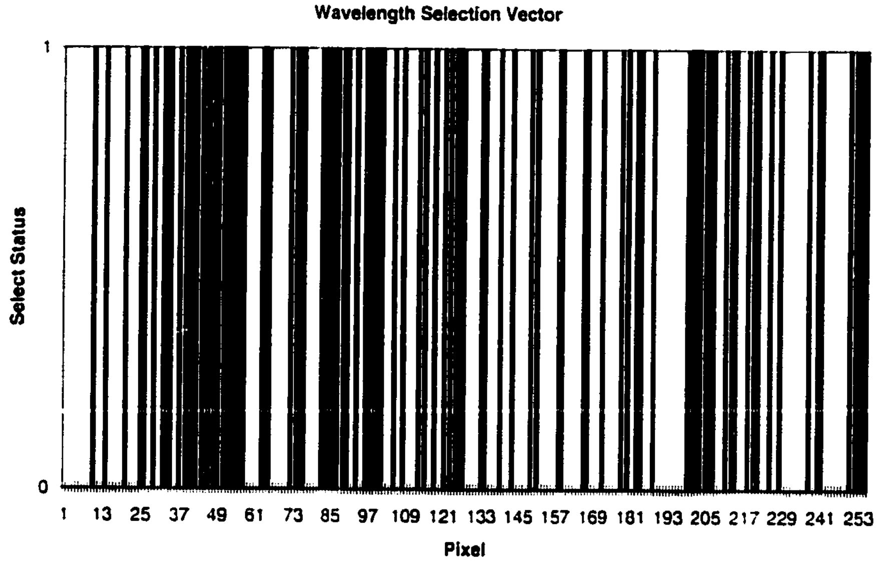


Figure 6.3: Wavelengths Selected for Data Set 1

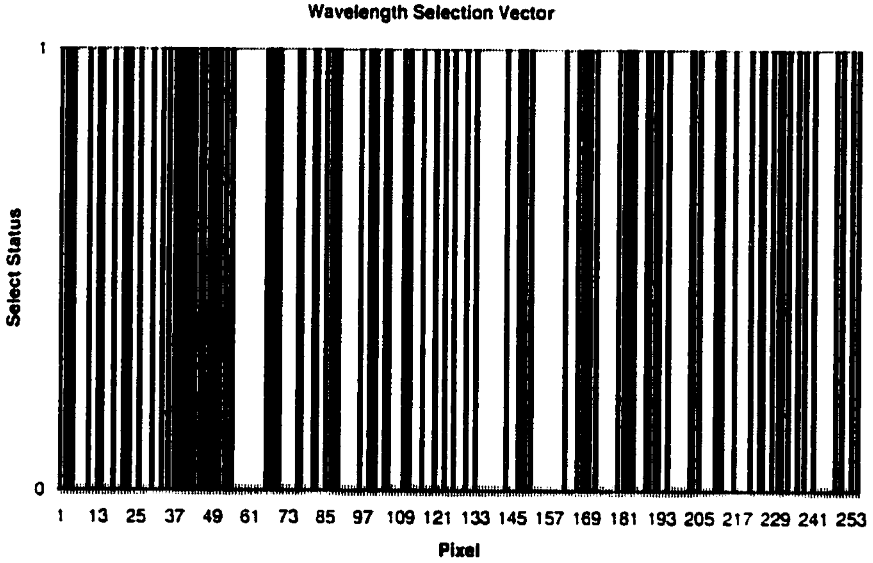


Figure 6.4: Wavelengths Selected for Data Set 3

Table 6.10: Error Rates using Wavelengths from Optimisation with another Data Set (with Autoscaling)

		GA - MD (Using Wavelengths from Same Data Set)				GA - MD (Using Wavelengths from Other Data Set)			
PC's Optimised for	Data Set	Training Error Rate (%)	Pred'n Error Rate (%)	# of λ's	# of PC's	Training Error Rate (%)	Pred'n Error Rate (%)	# of λ's	# of PC's
Training	1	10.61	27.86	96	20	13.13	30.71	99	27
	3	10.10	30.00	99	20	12.63	22.86	96	21
	Mean	10.36	28.93	97.5	20	12.88	26.79	97.5	24
	S.D.	0.36	1.51	2.1	0	0.35	5.55	2.1	4.2
Prediction	1	10.61	25.71	96	21	13.13	27.86	99	30
	3	13.13	27.86	99	23	12.63	22.86	96	21
	Mean	11.87	26.79	97.5	22	12.88	25.36	97.5	25.5
	S.D.	1.78	1.52	2.1	1.4	0.35	3.54	2.1	6.36

6.2.3 Summary of the Results of Analyses Classifying All Species

To properly compare the performance of all three methods, i.e., KNN, MD and GA-MD, the same data including the same transformations (i.e., whether autoscaling is used or not) are used. Table 6.11 shows this comparison using the transform SM11-2D11, PC's, and autoscaling. The mean prediction error rates are summarised into Table 6.12. MD was able to reduce the prediction error rate by 20.5% over that produced by KNN. GA-MD improved the prediction error rate by a further 7.0%. These are statistically significant improvements. These improvements are also conservative estimates, as MD performed better without autoscaling (refer to Tables 6.1 and 6.2). In fact, MD reduced the prediction error rate by 32.9% over that produced by KNN, comparing the best conditions for each. However, this study did not have a proper comparison for the GA-MD case without autoscaling, as the fitness criterion was not optimised for this case. The data, however, did provide evidence (refer to Table 6.7 - data set 3) that an improvement would be achieved for GA-MD without autoscaling, using an optimised fitness criterion.

Table 6.11: Comparison of KNN Vs. MD Vs. GA-MD Results Using the Same Data Transform (i.e., SM11-2D11, PC's, and Autoscaling)

		Specie	1	2	3	4	5	6	7	Error Rate (%)
KNN	Training	Mean	97.5	54.0	56.7	57.8	54.2	63.2	37.9	40.06
		S.D.	2.14	10.5	3.35	1.91	4.15	1.96	0	1.54
	Pred'n	Mean	93.3	60.0	61.7	40.0	56.7	62.5	51.7	38.33
		S.D.	2.89	5.0	12.6	8.66	2.89	8.22	2.89	1.80
MD	Training	Mean	100	82.8	77.8	85.6	94.4	92.0	75.9	13.47
		S.D.	0	6.90	5.08	1.96	2.37	1.96	3.45	1.45
	Pred'n	Mean	98.3	60.0	68.3	48.3	81.7	76.7	53.3	30.47
		S.D.	2.89	0	2.89	16.1	12.6	2.89	7.64	0.41
GA-MD	Training	Mean	100	86.2	84.4	91.1	93.1	93.1	79.3	10.61
		S.D.	0	3.45	5.10	6.97	2.37	3.45	9.15	0.51
	Pred'n	Mean	96.7	56.7	73.3	61.7	78.3	78.3	56.7	28.33
		S.D.	2.89	10.4	7.64	10.4	10.4	7.64	7.64	1.49

Table 6.12: Comparison of Mean Prediction Error Rates for KNN, MD and GA-MD under the Same Conditions

Method	KNN	MD	GA-MD
Prediction Error Rate (%)	38.33	30.47	28.33

Table 6.13 shows the best performance of the analyses, which used MD with transform: SM11-2D11, PC's, and no autoscaling. As already mentioned, if an optimised fitness criterion had been used for the no autoscaling case, GA-MD would likely have outperformed MD for the no autoscaling case also.

Table 6.13: Prediction Performance using Best Method (i.e., MD with transform SM11-2D11, PC's, No autoscale)

Specie	1	2	3	4	5	6	7	Mean Pred'n Error Rate (%)
Mean Predictability (%)	96.7	58.3	71.7	66.7	85.0	81.7	60.0	25.72

MD is able to classify urine (specie 1) to the 95% accuracy rate, while oxalate (specie 5) and iodoacetate (specie 6) are close to the 95% accuracy rate; viz., at 85.0% and 81.7% respectively. Serum (specie 2), citrate (specie 3), EDTA (specie 4), and heparin (specie 7) appear to be part of the same class, for this wavelength range, as their MD's are relatively small. A small surprise was that oxalate produces such a good NIR response, since the possible vibrational overtones are not listed in a common NIR response table, as shown in Table 1.2.

6.3 Classification of Urine Versus Blood

Since only the first objective (i.e., classifying urine from blood) has been achieved by classifying all species simultaneously, there may be a simpler solution, than the one found by classifying all species. This is done by reducing the problem to the simpler case of classifying urine (specie 1) from blood (i.e., species 2-7) only. The analyses involving the classification of urine (specie 1) from blood (species 2-7) used the Mahalanobis Distance method and the GA-MD method. The data were not analysed using KNN since MD was proven superior in classifying all the species. Also, autoscaling was not used as the previous analyses had shown that it did not enhance the performance.

6.3.1 Mahalanobis Distance Analyses

Table 6.14 shows the results of the MD analyses. The mean predictability rates for urine and blood are 95% and 99.72% respectively. Therefore, the MD method is able to achieve 95% predictability in classifying urine from blood. Nine (9) PC's were required for classifying urine from blood versus more than 20 PC's when classifying all the species.

Table 6.14: Error Rates for Mahalanobis Distance with No Autoscaling for Classifying Urine Vs. Blood

	Mahalanobis Distance Method				
Data Set	Training Error Rate (%)	Pred'n Error Rate (%)	Predictability of Urine (%)	Predictability of Blood (%)	# of PC's
1	0.51	1.43	90.00	100.00	9
2	0.00	0.71	100.00	99.17	9
3	0.51	0.71	95.00	100.00	9
Mean	0.34	0.95	95.00	99.72	9
S.D.	0.29	0.42	5.00	0.48	0

6.3.2 Genetic Algorithm Optimised Mahalanobis Distance Analyses

6.3.2.1 Finding the Best Fitness Criterion for Urine Versus Blood Classification

A variety of fitness criteria were tested on data sets 1 and 2 to determine the best criterion. Table 6.15 shows the results. For data set 1, criteria FTE1N8 to FTE1NSQR required 26 wavelengths and 9 PC's for optimum predictability. FTE1NSR4 required fewer wavelengths, i.e., 24, but more PC's, i.e., 11. For data set 2, criteria FTE1N2 to FTE1NSQR, all required 24 wavelengths and 10 PC's. Therefore, there is little difference in the results with respect to fitness criteria. This is likely due to the fact that each criterion is able to push the reduction of wavelengths to a minimum and achieve perfect classification, i.e., 100% for the training data. Therefore, the criterion chosen as optimum was FTE1N1, since it provided the most efficient calculation.

Table 6.15: Predictability of Various Fitness Criteria for Urine Vs. Blood Using GA-MD with No Autoscaling

		GA-MD					
Fitness Criterion	Data Set	Training Error Rate (%)	Pred'n Error Rate (%)	Predictability of Urine (%)	Predictability of Blood (%)	# of λ's	# of PC's
FTE1N8	1	0.00	0.71	95.0	100.00	26	9
FTE1N4	1	0.00	0.71	95.0	100.00	26	9
FTE1N2	1	0.00	0.71	95.0	100.00	26	9
FTE1N1	1	0.00	0.71	95.0	100.00	26	9
FTE1NSQR	1	0.00	0.71	95.0	100.00	26	9
FTE1NSR4	1	0.00	0.71	95.0	100.00	24	11
FTE1N2	2	0.51	0.71	95.0	100.00	24	10
FTE1N1	2	0.51	0.71	95.0	100.00	24	10
FTE1NSQR	2	0.51	0.71	95.0	100.00	24	10

6.3.2.2 Analyses Results using the Best Fitness Criterion

The results of applying FTE1N1 to all the data sets are shown in Table 6.16. The predictabilities of urine and blood are 95% and 100% respectively. This is achieved using an average of 23.7 wavelengths and 10.3 PC's. The predictabilities are slightly better, although not statistically significant, than those achieved by MD. However, GA-MD achieved it using significantly fewer wavelengths (i.e., 23.7 and 255 for GA-MD and MD respectively), although 1 more PC.

The GA analyses were then allowed to run past 500 generations, up to 999 generations, to determine if any more optimisation was possible. Figure 6.5 shows the predictability versus number of wavelengths for generations 500, 600, 800 and 999, for all 3 data sets. Although it was possible to lower the number of wavelengths and achieve 100% calibration, the predictability dropped to 90% for some of the analyses. Therefore, it appears that the minimum number of wavelengths is on the order of 25 or 26, to maintain a mean predictability of 95%.

Table 6.16: Predictability of Urine Vs. Blood with No Autoscaling Using GA-MD with Fitness Criterion, FTE1N1

Data Set	GA-MD					
	Training Error Rate (%)	Pred'n Error Rate (%)	Predictability of Urine (%)	Predictability of Blood (%)	# of λ's	# of PC's
1	0.00	0.71	95.00	100.00	26	9
2	0.51	0.71	95.00	100.00	24	10
3	0.00	0.71	95.00	100.00	21	12
Mean	0.17	0.71	95.00	100.00	23.7	10.3
S.D.	0.29	0	0	0	2.5	1.5

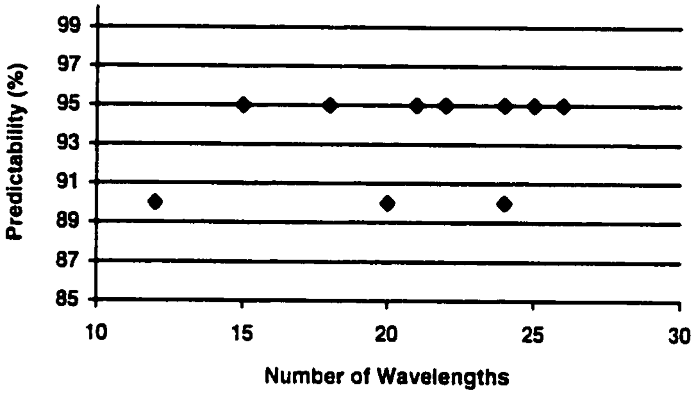


Figure 6.5: Predictability of Urine Versus Number of Wavelengths

Table 6.17 shows the interspecie Mahalanobis distances for data set 1 using the MD and the GA-MD methods. The mean interspecie MD's for the two methods are not significantly different. Both are above 6, consistent with achieving predictability better than 95%. These interspecie MD's, however, are significantly lower (i.e., 8.31 vs. 12.15), than those obtained when more PC's (i.e., approx. 25) are used, as shown in Table 6.5. Therefore, if better predictability than 95% is desired, more PC's should be used.

Table 6.17: Interspecie Mahalanobis Distances for Data Set 1 for MD and GA-MD Methods for Urine Vs. Blood Classification

Method	Specie Y	Specie X								
		1	2	3	4	5	6	7	Mean	S.D.
MD	1	0	7.38	8.17	7.22	9.43	9.53	8.14	8.31	0.98
GA-M.D	1	0	8.25	8.50	8.17	8.32	9.34	8.55	8.52	0.43

Figures 6.6 to 6.8 show the wavelengths selected for data sets 1 to 3 respectively. All three wavelength sets show wavelengths selected in the upper end of the spectrum, as expected from the spectral plots of the transformed data (refer to Figure 2.3). However, as in the "all specie" case, there are few common, coincidental wavelengths. For the same reasons given in Section 6.2.3.3, this can occur and not degrade predictability.

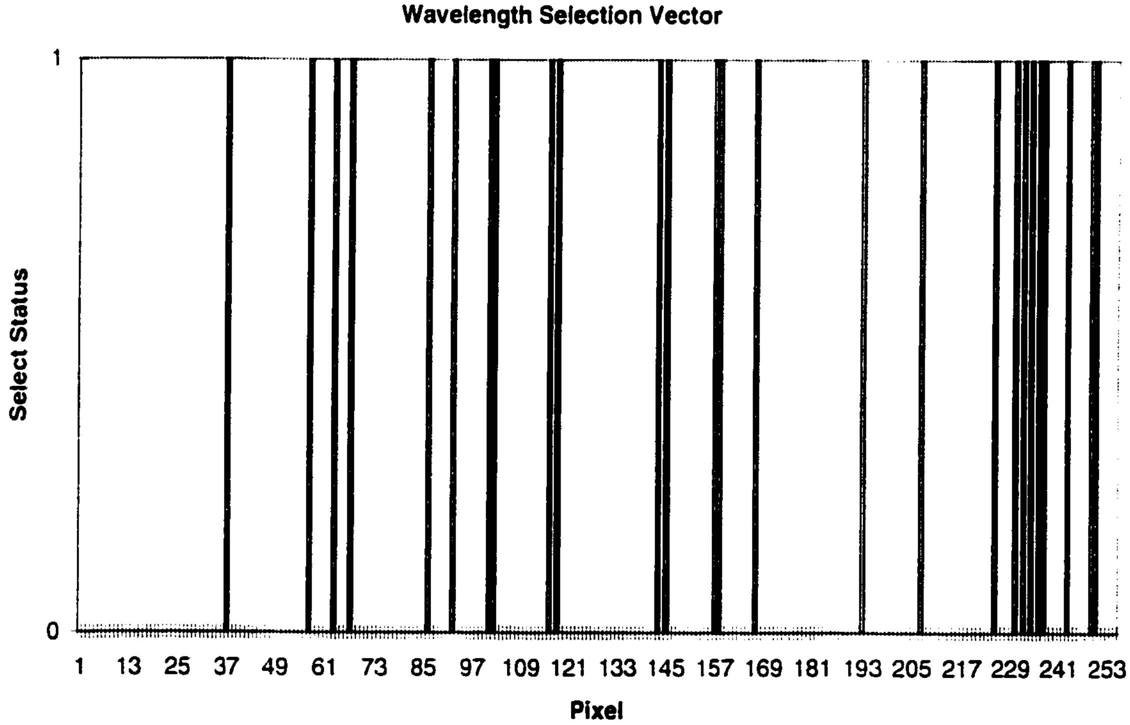


Figure 6.6: Wavelengths Selected for Data Set 1

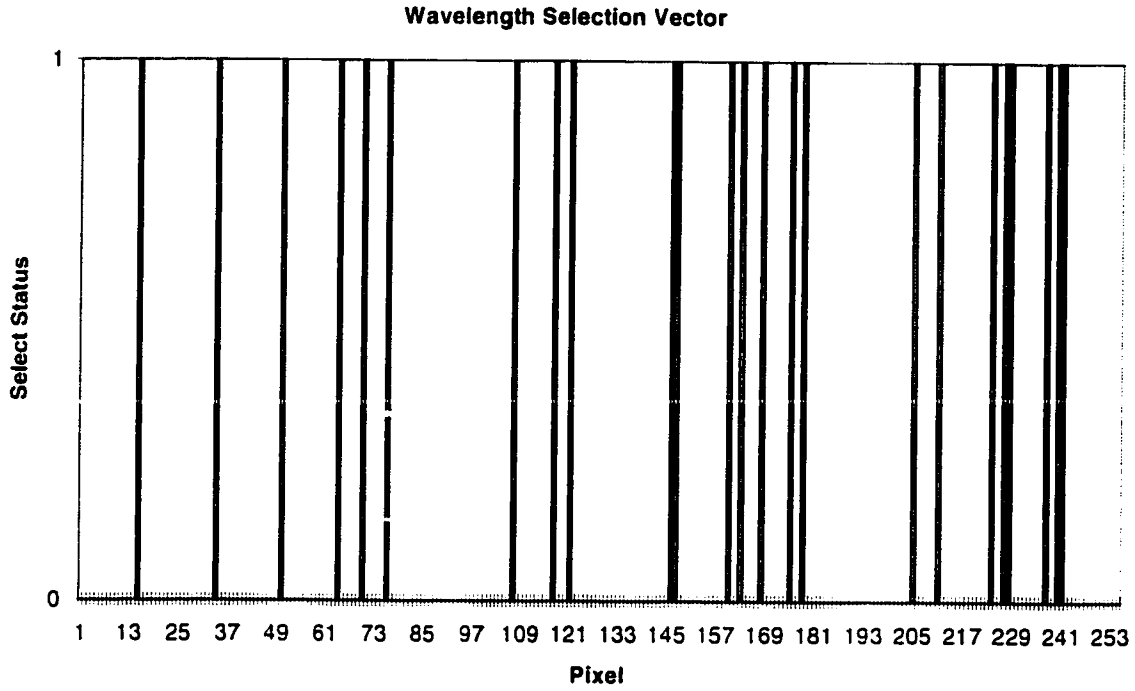


Figure 6.7: Wavelengths Selected for Data Set 2

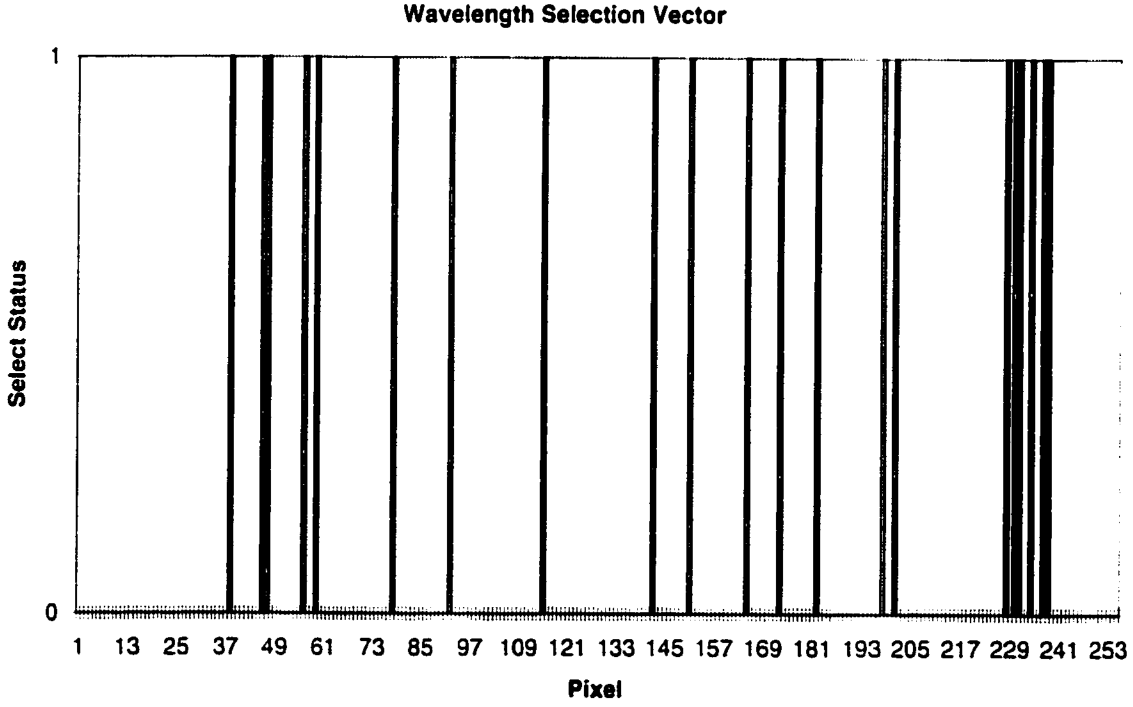


Figure 6.8: Wavelengths Selected for Data Set 3

6.3.3 Summary of Results in Classifying Urine from Blood

As mentioned previously, the data were analysed using MD and GA-MD using the transform SM11-2D11, no autoscaling, and PC's, for classifying urine from blood. The summary of the results is tabulated in Table 6.18.

Table 6.18 Summary of Results for Classifying Urine Versus Blood

	Prediction Error Rate (%)	Urine Prediction Rate (%)	Blood Prediction Rate (%)	Number of Wavelengths	Number of PC's
MD	0.95	95.00	99.72	255	9.0
GA-MD	0.71	95.00	100.00	23.7	10.3

These results indicate that urine and blood can be distinguished from each other at the 95% predictability level using less information than that determined when all species were classified simultaneously. The required number of features dropped from 97.7 wavelengths and 20 PC's, to 23.7 wavelengths and 10.3 PC's, without any loss in predictabilities. Therefore, the first objective has been achieved more efficiently, by optimising the classification for urine and blood species only.

Chapter 7

Conclusions

7.1 Were the Objectives Achieved?

The definition of success for each of the methodologies adopted in this study, is to achieve one or more of the objectives as stated in Section 1.3. Briefly, the first objective was to distinguish urine from blood; the second objective was to classify serum from plasma; and the third objective was to classify the five anticoagulants from each other; each at a classification accuracy of 95% or better.

The first objective was achieved relatively easily. Mahalanobis Distance and GA optimised MD were successful in achieving 95% predictability for classifying urine from blood. KNN was nearly successful in classifying urine from blood with predictabilities of 93.3% for urine and greater than 95% for blood. Therefore, the first objective, as outlined in Section 1.3, was achieved.

The second objective was not achieved, as the best predictability for serum was approximately 60%.

The third objective was being approached for oxalate and iodoacetate, using MD and GA-MD , with predictabilities of 85.0% and 81.7% respectively for the best case. The predictabilities of the other anticoagulants ranged between 60% and 72% for the best case.

Table 7.1 summarises the predictabilities for each of the species using the method that achieved the best results; i.e., MD with transform SM11-2D11, PC's, and no autoscaling.

Table 7.1: Prediction Performance using Best Method

Specie	1	2	3	4	5	6	7	Mean Pred'n Error Rate (%)
Mean Predictability (%)	96.7	58.3	71.7	66.7	85.0	81.7	60.0	25.72

A small surprise was that oxalate produces such a good NIR response, since the possible vibrational overtones are not listed in a common NIR response table.

7.2 Comparison of the Methodologies

The three methods used--KNN, MD and GA-MD--were compared for classifying all species simultaneously. Only MD and GA-MD were used to classify urine from blood. The classification performance results are summarised in Table 7.2. The data transform used was SM11-2D11, PC's, and autoscaling.

Table 7.2: Comparison of Mean Prediction Error Rates for KNN, MD, and GA-MD under the Same Conditions

Method	KNN	MD	GA-MD
Prediction Error Rate (%)	38.33	30.47	28.33

MD reduced the prediction error rate by 20.5% over that produced by KNN. GA-MD improved the prediction error rate by a further 7.0%. These are statistically significant improvements. These improvements are conservative estimates, as MD performed better without autoscaling. Unfortunately, this study did not have a proper comparison for the GA-MD case without autoscaling as the fitness criterion was not optimised for that case. The data, however, did show evidence that an improvement would be achieved for GA-MD without autoscaling, using an optimised fitness criterion determined for the no autoscaling case. This is the reason why the best results, overall, were achieved with MD without autoscaling, as shown in Table 7.1.

Table 7.3 shows how efficient MD and GA-MD performed with respect to the amount of information required. All the results in Table 7.3, except for GA-MD in classifying all species, did not use autoscaling. GA-MD required fewer wavelengths and fewer PC's than for MD (while achieving a 7.0% performance improvement); i.e., mean of 97.7 wavelengths and 20 PC's for GA-MD, and 255 wavelengths and 25 PC's for MD, for classifying all species. In classifying urine from blood, GA-MD, even more significantly, reduced the number of wavelengths over MD; from 255 to 23.7 with only a small increase in the number of PC's, i.e., from 9 to 10.3. There was no significant difference in predictability for the urine versus blood case, between that achieved by MD and that achieved by GA-MD.

Table 7.3: Number of Features Required to Achieve Optimum Results

	Classifying	Number of Wavelengths	Number of PC's
MD	All Species	255	25
GA-MD	All Species	97.7	20
MD	Urine from Blood	255	9
GA-MD	Urine from Blood	23.7	10.3

Therefore, GA optimisation of MD is able to achieve the same or better performance than MD alone, while reducing the amount of information required to achieve this.

Future Work

A number of items that can be considered for future investigational work are identified briefly in the following discussions.

The best GA-MD fitness criterion for the no autoscaling case could be determined, and complete analyses for this case could be performed. It is expected this would achieve even better results for the "all species" case.

To enhance interpretation of the wavelengths selected, it may be useful to group the wavelengths into sets of 5 or 10 wavelengths, and then select or deselect groups. The expectation is that the homogeneity of the wavelengths (in this case groups of wavelengths) selected would increase between data sets, and between runs, thus enhancing interpretation.

To optimise the number of PC's and enhance predictability, the data could be re-analysed using a separate validation data set (or use cross-validation) in the training process.

As indicated by the data, there was significant correlation between the number of wavelengths selected and the predictability. It may prove useful to modify the genetic algorithm to search for the best wavelengths given a particular number of wavelengths. The data would be analysed using a variety of the number of wavelengths, in order to find the optimum number of wavelengths. The process would be similar to finding the optimum fitness criterion except that the process would find the optimum number of wavelengths.

It may prove enlightening to optimise the classification of each specie separately, that is, determine which wavelengths are important for each specie.

To achieve more of the objectives it may require more sensitive instrumentation or a different wavelength range. These could be investigated.

Appendix A

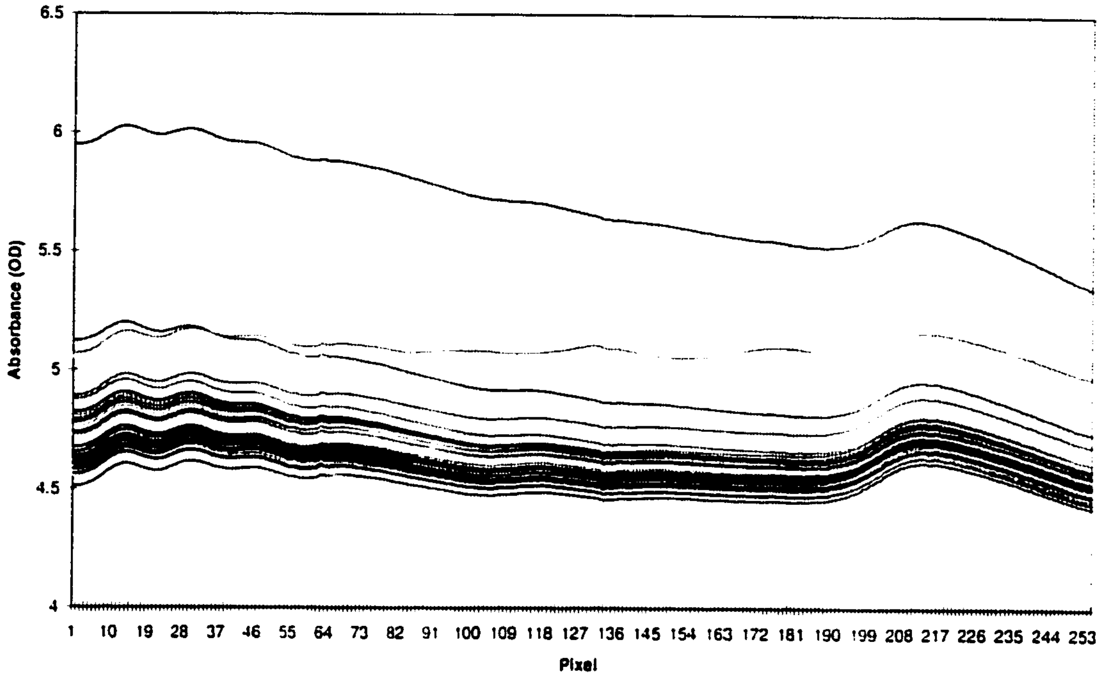
Spectrophotometer Wavelength Calibration

	Wavelength		Wavelength		Wavelength		Wavelength		Wavelength		Wavelength
Pixel	(nm)	Pixel	(nm)	Pixel	(nm)	Pixel	(nm)	Pixel	(nm)	Pixel	(nm)
1	602.83	51	689.31	101	775.79	151	862.27	201	948.75	251	1035.24
2	604.56	52	691.04	102	777.52	152	864.00	202	950.48	252	1036.97
3	606.29	53	692.77	103	779.25	153	865.73	203	952.21	253	1038.69
4	608.01	54	694.50	104	780.98	154	867.46	204	953.94	254	1040.42
5	609.74	55	696.23	105	782.71	155	869.19	205	955.67	255	1042.15
6	611.47	56	697.96	106	784.44	156	870.92	206	957.40		
7	613.20	57	699.69	107	786.17	157	872.65	207	959.13		
8	614.93	58	701.42	108	787.90	158	874.38	208	960.86		
9	616.66	59	703.14	109	789.63	159	876.11	209	962.59		
10	618.39	60	704.87	110	791.36	160	877.84	210	964.32		
11	620.12	61	706.60	111	793.09	161	879.57	211	966.05		
12	621.85	62	708.33	112	794.82	162	881.30	212	967.78		
13	623.58	63	710.06	113	796.55	163	883.03	213	969.51		
14	625.31	64	711.79	114	798.28	164	884.76	214	971.24		
15	627.04	65	713.52	115	800.00	165	886.49	215	972.97		
16	628.77	66	715.25	116	801.73	166	888.22	216	974.70		
17	630.50	67	716.98	117	803.46	167	889.95	217	976.43		
18	632.23	68	718.71	118	805.19	168	891.68	218	978.16		
19	633.96	69	720.44	119	806.92	169	893.41	219	979.89		
20	635.69	70	722.17	120	808.65	170	895.13	220	981.62		
21	637.42	71	723.90	121	810.38	171	896.86	221	983.35		
22	639.15	72	725.63	122	812.11	172	898.59	222	985.08		
23	640.88	73	727.36	123	813.84	173	900.32	223	986.81		
24	642.61	74	729.09	124	815.57	174	902.05	224	988.54		
25	644.34	75	730.82	125	817.30	175	903.78	225	990.27		
26	646.07	76	732.55	126	819.03	176	905.51	226	991.99		
27	647.80	77	734.28	127	820.76	177	907.24	227	993.72		
28	649.53	78	736.01	128	822.49	178	908.97	228	995.45		
29	651.26	79	737.74	129	824.22	179	910.70	229	997.18		
30	652.99	80	739.47	130	825.95	180	912.43	230	998.91		
31	654.72	81	741.20	131	827.68	181	914.16	231	1000.64		
32	656.44	82	742.93	132	829.41	182	915.89	232	1002.37		
33	658.17	83	744.66	133	831.14	183	917.62	233	1004.10		
34	659.90	84	746.39	134	832.87	184	919.35	234	1005.83		
35	661.63	85	748.12	135	834.60	185	921.08	235	1007.56		
36	663.36	86	749.85	136	836.33	186	922.81	236	1009.29		
37	665.09	87	751.57	137	838.06	187	924.54	237	1011.02		
38	666.82	88	753.30	138	839.79	188	926.27	238	1012.75		
39	668.55	89	755.03	139	841.52	189	928.00	239	1014.48		
40	670.28	90	756.76	140	843.25	190	929.73	240	1016.21		
41	672.01	91	758.49	141	844.98	191	931.46	241	1017.94		
42	673.74	92	760.22	142	846.70	192	933.19	242	1019.67		
43	675.47	93	761.95	143	848.43	193	934.92	243	1021.40		
44	677.20	94	763.68	144	850.16	194	936.65	244	1023.13		
45	678.93	95	765.41	145	851.89	195	938.38	245	1024.86		
46	680.66	96	767.14	146	853.62	196	940.11	246	1026.59		
47	682.39	97	768.87	147	855.35	197	941.84	247	1028.32		
48	684.12	98	770.60	148	857.08	198	943.56	248	1030.05		
49	685.85	99	772.33	149	858.81	199	945.29	249	1031.78		
50	687.58	100	774.06	150	860.54	200	947.02	250	1033.51		

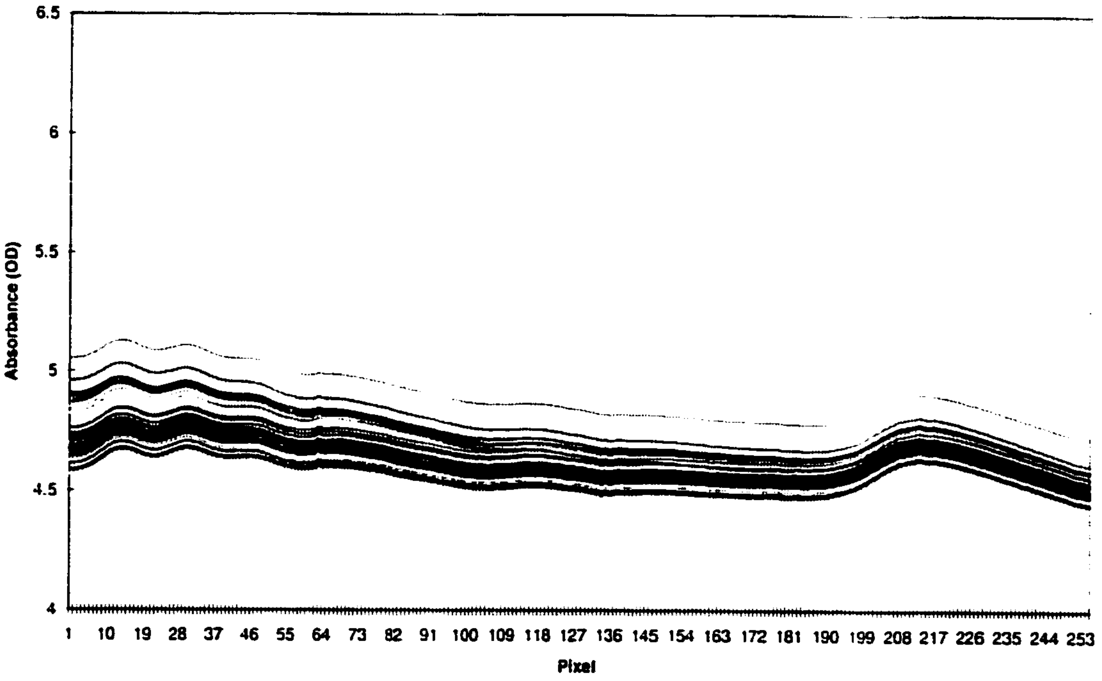
Appendix B

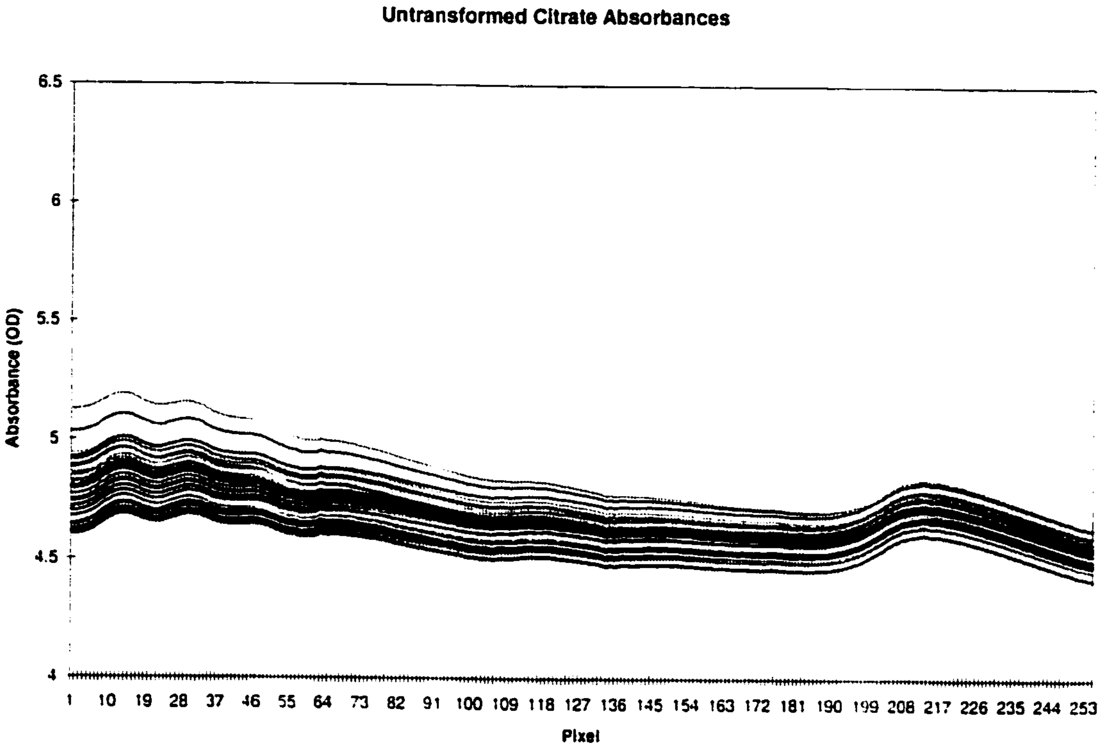
Untransformed Data

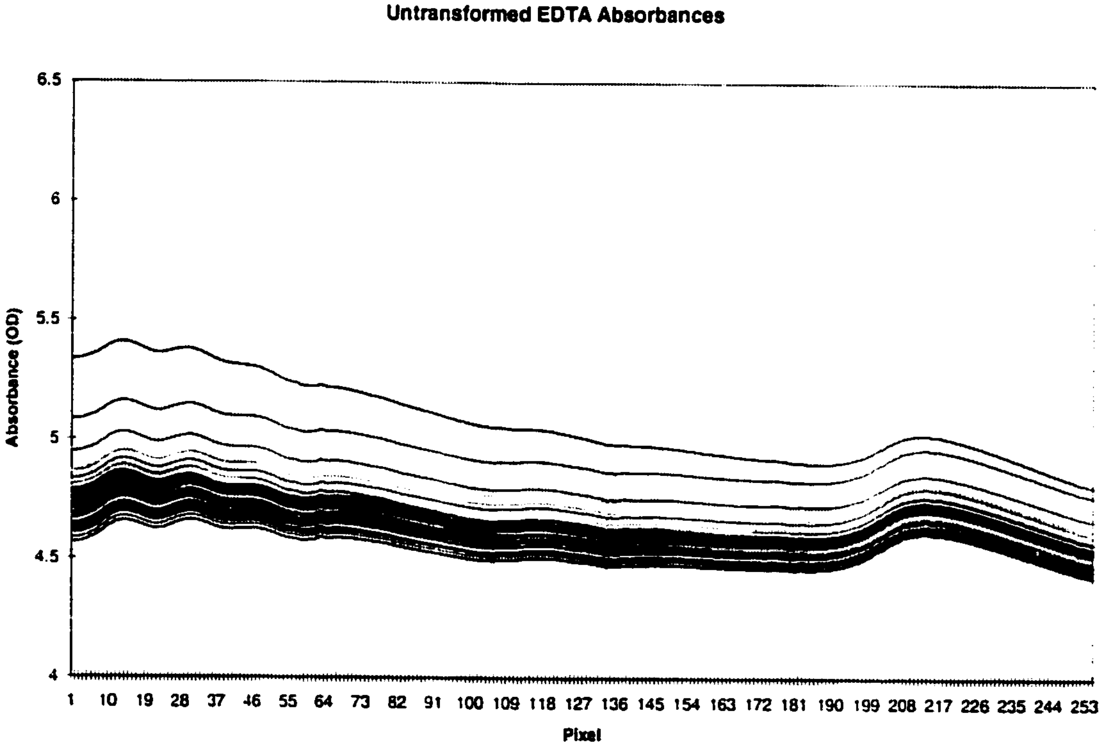
Untransformed Urine Absorbances



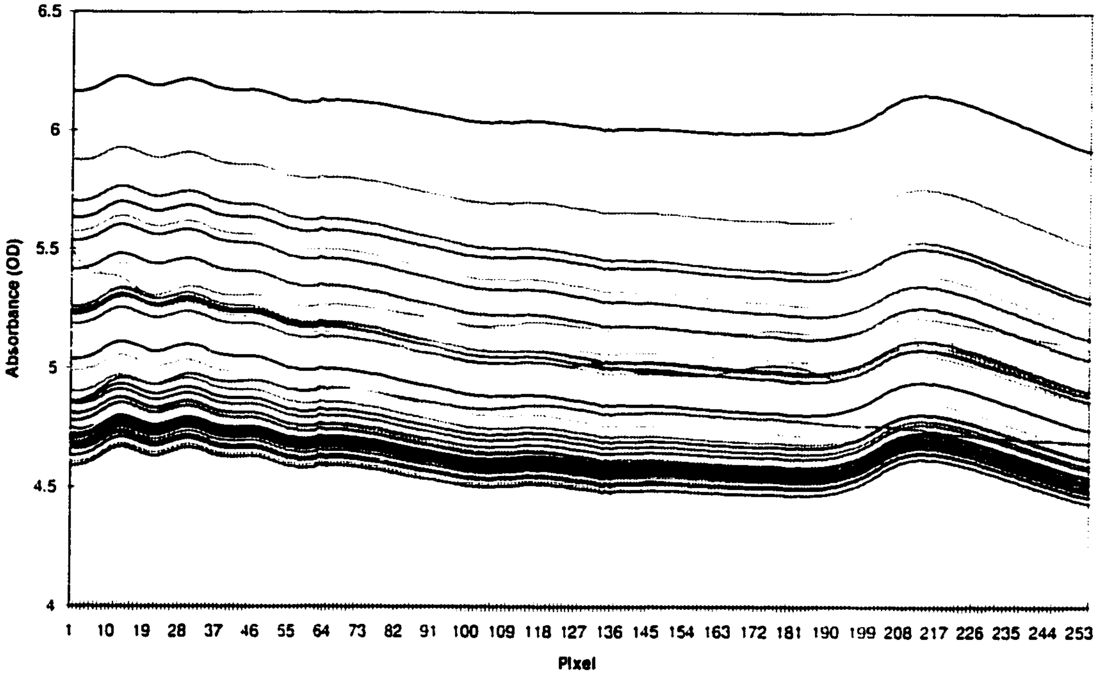
Untransformed Serum Absorbances



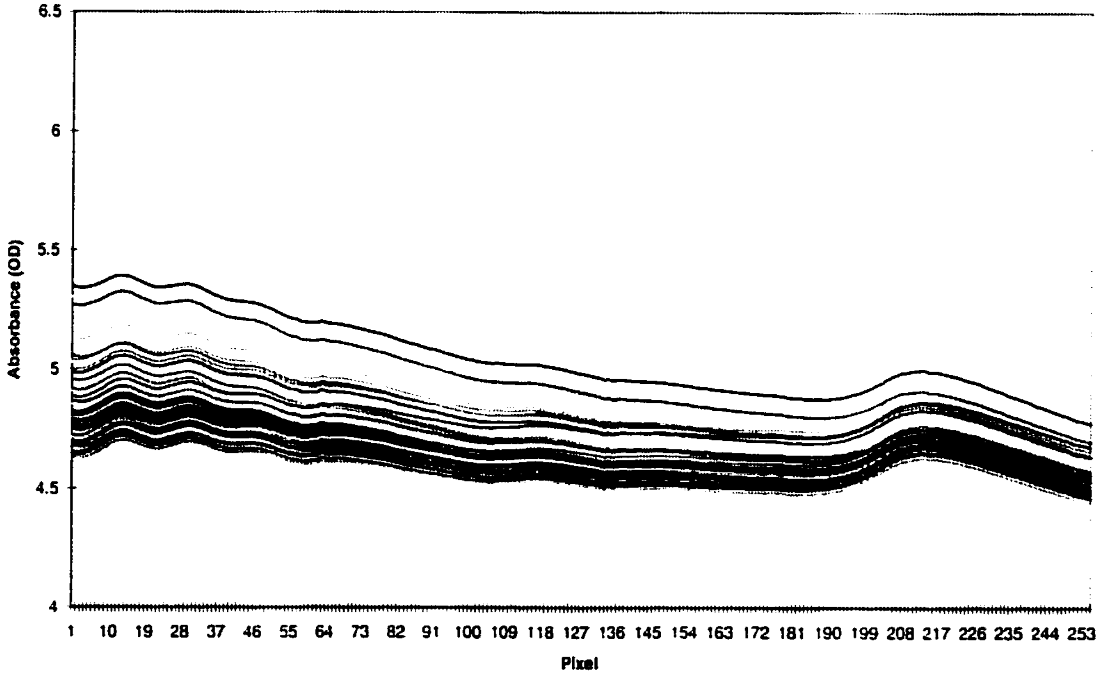

Untransformed Citrate Absorbances
Absorbance (OD)
6.5
6
5.5
5
4.5
4
1 10 19 28 37 46 55 64 73 82 91 100 109 118 127 136 145 154 163 172 181 190 199 208 217 226 235 244 253
Pixel

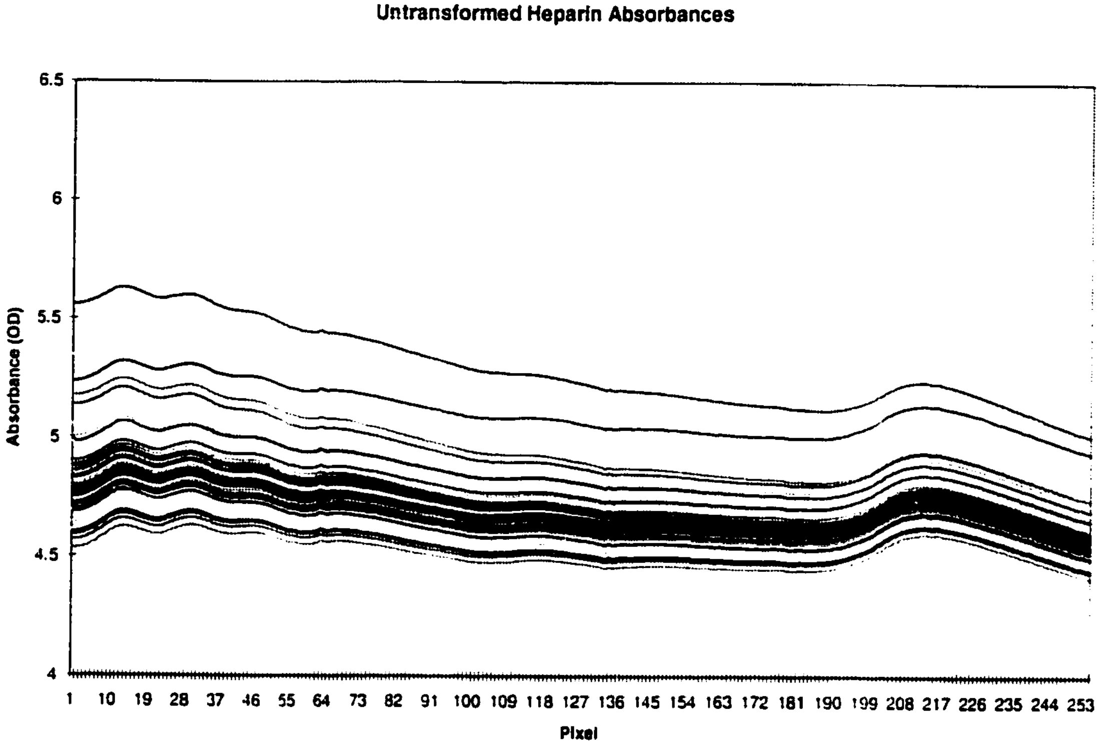

Untransformed EDTA Absorbances
Absorbance (OD)
6.5
6
5.5
5
4.5
4
1 10 19 28 37 46 55 64 73 82 91 100 109 118 127 136 145 154 163 172 181 190 199 208 217 226 235 244 253
Pixel

Untransformed Oxalate Absorbances



Untransformed Iodoacetate Absorbances

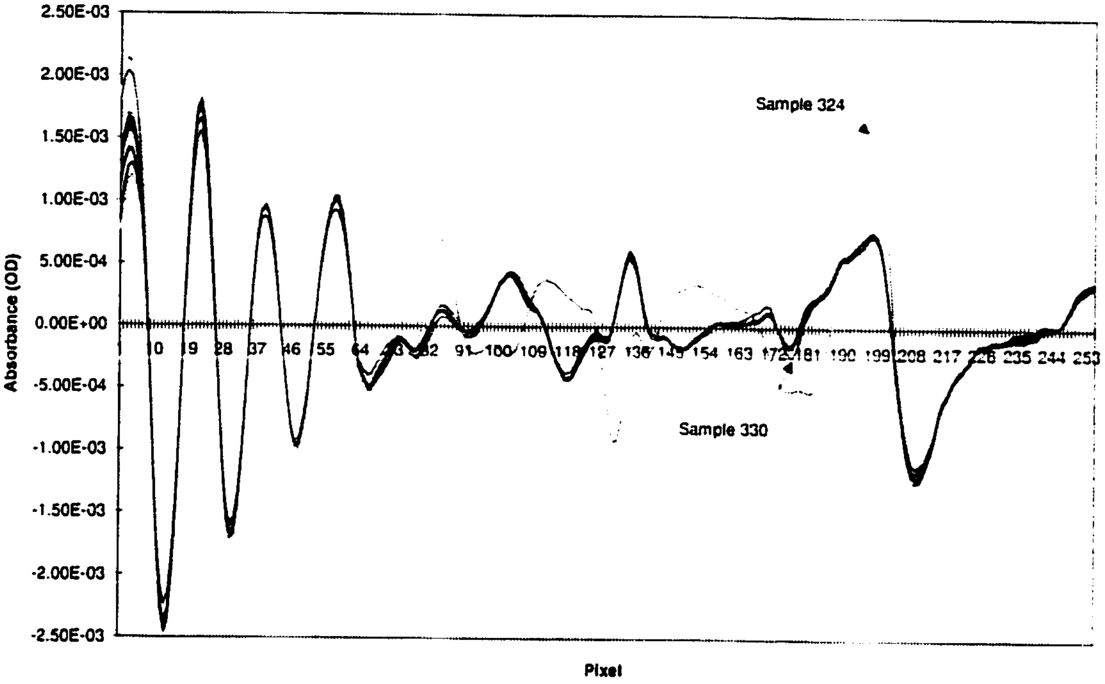



Untransformed Heparin Absorbances
Absorbance (OD)
6.5
6
5.5
5
4.5
4
1
10
19
28
37
46
55
64
73
82
91
100
109
118
127
136
145
154
163
172
181
190
199
208
217
226
235
244
253
Pixel

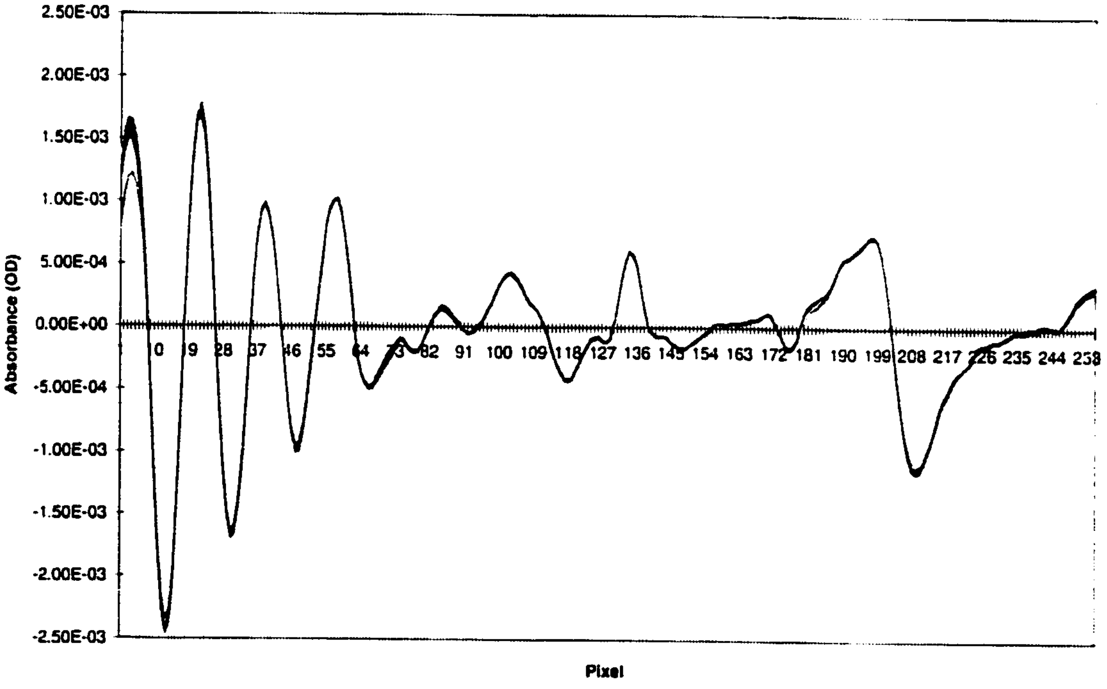
Appendix C

Transformed Data

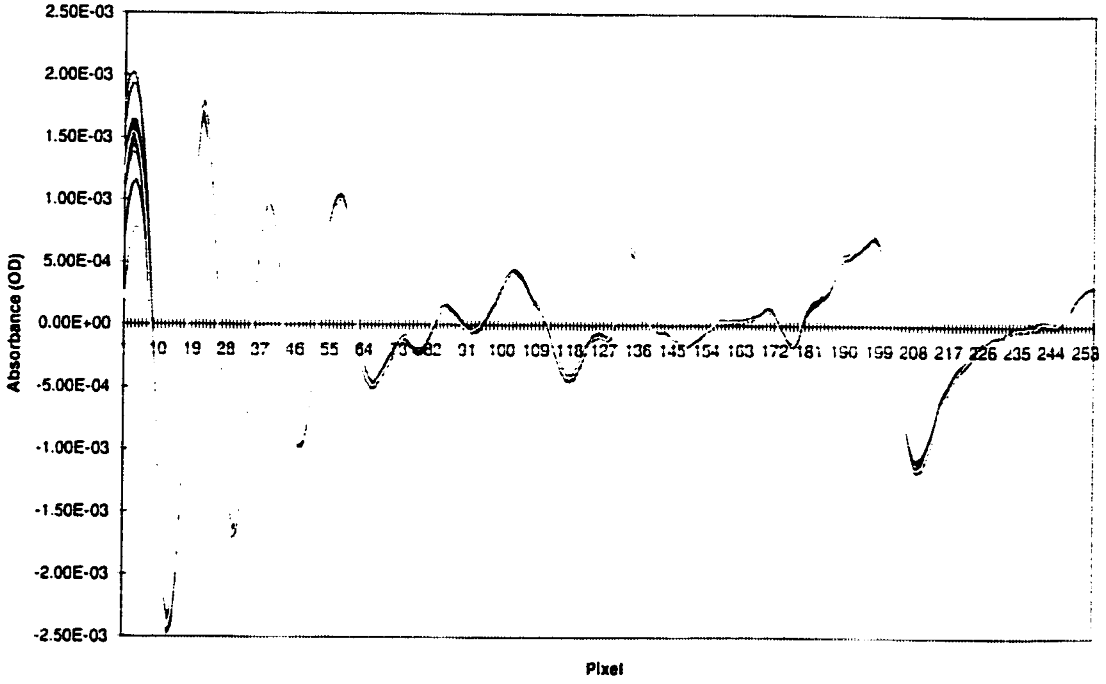
Transformed Urine Absorbances



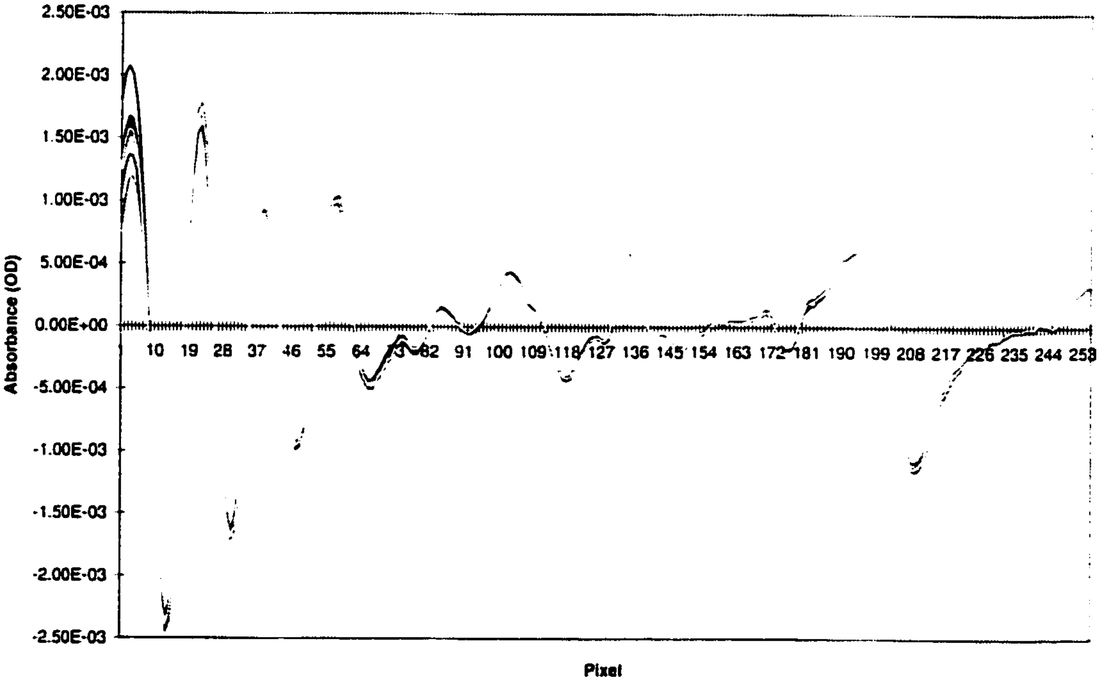
Transformed Serum Absorbances

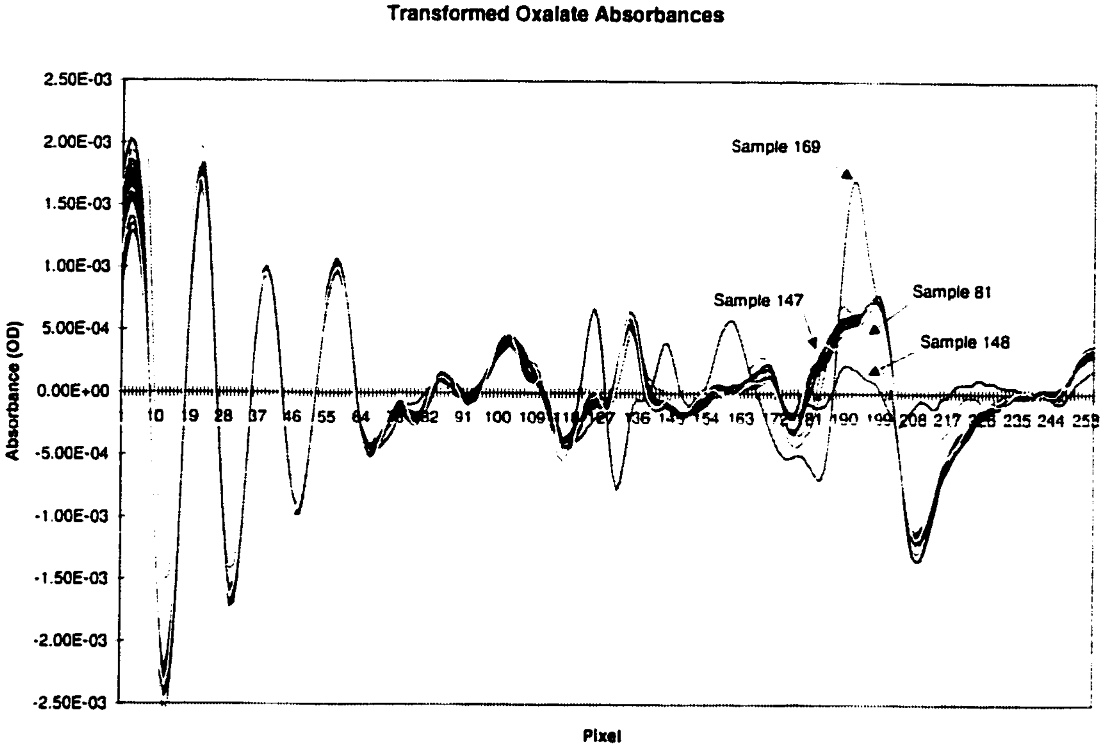


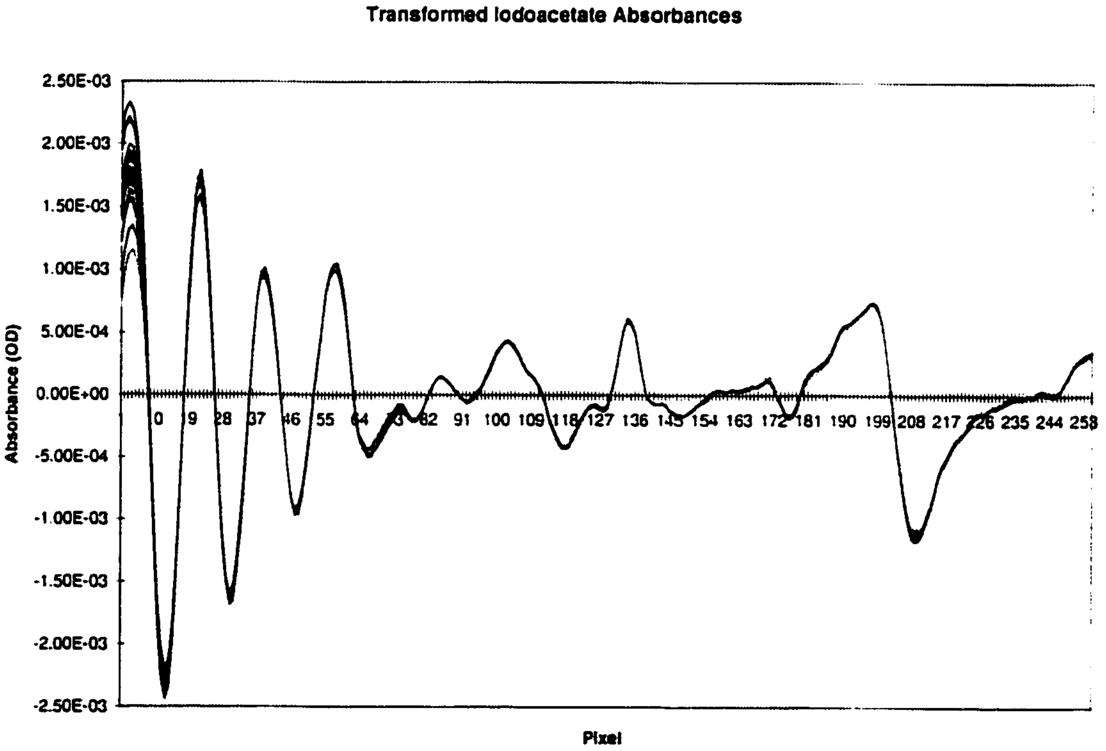
Transformed Citrate Absorbances

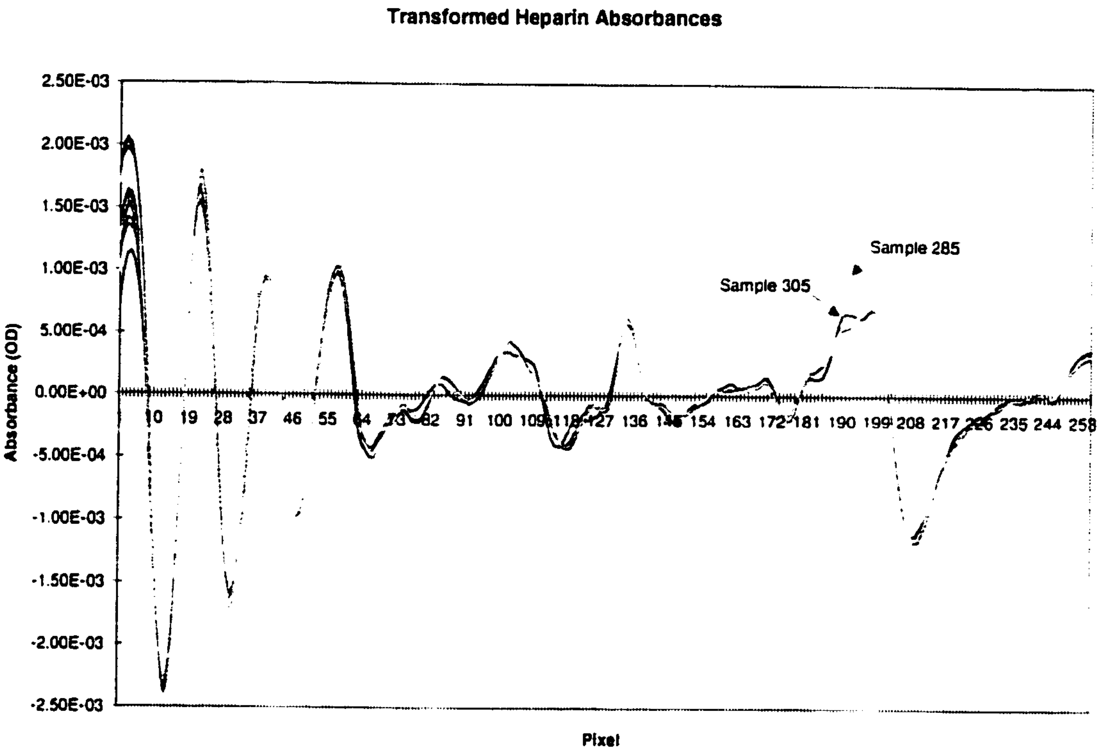


Transformed EDTA Absorbances



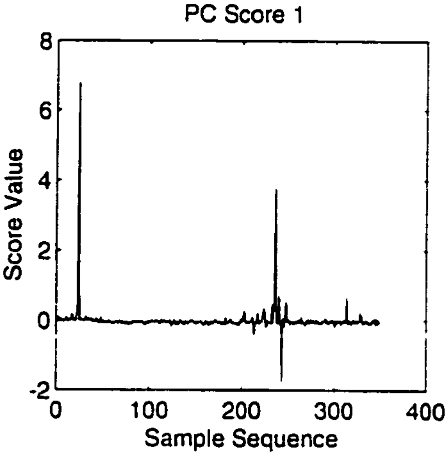

Transformed Oxalate Absorbances
Sample 169
Sample 147
Sample 81
Sample 148
2.50E-03
2.00E-03
1.50E-03
1.00E-03
5.00E-04
0.00E+00
-5.00E-04
-1.00E-03
-1.50E-03
-2.00E-03
-2.50E-03
Absorbance (OD)
1 10 19 28 37 46 55 64 73 82 91 100 109 118 127 136 145 154 163 172 181 190 199 208 217 226 235 244 253
Pixel

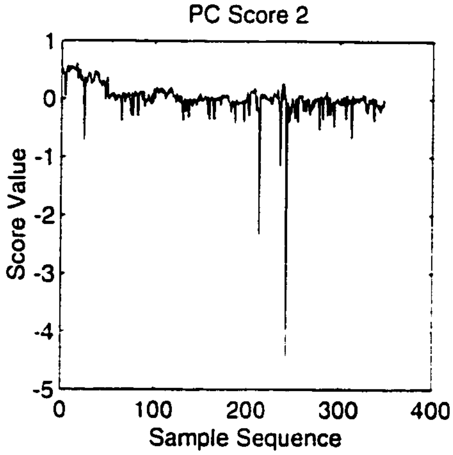

Transformed Iodoacetate Absorbances
2.50E-03
2.00E-03
1.50E-03
1.00E-03
5.00E-04
0.00E+00
-5.00E-04
-1.00E-03
-1.50E-03
-2.00E-03
-2.50E-03
Absorbance (OD)
10 19 28 37 46 55 64 73 82 91 100 109 118 127 136 145 154 163 172 181 190 199 208 217 226 235 244 253
Pixel

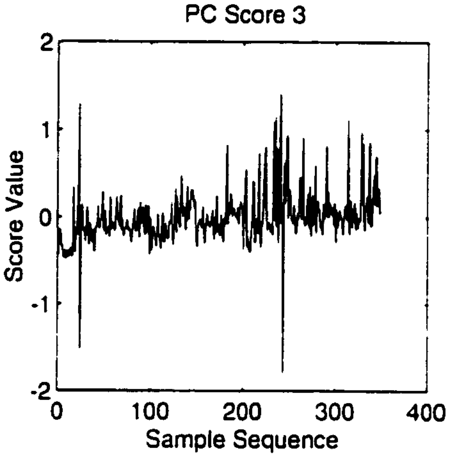

Transformed Heparin Absorbances
2.50E-03
2.00E-03
1.50E-03
1.00E-03
5.00E-04
0.00E+00
-5.00E-04
-1.00E-03
-1.50E-03
-2.00E-03
-2.50E-03
Absorbance (OD)
1 10 19 28 37 46 55 64 73 82 91 100 109 118 127 136 145 154 163 172 181 190 199 208 217 226 235 244 253
Pixel
Sample 285
Sample 305

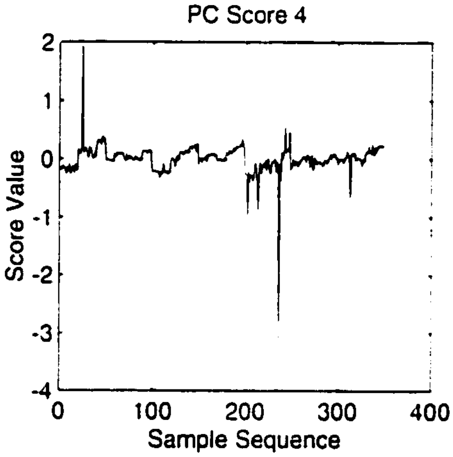
Appendix D

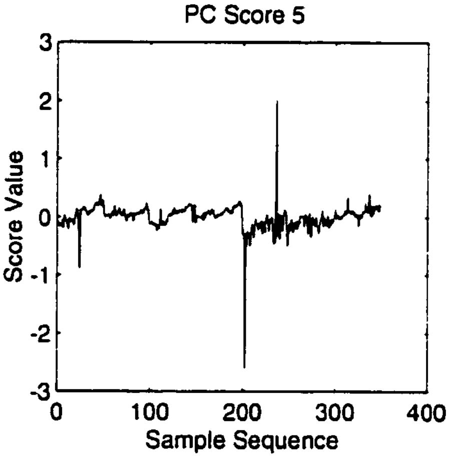
Principal Component Scores

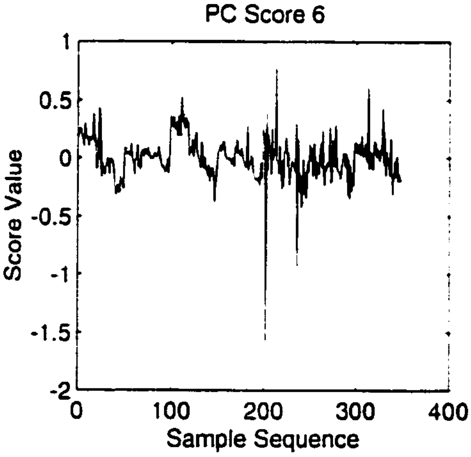

PC Score 1
Score Value
8
6
4
2
0
-2
0
100
200
300
400
Sample Sequence

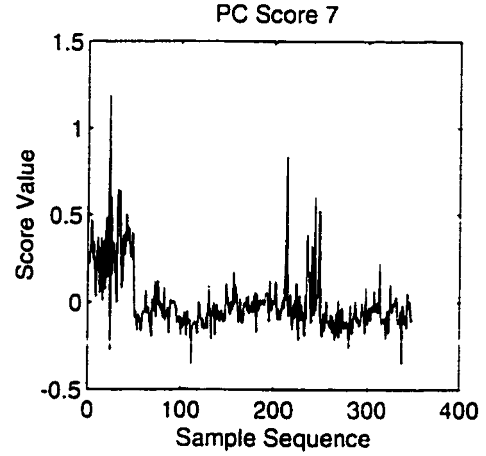

PC Score 2
Score Value
1
0
-1
-2
-3
-4
-5
0
100
200
300
400
Sample Sequence

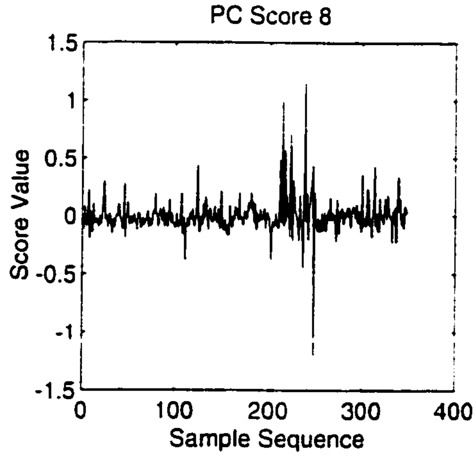

PC Score 3
Score Value
2
1
0
-1
-2
0
100
200
300
400
Sample Sequence

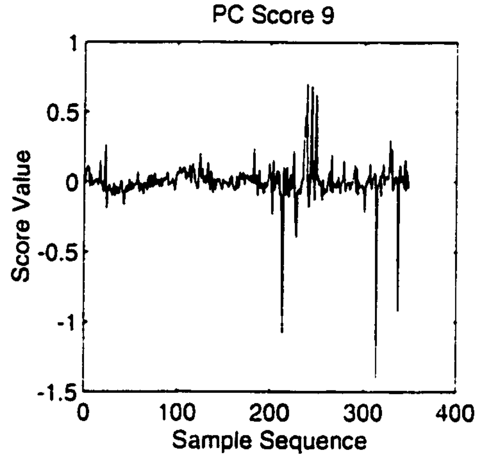

PC Score 4
Score Value
2
1
0
-1
-2
-3
-4
0
100
200
300
400
Sample Sequence

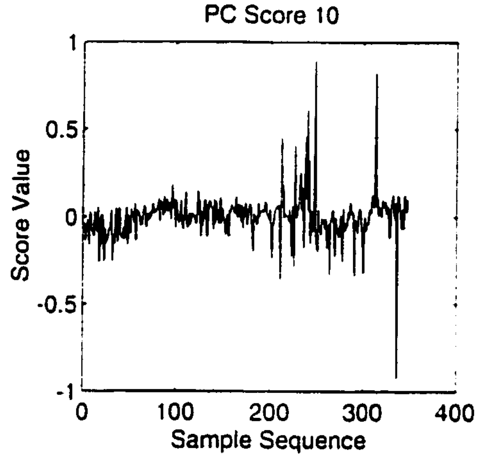

PC Score 5
Score Value
3
2
1
0
-1
-2
-3
0
100
200
300
400
Sample Sequence

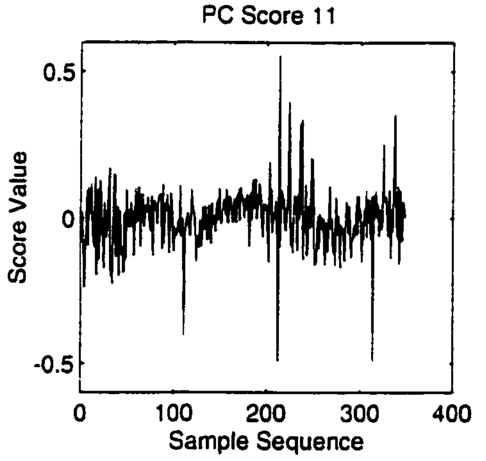

PC Score 6
Score Value
1
0.5
0
-0.5
-1
-1.5
-2
0
100
200
300
400
Sample Sequence

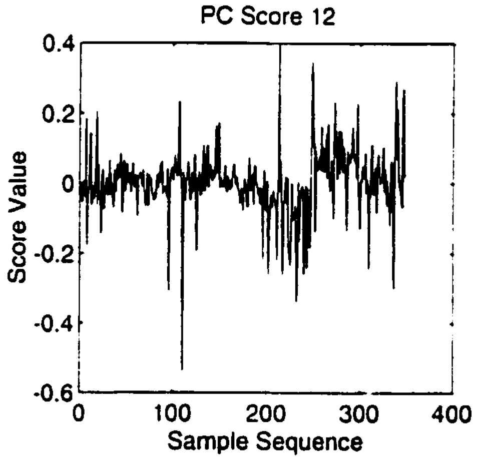

PC Score 7

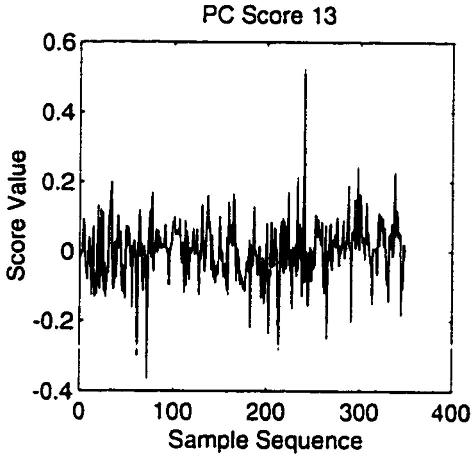

PC Score 8

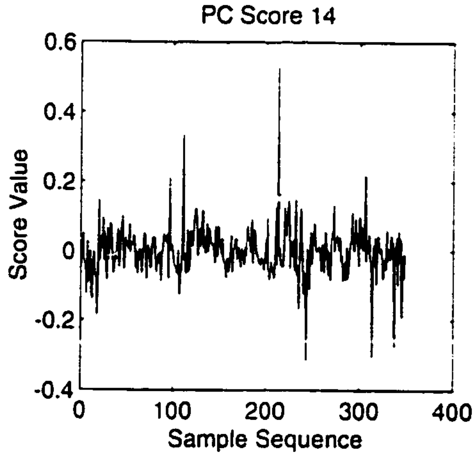

PC Score 9

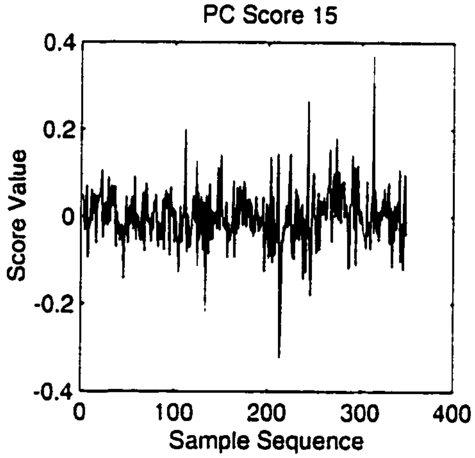

PC Score 10

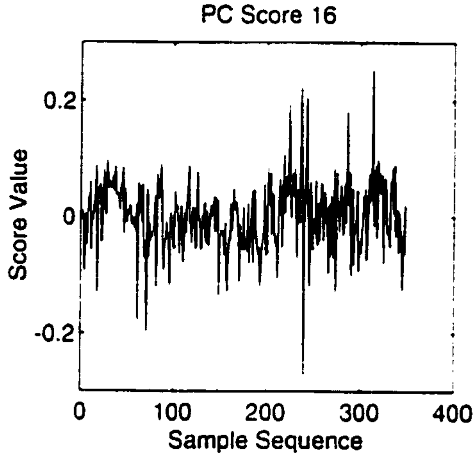

PC Score 11

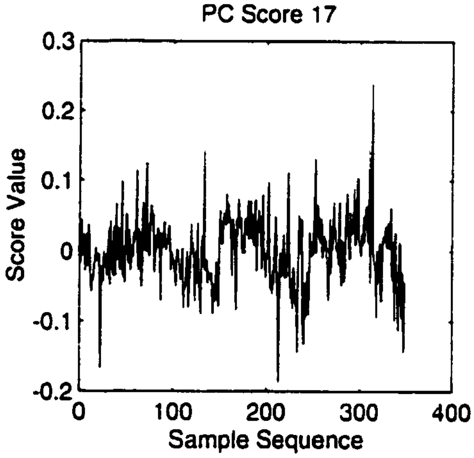

PC Score 12

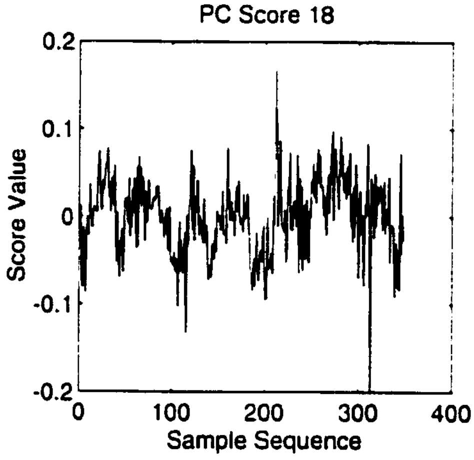

PC Score 13
0.6
0.4
0.2
0
-0.2
-0.4
Score Value
0
100
200
300
400
Sample Sequence

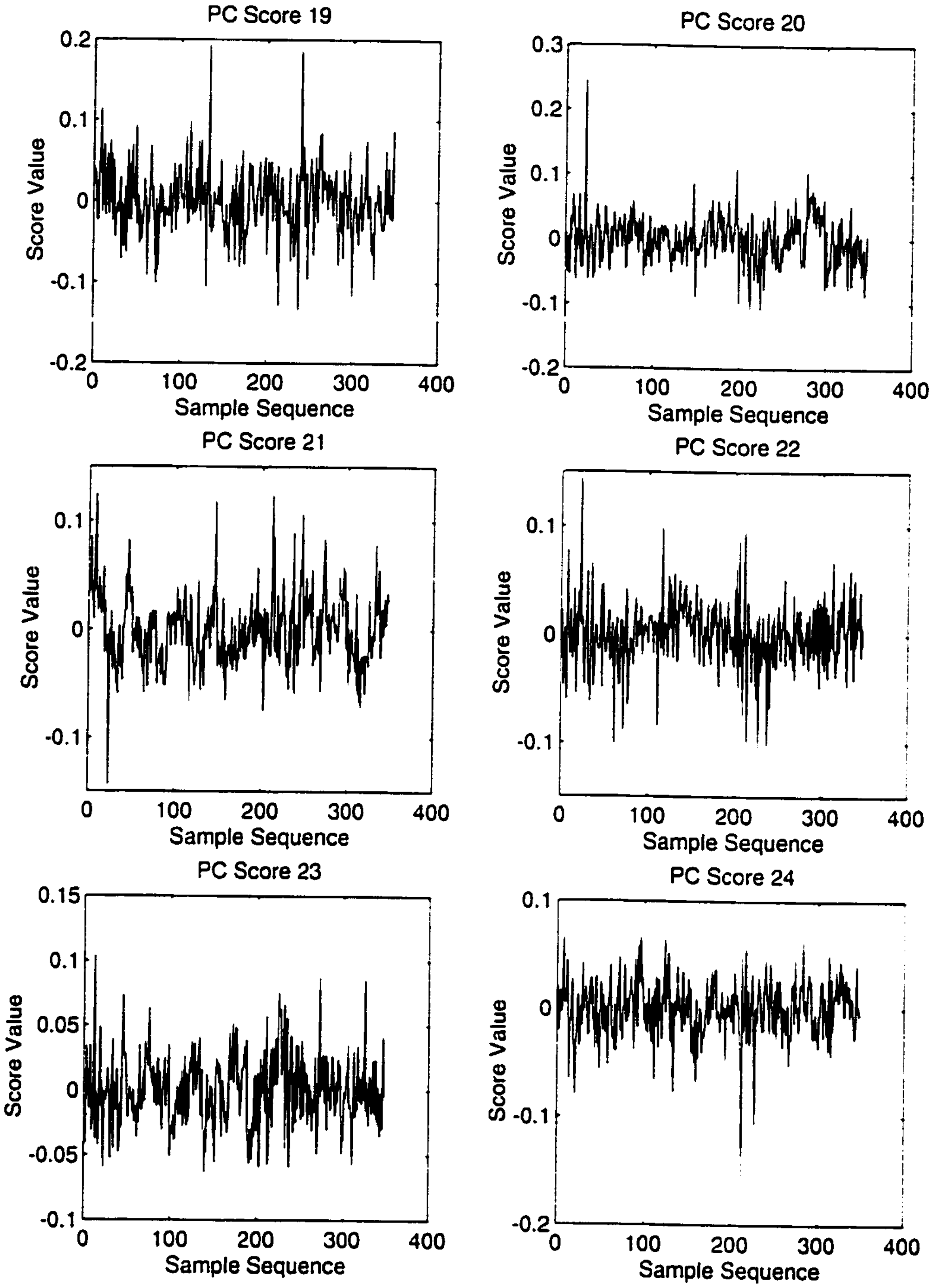

PC Score 14
0.6
0.4
0.2
0
-0.2
-0.4
Score Value
0
100
200
300
400
Sample Sequence


PC Score 15
0.4
0.2
0
-0.2
-0.4
Score Value
0
100
200
300
400
Sample Sequence


PC Score 16
0.2
0
-0.2
Score Value
0
100
200
300
400
Sample Sequence


PC Score 17
0.3
0.2
0.1
0
-0.1
-0.2
Score Value
0
100
200
300
400
Sample Sequence


PC Score 18
0.2
0.1
0
-0.1
-0.2
Score Value
0
100
200
300
400
Sample Sequence


PC Score 19
0.2
0.1
0
-0.1
-0.2
Score Value
0
100
200
300
400
Sample Sequence
PC Score 20
0.3
0.2
0.1
0
-0.1
-0.2
Score Value
0
100
200
300
400
Sample Sequence
PC Score 21
0.1
0
-0.1
Score Value
0
100
200
300
400
Sample Sequence
PC Score 22
0.1
0
-0.1
Score Value
0
100
200
300
400
Sample Sequence
PC Score 23
0.15
0.1
0.05
0
-0.05
-0.1
Score Value
0
100
200
300
400
Sample Sequence
PC Score 24
0.1
0
-0.1
-0.2
Score Value
0
100
200
300
400
Sample Sequence

Appendix E

Analysis Result Examples

Analysis Conditions

Classification Method	KNN	Data Processing	Second Derivative	11 pts
Data Set	1		S-G Smooth	11 pts
Number of Generations	N/A		Autoscale	Yes
Reference	Reslt07.xls		PCA	Yes

Analysis Results

Misclassification Matrix For Training Data

Actual	Predicted Specie								No. of
Specie	1	2	3	4	5	6	7	No. of Errors	Specimens
1	26	0	0	0	1	0	0	1	27
2	0	19	1	2	0	3	4	10	29
3	0	5	16	4	1	0	4	14	30
4	0	4	2	17	0	3	4	13	30
5	0	1	5	0	14	4	0	10	24
6	0	5	0	2	0	19	3	10	29
7	0	7	3	6	0	2	11	18	29
						Totals		76	198

Misclassification Rate Matrix For Training Data

Actual	Predicted Specie						
Specie	1	2	3	4	5	6	7
1	0.9630	0.0000	0.0000	0.0000	0.0370	0.0000	0.0000
2	0.0000	0.6552	0.0345	0.0690	0.0000	0.1034	0.1379
3	0.0000	0.1667	0.5333	0.1333	0.0333	0.0000	0.1333
4	0.0000	0.1333	0.0667	0.5667	0.0000	0.1000	0.1333
5	0.0000	0.0417	0.2083	0.0000	0.5833	0.1667	0.0000
6	0.0000	0.1724	0.0000	0.0690	0.0000	0.6552	0.1034
7	0.0000	0.2414	0.1034	0.2069	0.0000	0.0690	0.3793

Misclassification Matrix For Prediction Data

Actual	Predicted Specie								No. of
Specie	1	2	3	4	5	6	7	No. of Errors	Specimens
1	19	0	0	1	0	0	0	1	20
2	0	11	1	1	0	3	4	9	20
3	0	3	12	2	0	2	1	8	20
4	0	8	1	6	0	1	4	14	20
5	0	0	3	2	11	1	3	9	20
6	0	3	0	1	1	15	0	5	20
7	0	3	4	2	1	0	10	10	20
						Totals		56	140

Misclassification Rate Matrix For Prediction Data

Actual	Predicted Specie						
Specie	1	2	3	4	5	6	7
1	0.9500	0.0000	0.0000	0.0500	0.0000	0.0000	0.0000
2	0.0000	0.5500	0.0500	0.0500	0.0000	0.1500	0.2000
3	0.0000	0.1500	0.6000	0.1000	0.0000	0.1000	0.0500
4	0.0000	0.4000	0.0500	0.3000	0.0000	0.0500	0.2000
5	0.0000	0.0000	0.1500	0.1000	0.5500	0.0500	0.1500
6	0.0000	0.1500	0.0000	0.0500	0.0500	0.7500	0.0000
7	0.0000	0.1500	0.2000	0.1000	0.0500	0.0000	0.5000

Interspecie Mahalanobis Distances

Specie	Specie						
	1	2	3	4	5	6	7
1	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2	N/A	N/A	N/A	N/A	N/A	N/A	N/A
3	N/A	N/A	N/A	N/A	N/A	N/A	N/A
4	N/A	N/A	N/A	N/A	N/A	N/A	N/A
5	N/A	N/A	N/A	N/A	N/A	N/A	N/A
6	N/A	N/A	N/A	N/A	N/A	N/A	N/A
7	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Analysis Summary

Number of Wavelengths	255
Number of NNs	1
Number of PCs	26

Data Set	Error Rate
Train.	0.3838
Pred.	0.4000

Data	Predictability by Specie						
Set	1	2	3	4	5	6	7
Train.	0.963	0.655	0.533	0.567	0.583	0.655	0.379
Pred.	0.950	0.550	0.600	0.300	0.550	0.750	0.500

Analysis Conditions

Classification Method	Mahalanobis Distance	Data Processing	Second Derivative	11 pts
Data Set	1		S-G Smooth	11 pts
Number of Generations	NA		Autoscale	Yes
Reference	Reslt10T.xls		PCA	Yes

Analysis Results

Misclassification Matrix For Training Data

Actual Specie	Predicted Specie							No. of Errors	No. of Specimens
	1	2	3	4	5	6	7		
1	27	0	0	0	0	0	0	0	27
2	0	26	0	0	0	0	3	3	29
3	0	2	22	0	0	0	6	8	30
4	0	3	0	25	0	0	2	5	30
5	0	0	1	0	23	0	0	1	24
6	0	1	0	0	0	27	1	2	29
7	0	2	2	2	0	0	23	6	29
							Totals	25	198

Misclassification Rate Matrix For Training Data

Actual Specie	Predicted Specie						
	1	2	3	4	5	6	7
1	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0000	0.8966	0.0000	0.0000	0.0000	0.0000	0.1034
3	0.0000	0.0667	0.7333	0.0000	0.0000	0.0000	0.2000
4	0.0000	0.1000	0.0000	0.8333	0.0000	0.0000	0.0667
5	0.0000	0.0000	0.0417	0.0000	0.9583	0.0000	0.0000
6	0.0000	0.0345	0.0000	0.0000	0.0000	0.9310	0.0345
7	0.0000	0.0690	0.0690	0.0690	0.0000	0.0000	0.7931

Misclassification Matrix For Prediction Data

Actual Specie	Predicted Specie							No. of Errors	No. of Specimens
	1	2	3	4	5	6	7		
1	19	1	0	0	0	0	0	1	20
2	0	12	2	1	0	2	3	8	20
3	0	1	14	3	0	1	1	6	20
4	0	9	1	6	0	0	4	14	20
5	0	0	0	0	19	1	0	1	20
6	0	3	0	0	0	16	1	4	20
7	0	3	2	4	0	0	11	9	20
							Totals	43	140

Misclassification Rate Matrix For Prediction Data

Actual Specie	Predicted Specie						
	1	2	3	4	5	6	7
1	0.9500	0.0500	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0000	0.6000	0.1000	0.0500	0.0000	0.1000	0.1500
3	0.0000	0.0500	0.7000	0.1500	0.0000	0.0500	0.0500
4	0.0000	0.4500	0.0500	0.3000	0.0000	0.0000	0.2000
5	0.0000	0.0000	0.0000	0.0000	0.9500	0.0500	0.0000
6	0.0000	0.1500	0.0000	0.0000	0.0000	0.8000	0.0500
7	0.0000	0.1500	0.1000	0.2000	0.0000	0.0000	0.5500

Interspecie Mahalanobis Distances

Specie	Specie						
	1	2	3	4	5	6	7
1	0.00	11.02	11.18	11.11	11.94	12.86	11.79
2	-	0.00	2.78	2.23	5.04	4.32	2.03
3	-	-	0.00	2.89	4.93	5.10	2.16
4	-	-	-	0.00	5.70	4.85	2.46
5	-	-	-	-	0.00	5.68	4.99
6	-	-	-	-	-	0.00	5.52
7	-	-	-	-	-	-	0.00

Analysis Summary

Number of Wavelengths	255
Number of PCs	26
Optimized for Training Data Predictability	

Data Set	Error Rate
Train.	0.1263
Pred.	0.3071

Data Set	Predictability by Specie						
	1	2	3	4	5	6	7
Train.	1.000	0.897	0.733	0.833	0.958	0.931	0.793
Pred.	0.950	0.600	0.700	0.300	0.950	0.800	0.550

Analysis Conditions

Classification Method	G.A.-M.D.	Data Processing	Second Derivative	11 pts
Data Set	1		S-G Smooth	11 pts
Fitness Criterion	FTEALLN3		Autoscale	Yes
Number of Generations	500		PCA	Yes
Reference	Reslt18T.xls			

Analysis Results

Misclassification Matrix For Training Data

Actual	Predicted Specie							No. of	No. of Spec-
Specie	1	2	3	4	5	6	7	Errors	imens
1	27	0	0	0	0	0	0	0	27
2	0	25	1	1	0	0	2	4	29
3	0	1	24	0	0	0	5	6	30
4	0	3	0	25	0	0	2	5	30
5	0	0	1	0	22	0	1	2	24
6	0	1	0	0	0	28	0	1	29
7	0	2	0	1	0	0	26	3	29
							Totals	21	198

Misclassification Rate Matrix For Training Data

Actual	Predicted Specie						
Specie	1	2	3	4	5	6	7
1	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0000	0.8621	0.0345	0.0345	0.0000	0.0000	0.0690
3	0.0000	0.0333	0.8000	0.0000	0.0000	0.0000	0.1667
4	0.0000	0.1000	0.0000	0.8333	0.0000	0.0000	0.0667
5	0.0000	0.0000	0.0417	0.0000	0.9167	0.0000	0.0417
6	0.0000	0.0345	0.0000	0.0000	0.0000	0.9655	0.0000
7	0.0000	0.0690	0.0000	0.0345	0.0000	0.0000	0.8966

Misclassification Matrix For Prediction Data

Actual	Predicted Specie							No. of	No. of Spec-
Specie	1	2	3	4	5	6	7	Errors	imens
1	19	1	0	0	0	0	0	1	20
2	0	9	3	3	0	2	3	11	20
3	0	2	13	4	0	1	0	7	20
4	0	4	1	14	0	0	1	6	20
5	1	0	0	0	18	1	0	2	20
6	0	2	0	1	0	17	0	3	20
7	0	3	2	4	0	0	11	9	20
							Totals	39	140

Misclassification Rate Matrix For Prediction Data

Actual	Predicted Specie						
Specie	1	2	3	4	5	6	7
1	0.9500	0.0500	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0000	0.4500	0.1500	0.1500	0.0000	0.1000	0.1500
3	0.0000	0.1000	0.6500	0.2000	0.0000	0.0500	0.0000
4	0.0000	0.2000	0.0500	0.7000	0.0000	0.0000	0.0500
5	0.0500	0.0000	0.0000	0.0000	0.9000	0.0500	0.0000
6	0.0000	0.1000	0.0000	0.0500	0.0000	0.8500	0.0000
7	0.0000	0.1500	0.1000	0.2000	0.0000	0.0000	0.5500

Interspecie Mahalanobis Distances

Specie	Specie						
	1	2	3	4	5	6	7
1	0.00	9.56	9.76	9.48	10.57	11.42	10.22
2	-	0.00	2.71	1.94	4.69	4.34	1.99
3	-	-	0.00	2.70	4.57	4.99	2.02
4	-	-	-	0.00	4.93	4.75	2.25
5	-	-	-	-	0.00	5.07	4.44
6	-	-	-	-	-	0.00	4.60
7	-	-	-	-	-	-	0.00

Analysis Summary

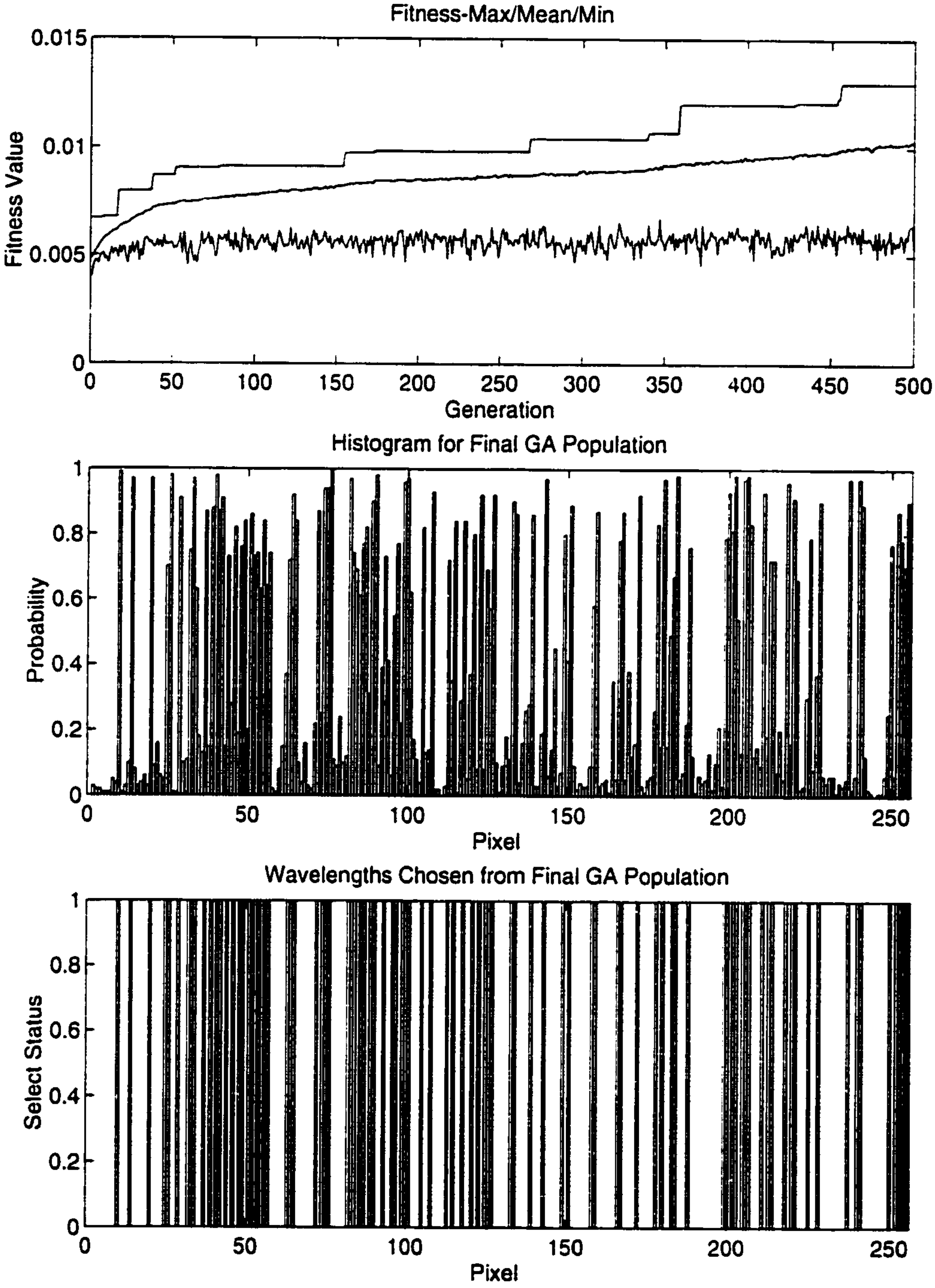
Number of Wavelengths	96
Number of PCs	20

Optimized for
Training Data Predictability

Data Set	Error Rate
Train.	0.1061
Pred.	0.2786

Data	Predictability by Specie						
Set	1	2	3	4	5	6	7
Train.	1.000	0.862	0.800	0.833	0.917	0.966	0.897
Pred.	0.950	0.450	0.650	0.700	0.900	0.850	0.550

RESULT18 (GA046)



Analysis Conditions

Classification Method	GA - MD		Data Processing	Second Derivative	11 pts
Data Set	1			S-G Smooth	11 pts
Fitness Criterion	FTE1N1			Autoscale	No
Number of Generations	500			PCA	Yes
Reference	Reslt33.xls (GA068)				

Analysis Results

Misclassification Matrix For Training Data

Actual Specie	Predicted Specie 1	2 to 7	No. of Errors	No. of Specimens
1	27	0	0	27
2	0	29	0	29
3	0	30	0	30
4	0	30	0	30
5	0	24	0	24
6	0	29	0	29
7	0	29	0	29
		Totals	0	198

Misclassification Rate Matrix For Training Data

Actual Specie	Predicted Specie 1	2 to 7
1	1.0000	0.0000
2	0.0000	1.0000
3	0.0000	1.0000
4	0.0000	1.0000
5	0.0000	1.0000
6	0.0000	1.0000
7	0.0000	1.0000

Misclassification Matrix For Prediction Data

Actual Specie	Predicted Specie 1	2 to 7	No. of Errors	No. of Specimens
1	19	1	1	20
2	0	20	0	20
3	0	20	0	20
4	0	20	0	20
5	0	20	0	20
6	0	20	0	20
7	0	20	0	20
		Totals	1	140

Misclassification Rate Matrix For Prediction Data

Actual Specie	Predicted Specie 1	2 to 7
1	0.9500	0.0500
2	0.0000	1.0000
3	0.0000	1.0000
4	0.0000	1.0000
5	0.0000	1.0000
6	0.0000	1.0000
7	0.0000	1.0000

Interspecie Mahalanobis Distances

Specie	Specie 1	2	3	4	5	6	7
1	0.00	8.25	8.50	8.17	8.32	9.34	8.55
2	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-
4	-	-	-	-	-	-	-
5	-	-	-	-	-	-	-
6	-	-	-	-	-	-	-
7	-	-	-	-	-	-	-

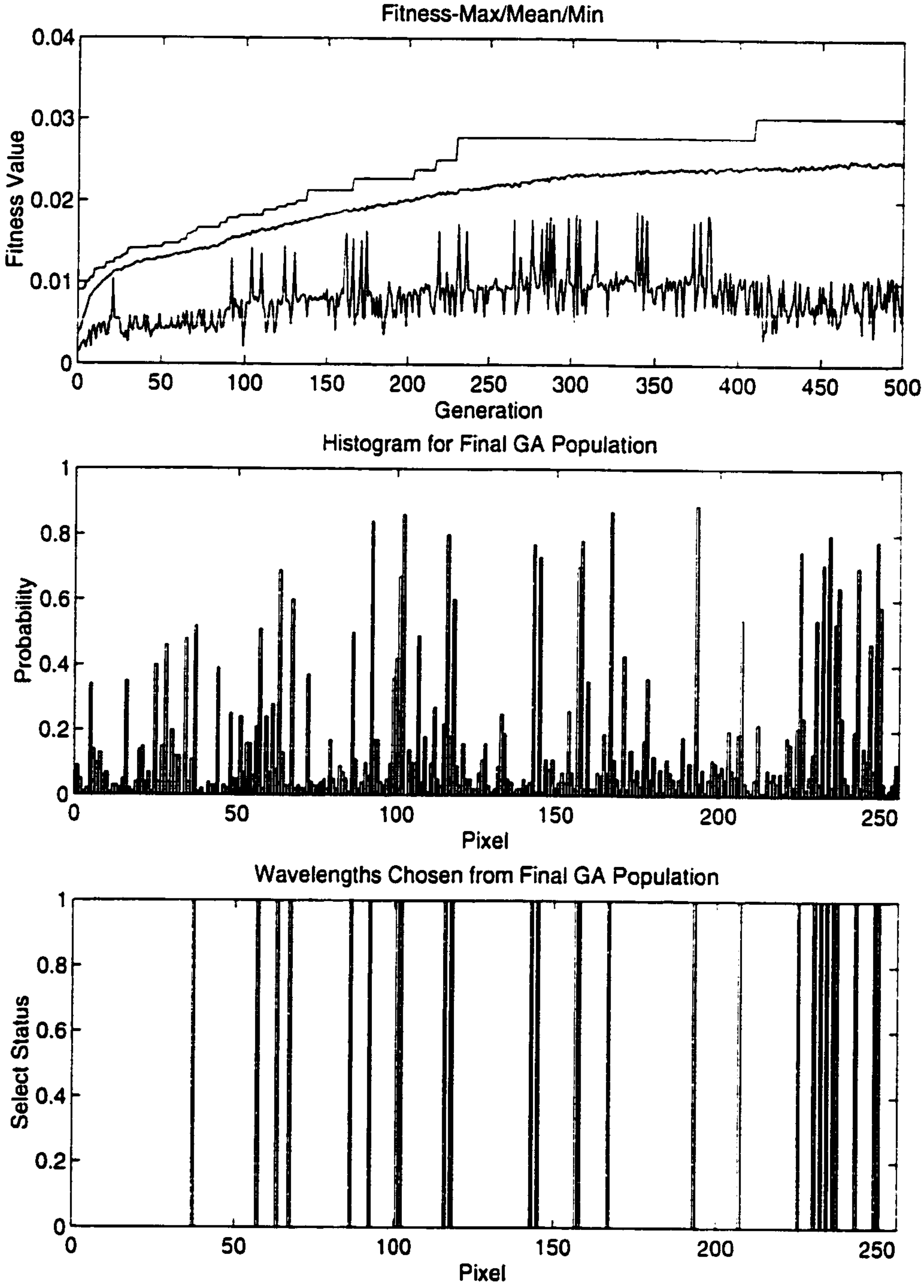
Analysis Summary

Number of Wavelengths	26
Number of PCs	9

Data Set	Error Rate
Train.	0.0000
Pred.	0.0071

Data Set	Predictability by Specie 1	2	3	4	5	6	7
Train.	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Pred.	0.950	1.000	1.000	1.000	1.000	1.000	1.000

GA068



Fitness-Max/Mean/Min

Histogram for Final GA Population

Wavelengths Chosen from Final GA Population

Bibliography

1. Southerts, J.; *How MDS Cut Lab Costs.* The Globe and Mail, Toronto, March 10 issue, p. B6, 1997.

2. Given, D.G. and Orr, K.H.; *CASS Operator's Manual.* CME Telemetrix Inc., Waterloo, Ontario, 1995.

3. Stark, E., Luchter, K. and Margoshes, M.; *Near-Infrared Analysis (NIRA): A Technology for Quantitative and Qualitative Analysis.* Applied Spectroscopy Reviews., 22(4), 335-399, 1986.

4. Halliday, D. and Resnick, R.; *Physics, Parts I and II Combined (Third Edition).* John Wiley & Sons, New York, pp. 1091-1116, 1978.

5. Banwell, C.N.; *Fundamentals of Molecular Spectroscopy (Third Edition).* McGraw-Hill, New York, 1983.

6. Kirsch, J.D. and Drennen, J.K.; *Near-Infrared Spectroscopy: Applications in the Analysis of Tablets and Solid Pharmaceutical Dosage Forms.* Applied Spectroscopy Reviews, 30(3), pp. 139-174, 1995.

7. Osborne, B.G., Fearn, T. and Hindle, P.H.; *Practical NIR Spectroscopy with Applications in Food and Beverage Analysis (Second Edition).* Longman Group, Essex, England, 1993.

8. Williams, P.C. and Norris, K. Editors; *Near-Infrared Technology in the Agricultural and Food Industries.* American Association of Cereal Chemists, Inc., St. Paul, Minnesota, 1987.

9. Shaw, R.A., Kotowich, S., Mantsch, H.H. and Leroux, M.; *Quantitation of Protein, Creatinine, and Urea in Urine by Near-Infrared Spectroscopy.* Clinical Biochemistry, Vol. 29, No. 1, pp. 11-19, 1996.

10. Hall, J.W. and Pollard, A.; *Near-Infrared Spectroscopic Determination of Serum Total Proteins, Albumin, Globulins, and Urea.* Clinical Biochemistry, Vol. 26, pp. 483-490, 1993.

11. Kuenstner, J.T., Norris, K.H., and McCarthy, W.F.; *Measurement of Hemoglobin in Unlysed Blood by Near-Infrared Spectroscopy.* Applied Spectroscopy, Vol. 48, No. 4, pp. 484-488, 1994.

12. Barnhart, R.K.; *The American Heritage Dictionary of Science.* Houghton Mifflin Company, Boston, Mass., 1986.

13. Guyton, A.C. and Hall, J.E.; *Textbook of Medical Physiology (Ninth Edition).* W.B. Saunders Company, Philadelphia, Pennsylvania, 1996.

14. Lehninger, A.L.; *Principles of Biochemistry.* Worth Publishers, Inc., New York, New York, 1982.

15. Budavari, S. Editor; *The Merck Index (Eleventh Edition).* Merck & Co., Inc., Rahway, New Jersey, 1989.

16. Given, D.G.; *CASS Instrumentation Principles (ER-145-001 Issue 1).* CME Telemetrix Inc., Waterloo, Ontario, 1995.

17. Martin, K.A.; *Recent Advances in Near-Infrared Reflectance Spectroscopy.* Applied Spectroscopy Reviews, 27(4), 325-383, 1992.

18. Geladi, P., MacDougall, D. and Martens, H.; *Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat.* Applied Spectroscopy, Vol. 39, No. 3, 1985.

19. Manly, B.F.J.; *Multivariate Statistical Methods - A Primer (Second Edition).* Chapman & Hall, London, UK, 1994.

20. Given, D.G.; *Biological Spectroscopy Using Statistical Pattern Recognition.* Course Paper for Statistical Pattern Recognition, University of Waterloo, Waterloo, Ontario, 1991.

21. Given, D.G.; *Information Theoretic Prediction Method for Spectroscopy.* Course Paper for Information Theory in Pattern Analysis and Synthesis, University of Waterloo, Waterloo, Ontario, 1990.

22. Fukunaga, K.; *Statistical Pattern Recognition (Second Edition).* Academic Press, San Diego, California, 1990.

23. Gemperline, P.J. and Shah, N.K.; *Combination of the Mahalanobis Distance and Residual Variance Pattern Recognition Techniques for Classification of Near-Infrared Reflectance Spectra.* Anal. Chem., 62:465-470, 1990.

24. Goldberg, D.E.; *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley, Reading, MA, 1989.

25. Lucasius, C.B. Jr.; *Towards Genetic Algorithm Methodology in Chemometrics.* PhD Thesis, Katholieke Universiteit Nijmegen, Netherlands, 1993.

26. Leardi, R.; *Application of a Genetic Algorithm to Feature Selection under Full Validation Conditions and to Outlier Detection.* Journal of Chemometrics, Vol. 8, 65-79, 1994.

27. Goldberg, D.E.; *Sizing Populations for Serial and Parallel Genetic Algorithms.* Proceedings of the Third International Conference on Genetic Algorithms, IEEE, San Mateo, California, pp. 70-79, 1989.

28. Savitzky, A. and Golay, M.J.E.; *Smoothing and Differentiation of Data by Simplified Least Squares Procedure.* Analytical Chemistry, Vol. 36, No. 8, pp. 1627-1638, 1964.

29. Steiner, J., Termonia, Y. and Deltour, J.; *Comments on Smoothing and Differentiation of Data by Simplified Least Square Procedure.* Analytical Chemistry, Vol. 44, No. 11, pp. 1906-1909, 1972.

30. Gorry, P.A.; *General Least-Squares Smoothing and Differentiation by the Convolution (Savitzky-Golay) Method.* Analytical Chemistry, Vol. 62, pp. 570-573, 1990.

31. Martens, H. and Naes, T.; *Multivariate Calibration.* John Wiley & Sons, New York, 1989.

32. Burgard, D.R. and Kuznicki, J.T.; *Chemometrics: Chemical and Sensory Data.* CRC Press, Boca Raton, Florida, 1990.

33. Mark, H.L. and Tunnel, D.; *Qualitative Near-Infrared Reflectance Analysis Using Mahalanobis Distances.* Analytical Chemistry, 57, 1449-1456, 1985.

34. Miller, I. and Freund, J.E.; *Probability and Statistics for Engineers.* Prentice-Hall, New Jersey, 1977.

35. Jouan-Rimbaud, D. and Massart, D.; *Genetic Algorithms as a Tool for Wavelength Selection in Multivariate Calibration.* Analytical Chemistry, Vol. 67, No. 23, pp. 4295-4301, 1995.

36. Xu, L. and Schechter, I.; *Wavelength Selection for Simultaneous Spectroscopic Analysis. Experimental and Theoretical Study.* Analytical Chemistry, Vol. 68, No. 14, pp. 2392-2400, 1996.