COMPARISON OF DATA CLASSIFICATION PROCEDURES IN APPLIED GEOCHEMISTRY USING MONTE CARLO SIMULATION

By

CLIFFORD R. STANLEY

A. B. (Earth Science) Dartmouth College, 1980

M. Sc. (Geological Sciences) University of British Columbia, 1984

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

GEOLOGICAL SCIENCES

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

October 1988

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Geological Sciences
The University of British Columbia
2075 Wesbrook Place
Vancouver, Canada
V6T 1W5

Date: October 11, 1988

Abstract

In geochemical applications, data classification commonly involves 'mapping' continuous variables into discrete descriptive categories, and often is achieved using thresholds to define specific ranges of data as separate groups which then can be compared with other categorical variables. This study compares several classification methods used in applied geochemistry to select thresholds and discriminate between populations or to recognize anomalous observations. The comparisons were made using *monte carlo* simulation to evaluate how well different techniques perform using different data set structures.

A comparison of maximum likelihood parameter estimates of a mixture of normal distributions using class interval frequencies *versus* raw data was undertaken to study the quality of the corresponding results. The more time consuming raw data approach produces optimal parameter estimates while the more rapid class interval approach is the approach in common use. Results show that provided there are greater than 50 observations per distribution and (on average) 10 observations per class interval, the maximum likelihood parameter estimates by the two methods are practically indistinguishable.

Univariate classification techniques evaluated in this study include the 'mean plus 2 standard deviations', the '95^{th} percentile', the gap statistic and probability plots. Results show that the 'mean plus 2 standard deviations' and '95^{th} percentile' approaches are inappropriate for most geochemical data sets. The probability plot technique classifies mixtures of normal distributions better than the gap statistic; however, the gap statistic may be used as a discordancy test to reveal the presence of outliers.

Multivariate classification using the background characterization approach was simulated using several different functions to describe the variation in the background distribution. Comparisons of principal components, ordinary least squares regression and reduced major axis regression indicate that reduced major axis regression and principal components are not only consistent with assumptions about geochemical data, but are less sensitive to varying degrees of data set truncation than is ordinary least squares regression. Furthermore, correcting the descriptive statistics of a truncated data set and calculating the background functions using these statistics produces residuals and scores which are predictable and thus can be distinguished easily from residuals and scores calculated for data from another distribution.

Table of Contents

List of Tables

List of Figures

Glossary of Abbreviations

PDF	Probability Density Function
BCA	Background Characterization Approach
RDML	Raw Data Maximum Likelihood
CIDML	Class Interval Data Maximum Likelihood
PC	Principal Components
OLS	Ordinary Least Squares
RMA	Reduced Major Axis

Acknowledgement

Technical funding for this project came from Natural Sciences and Engineering Research Council operating grant to Dr. Alastair J. Sinclair. A U.B.C. University Graduate Fellowship (2 years) and the Hugo E. Meilicke Graduate Fellowship (1 year) provided financial contributions of a more personal nature. Additionally, the Aaro E. Aho Memorial Scholarship contributed the funds used to purchase my first micro-computer, a contribution to which I am particularly indebted because, without inexpensive computer time, this study would have been prohibitively expensive.

Significant contributions to this thesis came from many of my friends and colleagues through words of encouragement, technical instruction, discussions and through other less tangible but equally important avenues. Several of the major contributors deserve special mention.

First of all, I'd like to thank Gordon Hodge for his assistance in drafting some of the figures, Frank Flynn for his help with the LaTeX macros, Dr. Kelly Russell for the use of his Compaq 386 micro-computer to typeset the thesis (mine does not have enough memory or speed), and the U.B.C. Groundwater and Petrology Groups for allowing me to use their laser printer to print the thesis. I also thank Dr. Thomas Brown for his assistance and frank discussions regarding functional stability, Dr. Kay Fletcher for his thoughts regarding the nugget effect, Dr. Rubin Zamar for his conversations regarding orthogonal regression and Dr. Glen Woodsworth for his recommendations regarding the probability plot software.

I also wish to express my sincerest gratitude to Dr. John Petkau for his tremendous assistance in guiding me into and through many of the statistical aspects of this thesis.

His help served to open my eyes regarding the importance of assumptions and the effects these assumptions can have on statistical analysis.

Finally, I'd like to thank Dr. Alastair J. Sinclair, who supervised this work. His straightforward approach to data analysis has certainly become part of my data analysis philosophy, and his advice, comments, enthusiasm, and critical evaluation of my work has contributed greatly to this study.

Chapter 1

Introduction

"I have yet to see any problem, however complicated, which, when you looked at it the right way, did not become more complicated."

Poul Anderson (1969)

"A good terminology is half the game."

Arthur Koestler (1981)

1.1 Purpose

Central to any data analysis procedure is a conceptual model of the genetic or deterministic processes the data are thought to describe. Associated numerical models commonly are used to characterize these conceptual models in quantitative terms, acting to describe the nature and behavior of the data. These numerical models may consist of simple combinations of statistical, empirical and theoretical components. In applied geochemistry, numerical models described by statistical terms are becoming increasingly common (Bölviken 1971; Sinclair 1974, 1976; Miesch 1981; Garrett 1987, Stanley and Sinclair, 1988).

In order to be useful, numerical models must be general enough to span the variation observed in the data sets to which they are applied. In addition, these models must be consistent with constraints imposed on the data by the conceptual models of their postulated genetic processes. The number and severity of these constraints which

limit application of these models differ markedly, but generally manifest themselves as underlying assumptions. Statistical models describing geochemical data, especially element concentration data applied to mineral exploration, environmental contaminant detection and geochemical mapping are no exception.

This study focuses on element concentration data in applied geochemistry and their evaluation using certain exploratory statistical data analysis procedures. A general multi-component model is postulated which approximately describes the probability density functions (*PDF*'s) of natural element concentration data. However, because *PDF*'s of similar general form occur in other types of geologic data, as well as in other physical sciences, the application of the results of this study is not restricted to the field of applied geochemistry.

The purpose of this research is to compare various statistical procedures used for the classification of element concentration data. All procedures tested are consistent with the constraints of the *PDF* models, but either represent different approaches to data analysis or impose slightly different statistical assumptions. Testing of these procedures by *monte carlo* simulation allows quantification of their levels of performance. Comparisons among these results are used to determine the relative effectiveness (efficiency) of each technique on data sets with different *PDF* forms.

1.2 Nature of Geochemical Data

The method by which geologists describe geological materials (the geologic nomenclature) is based, at least partially, on modal mineralogy. Because element concentrations in minerals are defined by stoichiometry, an analysis of the element concentrations can lead to an understanding of the modal frequency of the minerals within geologic materials and, thus, may be used to help in their classification (Thompson 1982a, 1982b).

Since physical and chemical processes commonly manifest themselves as variations in the composition of geologic materials, these processes are represented by characteristic frequency distributions of element concentrations. As a result, element concentrations have specific expected values (means), variations (standard deviations) and inter-element relationships (correlations; or in statistical terms - mean vectors and covariance matrices) which are characteristic of the specific physical and chemical processes that have caused variation in the composition of the analyzed geological materials.

In any natural environment, multiple physical and chemical processes have occurred to produce the variations in modal mineralogy, and thus variations in the element abundances observed in geochemical samples. These processes, in turn, may manifest themselves as a mixture of *PDF*'s, with each distribution representing a specific set of processes. Thus, the existence of mixtures of distributions in a data set may indicate that multiple genetic processes, or a single genetic process of varying intensity, have acted on the related geologic materials. Mixtures of distributions are, therefore, an expected *PDF* form for element concentration data in a variety of natural settings.

The element concentrations, depending on the processes which controlled them, can also occur in a variety of *PDF* forms. Frequency distributions reflecting primary and secondary geologic and geochemical processes are poorly understood; however, many authors have postulated that geological materials commonly display *PDF*'s with normal (or log-normal) forms (e.g. – Ahrens 1954; Aitchison and Brown 1957; Rodionov 1961; Shaw 1961; Sinclair 1974, 1976; Miesch 1981; Garrett 1987). Unfortunately, since element concentrations are intensive variables, and thus bounded by zero and one, geochemical *PDF*'s cannot be theoretically or accurately described in their tails by normal or log-normal distributions because these distributions have at least one unbounded tail.

1.2.1 Binomial Distributions

If random processes are the cause of variation in a data set, individual frequencies of geochemical concentration data for most hydromorphically borne elements should be distributed **binomially** according to the following *PDF* :

$$f(x_e) = \binom{n_t}{x_e} (\frac{n_a}{n_t})^{x_e} (1 - \frac{n_a}{n_t})^{n_t - x_e}, \tag{1.1}$$

where n_t is the total number of atoms in a geochemical sample, n_a is the number of atoms of the determined element, x_e is the observed number of atoms of the determined element in the geochemical sample ($x_e = 0, 1, 2, \ldots, n_t$), $\frac{n_a}{n_t}$ is the actual concentration and $f(x_e)$ is the probability of observing x_e number of atoms of interest. The mean and variance of the number of atoms of interest in this distribution are n_a and $n_a(1 - \frac{n_a}{n_t})$, respectively, provided the element of interest has a relatively high abundance.

This function is valid as a *PDF* model for hydromorphically borne elements because the following properties of geochemical data are consistent with the constraints required for a variable to be distributed binomially :

- the 'weight percentage' values (ppb, ppm or %) commonly reported in geochemical surveys can be easily transformed to mole (or atom) proportion values through division by the atomic (or gram formula) weight of the determined element (or species); thus an element concentration is merely the probability of an atom of a specific element being selected at random from a geochemical sample;

- the act of removing one atom of a specific element from a geochemical sample does not significantly alter the probability of removing a second atom of that same element from the geochemical sample (the probabilities of successive selection outcomes are essentially independent – non-hypergeometric), because of the large number of atoms present (on the order of Avogadro's number = 6.022×10^{23}

atoms) and because, since the atoms have been hydromorphically borne, each can be thought to occur independently (not in nuggets);

- geochemical data are intensive and must be bounded by zero and one (or 0 and 100 %, 0 and 1×10^6 ppm, 0 and 1×10^9 ppb, etc.); and
- the number of atoms of any element of interest is generally large with respect to the total number of atoms in the geochemical sample.

Elements which commonly display binomial distributions include those which are chemically mobile in the secondary environment (Cu, Pb, Zn, Ag, As, Sb, Ca, K, Na, Sr, etc.).

1.2.2 Poisson Distributions

On the other hand, if the element of interest is contained within a rare mineral grain (n_a is small with respect to n_t), assuming the theory of equant grains (Visman 1969, 1972; Ingamells 1974, 1981; Gy 1982), the resulting frequency will be distributed according to the following **Poisson** *PDF* :

$$f(x_e) = e^{-n_a} \frac{n_a^{x_e}}{x_e!}, \tag{1.2}$$

where n_a is the actual number of grains of the determined element in a geochemical sample and x_e is the observed number of grains of the determined element ($x_e = 1, 2, 3, \ldots, n_t$). The mean and variance of the number of grains of interest for a Poisson distribution are both equal to n_a. The Poisson distribution is an appropriate *PDF* model for elements borne in resistate minerals because all but the last constraint described above for binomially distributed elements apply. In this case, a nugget, instead of the atom, is the unit of measurement and the number of nuggets containing the

element of interest are small with respect to the total number of particles in the geochemical sample.

Elements such as Au, Pt, W, Ba and Sn, which commonly occur in rare resistate minerals, display this form of frequency distribution, especially where, relative to the true concentrations of the elements, small geochemical sample sizes are analyzed. In cases such as these, where the ideal equant grain model (Visman 1969, 1972; Ingamells 1974, 1981; Gy 1982) can be applied, the atoms occur in collective clusters known as grains (or nuggets). These grains report concentrations which occur in discrete and integral multiples of a basic concentration, related to the contribution of one grain to the observed concentration (refered to here as a *quantum*). These *quanta* are representative of individual mineral grains, and thus are discrete units, allowing the application of a Poisson distribution model to frequency distributions of this type.

1.2.3 Normal Distributions

As $n_t \rightarrow \infty$, with fixed $\frac{n_a}{n_t}$, the binomial distribution converges to the normal distribution (Hald 1952). Thus, if n_t is large ($n_t(\frac{n_a}{n_t}) \geq 5$ and $n_t(1 - \frac{n_a}{n_t}) \geq 5$; Hald 1952) and $\frac{n_a}{n_t}$ is fixed, a reasonable approximation to a binomial distribution (along with its integrals and derivatives) is the normal distribution (with its integrals and derivatives). In practical terms, if the size of the geochemical sample is made large enough, the expected element concentration distribution can be considered normal. Collection of geochemical samples of inadequate size will result in a poor approximation of the normal distribution by the binomial distribution, and thus a positive or negative skewness in the observed distribution. Element concentration data easily satisfy the above stipulation because geochemical samples commonly contain an extremely large total number (n_t) of atoms relative to the concentration of elements.

For hydromorphically borne elements in a 2 gram geochemical sample, $n_t \frac{n_a}{n_t} \gg 10^{12}$ and $n_t(1 - \frac{n_a}{n_t}) \gg 10^{12}$, even at ppb levels. The large number of atoms in geochemical samples mitigates the level of inaccuracy caused by the approximation of a discontinuous (binomial) distribution by a continuous (normal) distribution. In addition, the above conditions cause the tails of the binomial distribution to coincide with the tails of the normal distribution, thus reducing the bias of this approximation. As a result, in most cases, binomially distributed element concentrations do not require transformation to a normal distribution using the following operator :

$$\log \frac{p}{1-p}, \tag{1.3}$$

where p is the concentration (Johnson and Wichern 1982), although this may be recommended if adequate sampling is not possible.

Similarly, because the Poisson distribution is the limiting distribution of the binomial distribution as $n_t \to \infty$ with fixed n_a, it can also be approximated by a normal distribution provided that the geochemical sample is, as above, large enough to ensure that a large number of grains of the resistate mineral borne element of interest are present (Hald 1952). Where the size of the geochemical sample is large, the number of rare resistate grains increases and the size of the *quantum* concentration is reduced (because each grain represents a smaller proportion of the geochemical sample). As the *quantum* concentration $\to 0$, the resulting (Poisson) distribution converges to a normal distribution.

In many applications, due to cost or logistical limitations, geochemical samples of adequate size cannot be collected to ensure that a large number of grains containing the element of interest are present (e.g. Au, Pt and other resistate elements which occur in discrete grains at low concentration). The assumption of a normal distribution is not valid in these cases; however, in this study, geochemical samples are assumed to

be large enough to avoid the possibility of error in the approximation of a (discrete) Poisson distribution by a (continuous) normal distribution.

Thus, regardless of whether an element of interest in geochemical samples is present in low (Poisson distributed) or high (binomially distributed) abundances, or is transported as resistate mineral grains or hydromorphically, provided that the geochemical samples are of sufficient size, the frequency distribution form of geochemical data can be approximated by mixtures of normal distributions, each of the form :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \tag{1.4}$$

where x is the observed concentration of the element of interest. For the normal distribution, μ is the mean concentration ($= \frac{n_a}{n_t}$) and σ is the standard deviation.

1.3 Treatment of Errors

Variation in element concentrations in geochemical samples may be caused by poor geochemical sample collection, inadequate or improper geochemical sample size reduction or analytical procedures, or may be due to random spatial variations. Although collection, reduction and analysis procedures of geochemical samples can provide relevent geologic information regarding the mineral grain size (Clifton et al. 1969; Ingamells 1974, 1981; Stanley 1986) and occurrence of specific elements, most hypotheses which have geological or geochemical implications concern spatial variations. These spatial variations often are related to different geological or geochemical processes operating in different parts of a survey area.

Variations contributed through the geochemical sample collection, geochemical sample size reduction and analytical errors may obscure real variations of geological or geochemical import. However, all sources of variation must be evaluated in geochemical data analysis procedures. If significant variations exist, these must be distinguished

from variations due to geological or geochemical processes in order to avoid irrelevant or spurious conclusions. Where variations due to geological or geochemical processes can be isolated, significant geologic or geochemical conclusions are possible. Understanding the causative contributions to the variance of geochemical data is at least one of the motivations for any exploratory data analysis procedure.

In this study, no effort will be made to distinguish the nature or cause of the individual errors and variations associated with geochemical data; however, the existence of these errors is recognized and statistical procedures are formulated assuming error is present in all geochemical data for all subsequent data analysis.

1.4 Distribution Model for Element Concentration Data

The preceding discussion leads to the postulation of the following paradigm to describe the frequency distribution of element concentration data :

Specific Geological or Geochemical Processes Acting on Geologic Materials Are Characterized by Specific Multi-Element Frequency Distributions.

The following rules describe this postulated model for the frequency distribution of geochemical data and encompass all types of variation observed in geochemical data frequency distributions :

- Frequency distributions are composed of a mixture of distributions,
- Each distribution has a specific mean vector (a set of means),
- Each distribution has a specific covariance matrix (a set of standard deviations and correlations),
- Each distribution is multivariate normal.

An important consequence results from this postulated distribution model. Specifically, **mixtures of normal distributions can have frequency forms which approximate other positively (or negatively) skewed** *PDF*'s (e.g. – log-normal, binomial, or Poisson distributions). This may occur where, in a general mixture of two distributions, one distribution has a larger standard deviation or comprises a smaller proportion of the data set than the other distribution. Situations where the distribution with the larger mean has a higher standard deviation can be expected where the total variance (geological, sampling and analytical) is proportional to concentration. Sampling and analytical errors are known to exhibit this proportionality effect (Thompson 1973; and Howarth 1973, 1976a, 1976b, 1978). Likewise, situations where the distribution with the larger mean comprises a smaller proportion of the data set are common, especially in mineral exploration and contaminant detection, because the geochemical anomaly generally underlies only a small proportion of the survey area. Obviously, confusion may result with data sets which exhibit positive skewness because the geoscientist may not be able to readily distinguish whether the skewness is a result of a mixture of distributions or is caused by inadequate sampling producing a poor approximation of the normal distribution by the binomial or Poisson distributions. Because distributions with positive skewness are common in both mineral exploration and environmental contaminant detection, identifying the cause of observed skewness in a data set is critical to subsequent interpretation.

The above model limits the number and types of data analysis techniques which can be used to evaluate geochemical data. As a result, this study will address, particularly, the performance of those techniques which are consistent with the proposed model.

1.5 Use of the Conceptual Model

As discussed above, determination of element concentrations in geochemical samples within a survey area can lead to the determination of the minerals present, and thus yield an understanding of the geology and physical and chemical processes which act to modify geological materials during formation or destruction (weathering). Geochemical mapping, contaminant detection and ore discovery follow directly from this process.

Unfortunately, element concentrations are continuous (real-valued) variables, whereas geological materials are defined categorically. A transformation is required to 'map' the continuous geochemical concentration data onto the geological material categories. In geochemical applications, this transformation has taken the form of classification through the selection of a threshold concentration (Rose, Hawkes and Webb 1979). Geochemical samples with concentrations exceeding this threshold are classified as representative of one type of geologic material, presumably differing from other geologic materials due to the action of at least one additional genetic process. Those geochemical samples with concentrations less than the threshold are classified as representative of another type of geologic material with a different genetic history. Other forms of geological or geochemical knowledge can then be used to test these exploratory 'classification' hypotheses (Popper 1968) and determine the nature of the geological or geochemical processes which acted to produce the observed variation in the geological materials.

Classification by way of threshold selection is an appropriate exploratory data analysis approach for data sets involving geochemical mapping, contaminant detection and mineral exploration. In the case of geochemical mapping, the component distributions of mixtures of *PDF*'s are thought to be related to the different source materials from which they were derived. Because the amount of detail collected in mapping is scale

dependent, the mappable features in a survey each cover significant proportions of the survey area. As a result, each distribution consists of a significant proportion of the data set. Classification procedures used in this application are termed 'population discrimination' techniques.

In the search for mineral deposits and the detection of contaminants, the distribution of interest (that related to mineralization or contamination) generally comprises a small proportion of the data set (especially during early reconnaissance stages of investigation). Classification procedures used in these applications are called 'outlier recognition' or 'anomaly recognition' techniques.

Both 'anomaly recognition' and 'population discrimination' procedures are actually end-members of a continuum of classification techniques. Intermediate cases occur due to variations in the proportions of the component distributions. Appropriate data classification techniques must be designed to recognize anomalous geochemical samples which may be representative of a mineral deposit, a contaminant source or a different type of geological material. For consistency in this study, all of these techniques are included in the more general term of 'classification procedures'.

1.6 The Background Characterization Approach

Classically, the term 'anomaly' is defined as "a deviation from the norm" (Rose, Hawkes and Webb 1979, p. 34). Thus, mineral deposits, contaminant sources and geological materials comprising a small proportion of the data set may exhibit anomalous geochemical signatures. In order to recognize this anomalous character, one must first determine the characteristics of 'the norm'. In geochemical data analysis, the norm is called 'background' (Rose, Hawkes and Webb 1979), and is commonly considered to be

the geochemical signal of the predominant geological material after possible modification by surficial weathering processes.

Generally, the models used in anomaly recognition and population discrimination procedures allow recognition of those geochemical samples derived from (the possibly several) 'background' geologic materials. As a result, the first objective of exploratory geochemical data analysis (after confirmation of data quality) is to identify these background geochemical samples and then use them to define the background geochemical signature(s) or model(s). If several geochemical variables exist, and a background model can be successfully developed, it may be used to classify the geochemical samples, which are not derived from the background distribution, by recognizing those geochemical samples which differ substantially (and in varying ways) from this background model. This technique is known as the '**background characterization approach**' (Stanley and Sinclair 1987).

This procedure involves the selection of thresholds to separate the background geochemical samples from the enigmatic geochemical samples (those which are anomalous or which are derived from another background distribution). In this way geochemical samples which cannot be confidently determined as representative of the background geochemical signature can be truncated from the data set so that subsequent calculation of a background model excludes their (possibly spurious) contribution (Figure 1.1). The procedures by which this may be accomplished are discussed below and include the 'mean plus 2 standard deviations', '95^{th} percentile', gap statistic and probability plot approaches. Then, after the selection of a threshold and definition of those geochemical samples which are enigmatic, two lines of investigation may follow.

Geoscientists have applied a regression approach to the problem of determining the background geochemical signature (Rose, Hawkes and Webb 1979; Matysek et al. 1982; Day et al. 1987; Stanley and Sinclair 1987). In this technique, several independent

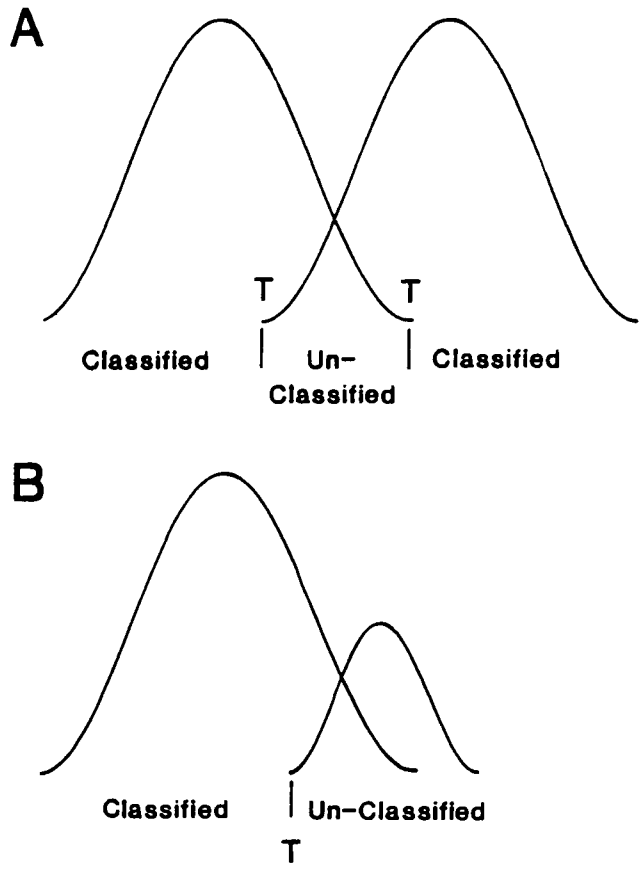


Figure 1.1: Threshold Selection to Separate Background Observations from Enigmatic Observations

A. Population Discrimination – the two thresholds (T) define a range of population overlap where further classification may be necessary. Two background models may be developed, one for each distribution.

B. Anomaly Recognition – the single threshold (T) defines a range where the data are enigmatic (possibly anomalous) and where further classification may also be necessary. A single background model may be developed because the second distribution is too small to adequately characterize its variation.

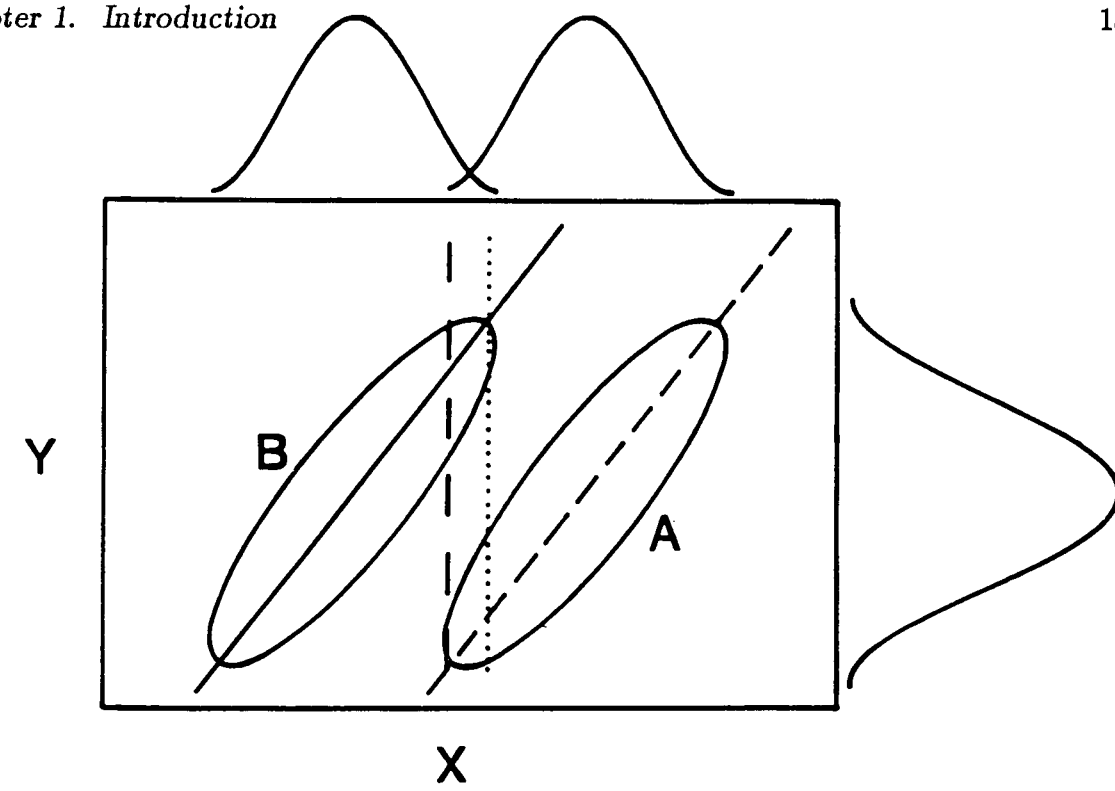


Figure 1.2: Bivariate Example of the Background Characterization Approach to Anomaly Recognition

This plot of X *versus* Y demonstates how the *BCA* can be used to discriminate populations. Ellipses represent the 95^{th} percentile density contour of the A and B distributions. Marginal histograms (located on the top (X) and side (Y) of the plot) show that approximate classification can be achieved using variable X, but not Y. The dashed vertical line defines boundary below which data from population A does not (significantly) occur. Thus, data below this threshold can be used to define a background model describing the variation in population B (the solid diagonal line). Applying this model to data above the threshold readily separates the data from population B from that of population A because the residuals (or scores) defined by the background model for these two distributions are substantially different.

If the proportion of population A was large, a threshold could also be defined for population A (the dotted vertical line) to identify the data to be used to define a background model for population A. Residuals (or scores) calculated from this background model (the dashed diagonal line) could also be used to classify the data from the region of population overlap (between the two vertical lines). This situation would be a population discrimination problem. If the proportion of population A was small, this situation would be an anomaly recognition problem and this second background model would not generally be defined from population A because of the sparsity of data.

In this example, the standard deviations and correlations for X and Y are the same for both distributions. This need not be true for application of the *BCA*.

geochemical variables are regressed against a dependent variable. Once this regression function has been defined (using **only** the background geochemical samples), the function is applied to the enigmatic geochemical samples. Residuals for these enigmatic geochemical samples are then evaluated in a univariate context (Figure 1.2) using similar threshold selection and classification procedures. If truly anomalous geochemical samples exist in a data set, these would be expected to 'react' differently to the regression function than the background geochemical samples. As a result, those enigmatic geochemical samples with residuals which differ substantially from what would be expected are considered to be anomalous.

Alternatively, instead of a regression function applied to the background geochemical samples, the principal components of the background data are determined (Roquin and Zeegers 1987, Lindqvist et al. 1987). This approach rigidly rotates the multivariate data in hyperspace to determine the the most important (uncorrelated) sources of variation. Then, instead of evaluating the residuals, the principal component scores are examined and those geochemical samples with scores which differ substantially from what would be expected are considered anomalous using the same rationale and classification procedures.

Both linear regression and principal components analysis serve to reduce the number of variables to be considered, allowing a more rapid evaluation of multivariate data sets using the univariate classification procedures described below. However, these two statistical approaches have important philosophical implications which may serve to limit their application under certain circumstances. The philosophy of the linear regression approach involves a conceptual model which postulates the linear control of one geochemical variable by others. Examples of these types of causal relationships in geochemical data include :

- bulk changes in the composition of the waters in equilibrium with geological materials, such as *pH*, *pe* and the concentrations of other disolved species, or physical properties, such as temperature and pressure, may control the solubility of elements, and

- variations in the crystal composition of the geological materials, such as the abundance and crystallinity of *Fe*- and *Mn*-oxyhydroxides, may cause different magnitudes of metal adsorbtion by the geological materials.

Principal components analysis implies a significantly different conceptual model, assuming that the observed variables are linear manifestations of an unobserved, underlying factor. Thus, the principal components approach produces a quantitative measure of this 'latent' factor. Geoscientists must consider the goal of the analysis, the variables available and their possible geologic and geochemical relationships before selecting regression or principal component analysis as the appropriate background geochemical function.

Care must also be exercised in the application of both of these techniques. If multiple background signatures exist in the data set, both the regression and principal component analyses will merely serve to recognize the variables which contribute the most to the variation among the geochemical signatures. In these cases, the background functions will not necessarily define those variables which contribute to variation within a single component distribution (such as background). Robust statistical approaches to avoid this potential problem have been recommended by Lindqvist et al. (1987) and Wurzer (1988). Alternatively, Stanley and Sinclair (1987) have recommended subsetting the data into groups defined by categorical variables (lithology, alteration, etc.) which may be responsible for the different background signatures, and evaluating each group independently.

1.7 Univariate Classification Techniques Used in Applied Geochemistry

The initial approach commonly taken in any form of geochemical data analysis is to evaluate each variable independently. If results from this analysis are unsatisfactory or incomplete, subsequent multivariate analysis may be performed. In cases where a large number of variables need to be evaluated, the univariate analysis phase may be skipped entirely, and only a multivariate analysis performed.

Although numerous univariate data analysis procedures have been applied to geochemical data, only a few are consistent with the above proposed paradigm for the frequency distribution of geochemical data. The following data analysis techniques have been advocated and used to evaluate geochemical data by geoscientists. The merits and shortcomings of each technique are discussed relative to the proposed model for the frequency distribution of geochemical data.

1.7.1 Experiential Selection

Selection of a geochemical threshold based solely on the 'experience' of the geoscientist is called experiential selection. This arbitrary technique does not require a distribution model to describe the variation in the element concentrations. Thresholds are chosen to 'distinguish anomalous geochemical samples from background geochemical samples' or to 'discriminate geochemical samples derived from different lithologies'. A typical example of this type of classification occurs when a mineral explorationist further explores only those portions of a survey area which have element abundances above a certain concentration, on the assumption that these are derived from a weathering mineral deposit. The arbitrary nature of this technique and lack of any quantitative frequency distribution model make it, by definition, un-scientific (Popper 1968) and, as a result, it is not considered in this study.

1.7.2 Mean ± X Standard Deviations

A widely advocated and utilized threshold selection technique in applied geochemistry is the choice of a threshold at the 'mean plus or minus some multiple of standard deviations' (Rose, Hawkes and Webb 1979; in common practice the 'mean plus 2 standard deviations'). This approach assumes that the anomalous geochemical samples have higher element concentrations than the norm (not generally true), and that the frequency distribution of the data is normal (also not generally true). The subjective choice of how many standard deviations to add or subtract to or from the mean to define a threshold and the assumption of a normal distribution limits the range of applications for which this technique is usable. Although no detailed evaluation of the performance of this anomaly recognition approach is undertaken, its projected classification performance will be discussed relative to other classification methods.

1.7.3 Threshold = Y^{th} Percentile

Another commonly used anomaly recognition approach involves selection of a threshold at some arbitrary 'cumulative percentile of the sorted data'. Although this approach makes no assumption about the form of the frequency distribution being considered (and thus is non-parametric), it still suffers from the subjectivity imposed by the selection of a percentile, commonly the 95^{th} percentile. The non-parametric nature of this technique is the basis for its favored use by some geoscientists.

The distribution model implied by the use of this technique generally takes the following form. Where a geoscientist can assume that the anomalous element concentrations occur in the positive tail of the distribution, selection of a threshold at an intermediate percentile between the anomalous and background concentrations is appropriate. Unfortunately, no method is generally employed to help determine the value

of this intermediate percentile; thus, its choice is arbitrary.

This technique may also be applied where economic or time contraints limit the amount of follow-up evaluation which may be done. In these cases, selecting a threshold at a fixed percentile guarantees that time and resources are available to ensure that all 'anomalies' can be further investigated because a fixed limitation is made on the maximum number of geochemical samples classified as anomalous. No detailed evaluation of the performance of this approach is undertaken, but its projected classification performance will also be discussed relative to the other classification methods.

1.7.4 Histograms

Another approach to threshold selection and classification involves the visual inspection and selection of anti-modes (low frequency classes on discrete histograms or low frequency ranges on continuous histograms which are bounded by higher frequency classes or ranges; Rose, Hawkes and Webb 1979). These anti-modes are thought to be **rough** approximations of thresholds which separate (or discriminate) the two adjacent distributions. In practice, this approach has been carried out in a subjective manner.

The histogram approach suffers from several serious theoretical and procedural problems which substantially affect the quality of the resulting classification. Two component mixtures of normal distributions have been shown by Eisenberger (1964) to have the following properties :

- they are unimodal if they satisfy the following sufficient condition :

$$(\mu_1 - \mu_2)^2 < \frac{27\sigma_1^2\sigma_2^2}{4(\sigma_1^2 + \sigma_2^2)}, \tag{1.5}$$

- they are bimodal if they satisfy the following sufficient condition :

$$(\mu_1 - \mu_2)^2 > \frac{8\sigma_1^2\sigma_2^2}{(\sigma_1^2 + \sigma_2^2)}, \tag{1.6}$$

- with $\mu_1 = \mu_2$, they are unimodal for all ϖ (the component percentage of one of the distributions), and
- for every set of possible μ_1, μ_2, σ_1 and σ_2 parameter values in a two-component mixture of normal distributions, values of ϖ exist which make the *PDF* unimodal.

Similarly, Behoodian (1970) has shown, for two component mixtures of normal distributions, that :

- a more constraining sufficient condition for unimodality is :

$$|\mu_1 - \mu_2| \leq 2 \times \min(\sigma_1, \sigma_2), \tag{1.7}$$

and

- if $\sigma_1 = \sigma_2 = \sigma$, the two-component mixtures of normal distributions are unimodal if :

$$|\mu_1 - \mu_2| \leq 2\sigma\sqrt{1 + \frac{|\log \varpi - \log(1-\varpi)|}{2}}. \tag{1.8}$$

Finally, McLachlan and Basford (1988) have shown that :

- if $\varpi = 50$ % and $\sigma_1 = \sigma_2 = \sigma$, then two-component mixtures of normal distributions are bimodal if and only if :

$$\frac{|\mu_1 - \mu_2|}{\sigma} > 2. \tag{1.9}$$

In the continuous unimodal histogram cases above, no threshold can be selected because no anti-mode exists, and the distribution with the smaller frequency appears as a shoulder on the side of the distribution with the larger frequency (Figure 1.3). Therefore, unimodality of a *PDF* does not suggest that a distribution is not a mixture of two or more normal components. Moreover, small statistical samples of a single normal

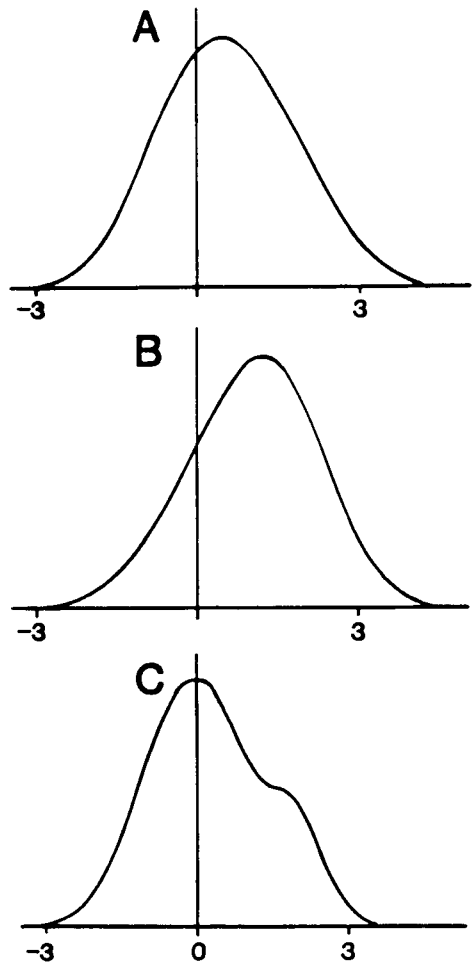


Figure 1.3: Examples of Mixtures of Normal Distributions Exhibiting Unimodal Probability Density Functions
In each case $\mu_1 = 0.0$ and $\sigma_1 = 1.0$.

A. $\mu_2 = 1.5$, $\sigma_2 = 1.0$ and $\varpi = 0.30$.

B. $\mu_2 = 1.5$, $\sigma_2 = 1.0$ and $\varpi = 0.60$.

C. $\mu_2 = 2.0$, $\sigma_2 = 0.5$ and $\varpi = 0.85$.

Case A appears to be a symmetrical distribution while case B exhibits negative skewness. Only in case C is it obvious that the distribution it actually a mixture of two distributions. Modified from Everitt and Hand (1981).

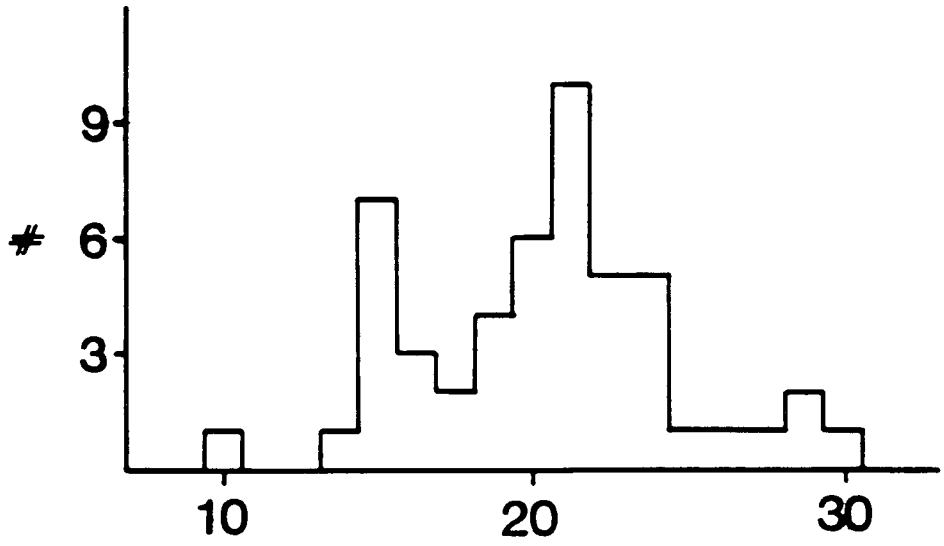


Figure 1.4: Example of a Single Normal Distribution Exhibiting a Multimodal Histogram
Four modes occur on this histogram of 50 observations derived from a single normal distribution with $\mu = 20$ and $\sigma = 5$.

distribution represented in class intervals on a discrete histogram may appear multimodal, suggesting the presence of more than one component distribution (Figure 1.4). Obviously, relying on the location of histogram anti-modes as estimates of thresholds can lead to a large number of classification errors.

Two-component mixtures of *PDF*'s which have very different component percentages or standard deviations also have anti-modes which do not define a threshold which adequately classifies the data (Figure 1.5). In these cases, a large number of classification errors (both of omission and inclusion) occur. Only in cases where the standard deviations and component percentages of the distributions are equal does the anti-mode define a threshold which minimizes the total number of classification errors. Thus, visual inspection and selection of thresholds using histograms cannot locate optimal thresholds in the majority of real cases in applied geochemistry.

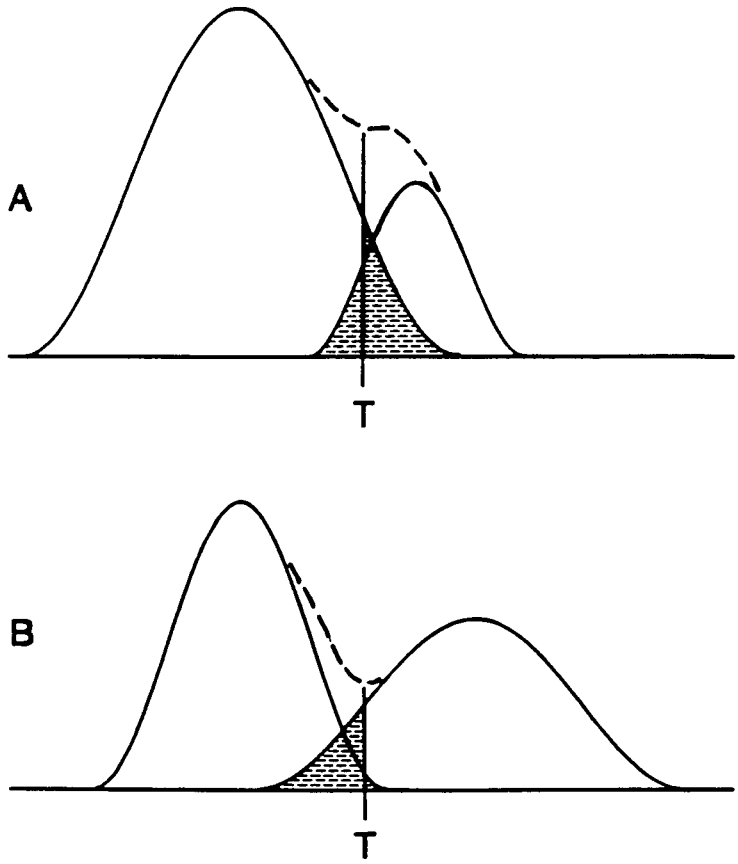


Figure 1.5: Distributions Where the Anti-Modes do Not Define Optimal Thresholds

In A, the distributions have the same σ, but different component proportions. In B, the distributions have different σ, but the same component proportions. In both cases, the shaded areas are proportional to the amount of misclassification produced by the selection of a threshold at the anti-mode. It can be shown that the threshold which minimizes the amount of misclassification divides this shaded region into two equal areas (see Chapter 2). Clearly, the thresholds defined by the anti-modes of these mixtures of distributions (T) do not split the shaded regions equally and thus do not classify the data with a minimum of error.

Furthermore, utilization of different class interval sizes and limits, regardless of whether they are uniform or irregular, may cause the histograms to exhibit entirely different anti-mode values. Finally, the highly subjective assignment of a threshold within any anti-modal class interval (a finite range in which no available criteria exist to define a best location), makes this procedure highly subjective. For the above reasons, use of the histogram approach to data classification is difficult at best and it will not be discussed because of its many theoretical similarities to the probability plot technique discussed below.

1.7.5 Probability Plots

The probability plot approach to classification is very similar to the histogram approach, except that the thresholds are defined either manually, visually (using the cumulative frequencies; Sinclair 1974, 1976; Figure 1.6) or numerically (using the individual or cumulated densities; Stanley 1987). This approach allows more rigorous approximation of thresholds in a graphical context, because all parameters (μ, σ and ϖ) can be visually estimated directly from the probability graph.

The manual approach requires that the cumulative frequency data be plotted on a probability graph and the percentiles of the points of inflection defined. The number of inflection points determine the number of component populations in the distribution (equal to the number of inflection points plus one). Using the formula :

$$C_m = \sum_{k=1}^{v} \varpi C_k, \tag{1.10}$$

where $\sum_{k=1}^{v} \varpi = 1$, C_m is the cumulative frequency of the mixture of distributions, v is the number of component distributions and the C_k's are the cumulative frequencies of the component distributions, a curvilinear model is fitted to the cumulative data with the graphical approach described by Sinclair (1974, 1976). The mean and standard

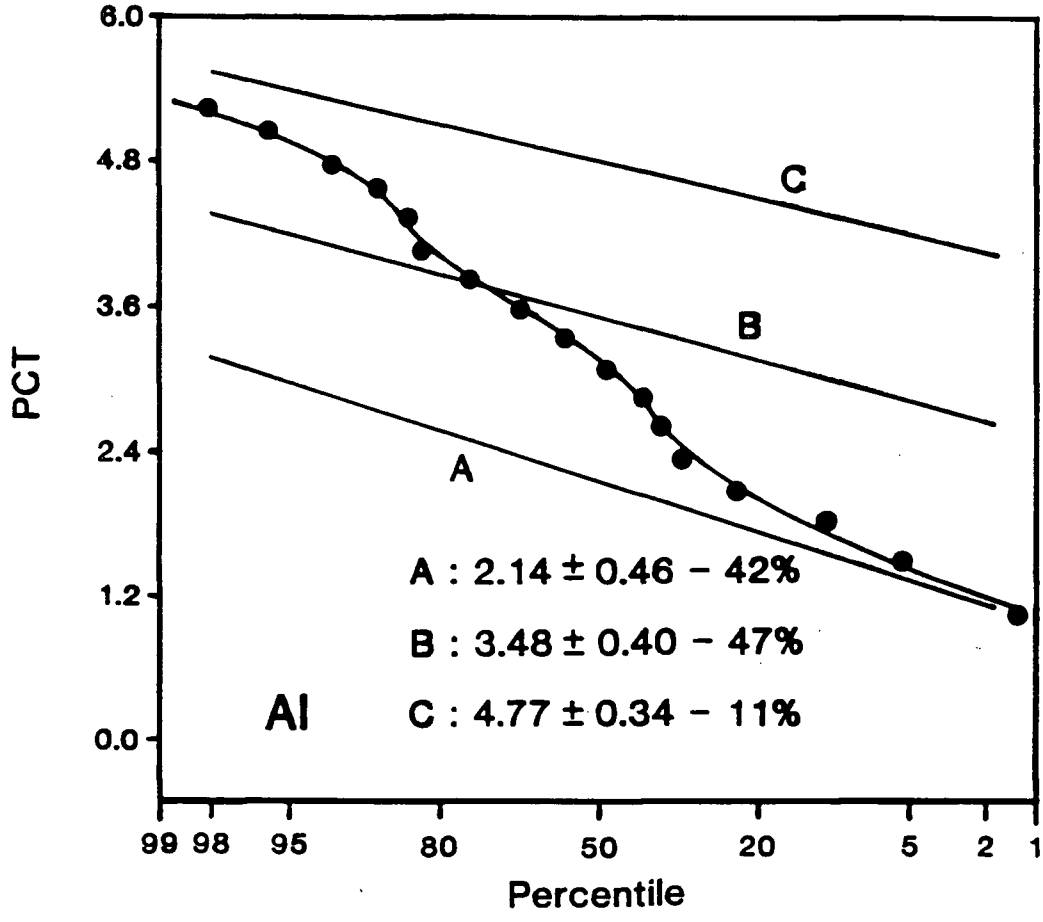


Figure 1.6: Example of Probability Plot Approach to Data Classification
This Al (pct) concentration data has been modeled as a mixture of three normal distributions in the ratio 42 : 47 : 11, with means and standard deviations as indicated ($n = 114$; from Stanley and Sinclair 1988).

deviation of each normal component distribution are the parameters of this curvilinear model and may be estimated directly from the probability graph (Figure 1.6).

Numerical optimization requires that the mixture of normal distribution *PDF* be fitted to the frequency data by maximizing the likelihood function (Stanley 1987). Likelihood ratio tests can be performed to determine the number of distributions used in the numerical optimization which are significant at a certain confidence level. This approach is discussed in detail in Chapter 2.

The parameters of the distribution model (the means, standard deviations and component proportions (ϖ)) may then be used to select a single threshold which defines the smallest number of total classification errors based on the theoretical characteristics of the model. This parametric approach involves no subjective decision other than how to determine the number of component distributions and the weights assigned to the different types of classification errors (omission and inclusion). This approach is an improvement over the histogram approach because all parameters and thresholds may be determined (or calculated) directly from the probability graph.

1.7.6 Gap Statistic

The gap statistic (Miesch 1981), another relatively objective technique that has not been widely used in geochemistry, involves a different set of assumptions about the *PDF* of geochemical data. Instead of assuming that the *PDF* is multi-normal, this technique tests the null hypothesis that 'the data are 3-parameter log-normally distributed'. Scores (gaps) are computed for each adjacent pair of concentration values. The gaps are each related to a corresponding value equal to the midpoint of the pair of concentrations associated with that gap. If the largest gap exceeds a critical value at some confidence level, defined by what would be expected from random sampling of

a 3-parameter log-normal distribution, the null hypothesis is rejected and the distribution is not considered to have a 3-parameter log-normal form. The midpoint location related to this gap (whether significant or not) is defined as the threshold (Miesch 1981). This technique is discussed in detail in Chapter 4.

1.8 Classification Techniques Addressed

Although several of the above univariate techniques commonly are applied to the analysis of geochemical data, only two of these are addressed in detail in this study. Those approaches evaluated include the probability plot and gap statistic techniques; however, a discussion of how these results would compare with those from the 'mean plus 2 standard deviations' and the '95^{th} percentile' techniques is also presented.

Linear regression and principal components analysis techniques are addressed in the context of the 'background characterization approach'. They are used solely to aid in classification of a single geochemical variable where population overlap prevents adequate classification of the data using a single threshold. Multivariate background models can then be developed to characterize the background variation and help to classify the enigmatic observations. These procedures are used to improve classification of the marginal distributions where conditions of extreme overlap are present and univariate classification results in a large number of classification errors.

Chapter 2

Theory of Probability Plot Analysis

"The difference between art and science is that science is what we understand well enough to explain to a computer. Art is everthing else."

Donald Knuth (1987)

"Build a system that even a fool can use, and only a fool will want to use it."

Shaw's Principle (1979)

2.1 Mixtures of Distributions

The recognition of the existence of mixtures of distributions in geochemical data was, at least implicitly, accepted at such an early stage in the development of the science of applied geochemistry, that the definitions of 'background', 'threshold' and 'anomaly' are all embodied within it (Rose et al. 1979). Its overall acceptance as an element concentration distribution model led to its mathematical formulation in a specific *PDF*. Consequently, geoscientists have numerically optimized this function to 'fit' the observed frequency distribution and estimate the corresponding parameters. This, in turn, allows the calculation of thresholds which optimally classify individual observations as part of the component distributions.

The general formula for a normal frequency distribution mixture model which is

'fitted' to the observed frequency data has the following *PDF* :

$$p(x) = \sum_{k=1}^{v} \frac{\varpi_k}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}, \tag{2.11}$$

or, considering only the independent parameters :

$$p(x) = \sum_{k=1}^{v-1} \frac{\varpi_k}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2} + \frac{(1-\sum_{k=1}^{v-1}\varpi_k)}{\sqrt{2\pi}\sigma_v} e^{-\frac{1}{2}\left(\frac{x-\mu_v}{\sigma_v}\right)^2}, \tag{2.12}$$

where v is the number of normal distribution components, μ_k and σ_k are the means and standard deviations of each component distribution, ϖ_k are the percentages of the normal distribution components, and $\sum_{k=1}^{v} \varpi_k = 1$. For v normal distribution components, there are $3v - 1$ independent parameters required to describe their combined *PDF*.

2.2 Algorithms for Parameter Optimization

Numerous statistical techniques have been applied to discriminate, partition or decompose mixtures of distributions into their component distributions. Most partitioning approaches involve fitting a normal distribution mixture model to frequency data and, in the process, produce parameters estimates for the distribution model. Several techniques to determine estimates of the parameters of a mixture of normal distributions have been advocated. These include (Silverman 1981; Everitt and Hand 1981; Titterington et al. 1985; McLachlan and Basford 1988;) :

- Minimum Distance Techniques,
- Graphical Methods,
- Method of Moments,
- Bayesian Methods, and

- Maximum Likelihood Optimization.

Parameter estimation using these techniques is not simple, largely because of the dual nature of the parameters to be estimated. The component proportions (ϖ_k) are bounded by 0 and 1, whereas the means and standard deviations (μ_k and σ_k) are parameters of normal distributions. This dichotomy, as well as the constraints on the parameters ($0 \leq \varpi_k \leq 1$, $\sum_{k=1}^{v} \varpi_k = 1$ and $\sigma_k > 0$), can cause numerous numerical stability problems in all of these parameter estimation techniques.

The properties of these techniques, as well as their advantages and disadvantages, are described below (Fryer and Robertson 1972; Tan and Chang 1972; Quandt and Ramsey 1978).

2.2.1 Minimum Distance Techniques

Minimum distance methods for determining the equation of a function through a set of points have been used to determine the parameters describing the *PDF* of frequency data (Mundry 1972; Clark 1976). Generally, this approach is invoked by minimizing some 'distance' criterion which quantifies the extent of deviation of the distribution model from the observed frequency data. In some cases, this is equivalent to 'least squares' optimization. Advantages of this approach to parameter estimation include its applicability to discrete and continuous variables and to multivariate distributions.

Calculations of the minima generally are made using the multivariate Newton-Raphson algorithm :

$$\vec{\Psi}^{r+1} = \vec{\Psi}^{r} - \tilde{F}''(\Psi^{r})^{-1}\vec{F}'(\Psi^{r}), \tag{2.13}$$

where $r = 0, 1, 2, \ldots$, $\vec{F}'(\Psi^r)$ is the vector of first derivatives (Jacobian vector) and $\tilde{F}''(\Psi^r)$ is the Hessian matrix of the likelihood function with respect to a vector of the parameters $\vec{\Psi}^r$. This procedure finds the roots of the estimating equations and these

are screened to separate the minima from the maxima.

Minimizing functions which have been used are presented in Table 2.1 (Kullback and Leibler 1951, Wolfowitz 1957, Rao 1965, Choi 1969, MacDonald 1975, MacDonald and Pitcher 1979, Parr 1981; after Titterington et al. 1985, p. 116). In certain applications, some of these distance measures have specific advantages over others. Specifically, the similarity of the Kullback-Leibler distance to the natural logarithm of the likelihood function (see below), in terms of both form and performance, "makes it the optimal choice for likelihood adherents" (Titterington et al. 1985). Likewise, in applications where close approximations are required for the tails of the distribution, the χ^2 and Modified χ^2 techniques are favored.

However, problems can occur during the minimization if the denominator terms of the Kullback-Leibler, χ^2, and Modified χ^2 distance measures equal or approach zero and the functions become extremely unstable. As a result, depending on the numerical algorithm employed, solution may not be possible or an unacceptably large number of iterations may be required for convergence to the correct solution.

2.2.2 Graphical Methods

Graphical techniques have also been widely used (Harding 1949; Preston 1953; Cassie 1954; Strömgren 1954) to decompose mixtures of distributions. The purposes of these informal procedures are :

- to determine whether a certain mixture of distributions conforms to the observed frequency data, and
- to provide estimates of the means, standard deviations and percentages of the component distributions.

Table 2.1: Minimum Distance Functions Commonly Used to Determine Optimum *PDF* Parameter Values

Function Description	Continuous Variable	Discrete Variable
Integrated *PDF*'s	$\int[F_o(x) - F_m(x)]^2 dx$	$\sum_{j=1}^{m}(\sum_{i=c_{j-1}+1}^{c_j} q_{o_i} - \sum_{i=c_{j-1}+1}^{c_j} q_{m_i})^2$
PDF's	$\int[f_o(x) - f_m(x)]^2 dx$	$\sum_{i=1}^{n}(q_{o_i} - q_{m_i})^2$
Weighted Integrated *PDF*'s	$\int[F_o(x) - F_m(x)]^2 w(x) dx$	$\sum_{j=1}^{m}(\sum_{i=c_{j-1}+1}^{c_j} q_{o_i} - \sum_{i=c_{j-1}+1}^{c_j} q_{m_i})^2 w_j$
Weighted *PDF*'s	$\int[f_o(x) - f_m(x)]^2 w(x) dx$	$\sum_{i=1}^{n}(q_{o_i} - q_{m_i})^2 w_i$
χ^2	$\int[f_o(x) - f_m(x)]^2 dx / f_m(x)$	$\sum_{i=1}^{n}(q_{o_i} - q_{m_i})^2 / q_{m_i}$
Modified χ^2	$\int[f_o(x) - f_m(x)]^2 dx / f_o(x)$	$\sum_{i=1}^{n}(q_{o_i} - q_{m_i})^2 / q_{o_i}$
Wolfowitz Distance	$\int \lvert F_o(x) - F_m(x)\rvert dx$	$\sum_{j=1}^{m}\lvert\sum_{i=c_{j-1}+1}^{c_j}(q_{o_i} - q_{m_i})\rvert$
Kullback-Leibler	$\int \log[dF_o(x)/dF_m(x)]dF(x)$	$\sum_{i=1}^{n} q_{o_i} \log(q_{o_i}/q_{m_i})$

$F_o(x), F_m(x)$ = integrated densities of observed data value and distribution model,
$f_o(x), f_m(x)$ = densities of observed data value and distribution model,
$w(x)$ = weight for each observation,
q_o, q_m = discrete densities of observed data value and distribution model,
m = number of discrete variable values,
n_j = number of observations for each discrete variable value,
c_j = cumulative number of observations up to each discrete variable value,
n = number of observations,
where $\sum_{j=1}^{m} n_j = n$,
and $\sum_{i=1}^{j} n_j = c_j$.

These techniques are largely restricted to univariate normal and log-normal applications, and comprise the earliest attempts at statistically evaluating and decomposing mixtures of distributions. Two main type of plots have been used : those based on the density function (histograms), and those based on the distribution function (probability graphs). "In general, in order to be of much use, the former require more data than the latter." (Titterington et al. 1985, p. 52).

The first objective, when graphically decomposing mixtures of distributions, is to determine the number of component distributions. Three density-based graphical approaches to this problem have been advocated (Tanner 1959; Tanaka 1962; Bhattacharya 1967).

Tanner (1959) plots the first and second differences of the histogram counts (n_j), and uses the existence of local maxima (modes) and local minima (anti-modes) to determine the number of component distributions. These can also be used to determine crude parameter estimates. The approach suffers from the problems described in Chapter 1, in that mixtures of distributions may not exhibit multi-modality.

The more elegant approach of Bhattacharya (1967) is based on two facts :

- the natural logarithm of the normal density is a concave quadratic in x, and its derivative is linear with negative slope,
- if n is large and the class intervals on a histogram are small, the histogram heights are proportional to the density.

Thus, a 'Bhattacharya Plot' of the first differences of the natural logarithms of the histogram frequencies will display a series of negatively sloping linear trends. The number of trends equals the number of component distributions. The positions and orientations of these lines can be used to determine crude estimates of the parameters

(Bhattacharya 1967) by the following formulae :

$$\hat{\mu}_k = \lambda_k + \frac{w}{2}, \tag{2.14}$$

and :

$$\hat{\sigma}_k^2 = w \cot \alpha_k - \frac{w^2}{12}, \tag{2.15}$$

where w is the class interval width, α_k is the angle between the k^{th} line and the negative direction of the x axis and λ_k is the x intercept of the k^{th} line. Unfortunately, this approach does not lead directly to estimates of ϖ_k.

A third approach similar to that of Bhattacharya has been described by Tanaka (1962). The *PDF* for the k^{th} component distribution is multiplied by the estimated number of observations derived from that distribution and then its natural logarithm is taken, giving :

$$\ln f_k(x) n_k = \frac{-(x-\mu_k)^2}{2\sigma_k^2} + \ln\left(\frac{n_k}{\sqrt{2\pi\sigma_k^2}}\right). \tag{2.16}$$

This is a quadratic of the form :

$$h(x) = ax^2 + bx + c, \tag{2.17}$$

where :

$$a = \frac{-1}{2\sigma_k^2}, \tag{2.18}$$

$$b = \frac{\mu_k}{\sigma_k^2}, \tag{2.19}$$

and :

$$c = \frac{-\mu_k^2}{2\sigma_k^2} + \ln n_k - \ln \sqrt{2\pi\sigma_k^2}. \tag{2.20}$$

As a result, the parameters can be estimated by the following :

$$\sigma_k^2 = \frac{-1}{2a}, \tag{2.21}$$

$$\mu_k = \frac{-b}{2a}, \tag{2.22}$$

$$\ln \varpi_k = c + \frac{b^2}{4a} + \ln \frac{\sqrt{-\pi/a}}{n}. \tag{2.23}$$

Tanaka uses quadratic templates to select visual estimates of the coefficients of the quadratic and allow calculation of the parameter values.

All of the above graphical approaches based on *PDF*'s suffer severe estimation problems where overlap between component populations is large. Iterative re-approximation of the parameters through subtraction of the overlapping extrema of the component distributions during estimation is likely to help alleviate these inaccuracies (Titterington et al. 1985).

Alternative graphical approaches for determining the parameters of a mixture of distributions involve cumulative distribution plots. The normal quantile-quantile (Q-Q) plot diagram relates $P(q)$ against $\Phi(q)$, where $0 < q < 1$, $P(\cdot)$ is the empirical distribution function and $\Phi(\cdot)$ is the standard (cumulative) normal. On this plot, observations from a normal distribution define a straight line with slope of 1 and $P(q) = \Phi(q)$. Observations from multi-modal mixtures of overlapping normal distributions plot as multiple sigmoidal curves with inflection points at percentages which define the percentages of the component distributions along the $\Phi(\cdot)$ axis. The means and standard deviations of the component populations cannot be estimated from this plot.

A slight modification of the Q-Q plot is the P-P plot where $\Phi((x_i - \overline{x})/s) - q_i$ is plotted against $(x_i - \overline{x})/s$, where q_i is the cumulative percentile of the i^{th} observation. This produces a similar plot, except that the resulting curve is not monotonic. A horizontal line defines a normal distribution (because q_i is subtracted from $\Phi((x_i - \overline{x})/s)$), and deviations from this indicate the possible presence of multiple distributions (Fowlkes 1979).

An alternative to the Q-Q and P-P plots is the probability graph, where the value of the variable x, instead of its cumulative frequency $P(\cdot)$ is plotted against $\Phi(\cdot)$. On plots of this type, the inflection points of sigmoidal curves also define the percentages of the component populations. Although early workers using this technique resorted to visual estimation of the parameters (Harding 1949; Court 1949; Cassie 1954), straight lines can be derived through simple hand calculator operations and graphical projections (Sinclair 1976). This allows the definition of the individual normal components, and these straight lines plot as asymptotic limits to the sigmoidal curves. Furthermore, since the actual values of the variable are plotted, the mean of each component can be visually defined as the value on the x axis where the straight component distribution line crosses the 50^{th} percentile on the $\Phi(\cdot)$ axis. Similarly, the standard deviation of each component distribution is equal to the difference between values on the x axis which correspond to the intersection of the straight component population line with the 84^{th} and 50^{th} percentiles on the $\Phi(\cdot)$ axis. Thus, probability graphs provide more information than Q-Q or P-P plots (or histograms) because **all** parameters can be estimated directly from the plot.

Use of probability plots in the geological sciences is widespread (Tennant and White 1959; Ageno and Frontali 1963; Williams 1967; LePeltier 1969; Folk 1971; Van Andel 1973; Parslow 1974; Sinclair 1974; Clark 1976; Sinclair 1976; Brazier et al. 1983; Stanley 1984; Stanley and Sinclair 1988).

2.2.3 Method of Moments

Determination of the parameters of the mixture of two normal distributions model which best 'fits' the observed distribution can be accomplished using the method of moments technique (Pearson 1894; Charlier and Wickersell 1924; Cohen 1967). Results using this method are obtained by determining the first 5 central moments (ν_g) for the

observed distribution by the following formulae for the first moment :

$$\nu_1 = \frac{\sum_{i=1}^{n} x_i}{n}, \tag{2.24}$$

and :

$$\nu_g = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^g}{n}, \tag{2.25}$$

for the second through fifth moments. These moments are then equated using the following equations :

$$\nu_1 = \varpi\mu_1 + (1-\varpi)\mu_2, \tag{2.26}$$

$$\nu_2 = \varpi(\mu_1^2 + \sigma_1^2) + (1-\varpi)(\mu_2^2 + \sigma_2^2), \tag{2.27}$$

$$\nu_3 = \varpi\mu_1(\mu_1^2 + 3\sigma_1^2) + (1-\varpi)\mu_2(\mu_2^2 + 3\sigma_2^2), \tag{2.28}$$

$$\nu_4 = \varpi(\mu_1^4 + 6\mu_1^2\sigma_1^2 + 3\sigma_1^4) + (1-\varpi)(\mu_2^4 + 6\mu_2^2\sigma_2^2 + 3\sigma_2^4), \tag{2.29}$$

$$\nu_5 = \varpi\mu_1(\mu_1^4 + 10\mu_1^2\sigma_1^2 + 15\sigma_1^4) + (1-\varpi)\mu_2(\mu_2^4 + 10\mu_2^2\sigma_2^2 + 15\sigma_2^4). \tag{2.30}$$

Solving these simultaneously gives estimates for ϖ, μ_1, μ_2, σ_1 and σ_2. Pearson showed that algebraic manipulation of these equations and elimination of these variables produces the following nonic polynomial in ρ :

$$\begin{aligned} 0 \;=\; & 24\rho^9 + 84c_4\rho^7 + 36\nu_3^2\rho^6 + (90c_4^2 + 72c_5\nu_3)\rho^5 + \\ & (444c_4\nu_3^2 - 18c_5^2)\rho^4 + (288\nu_3^4 - 108c_4c_5\nu_3 + 27c_4^3)\rho^3 - \\ & (63c_4^2\nu_3^2 + 72c_5\nu_3^3)\rho^2 - 96c_4\nu_3^4\rho - 24\nu_3^6, \end{aligned} \tag{2.31}$$

where $c_4 = \nu_4 - 3\nu_2^2$ and $c_5 = \nu_5 - 10\nu_2\nu_3$, the 4^{th} and 5^{th} (statistical) sample cumulants and $\rho = (\mu_1 - \nu_1)(\mu_2 - \nu_1)$. After determining the negative real root (ρ_0) to this equation, we can use the following procedure to determine the 5 parameter estimates.

First, define :

$$\eta = \frac{-6\nu_3\rho_0^3 + 2c_5\rho_0^2 + 9c_4\nu_3 + 6\nu_3^3}{2\rho_0^3 + 3c_4\rho_0 + 4\nu_3^2}, \tag{2.32}$$

and :

$$\lambda = \eta - \nu_3. \tag{2.33}$$

Then by computing $\varrho = \lambda/\rho_0$ and solving the quadratic equation :

$$0 = d^2 - \varrho d + \rho_0, \tag{2.34}$$

(giving roots d_1 and d_2, where $d_1 < 0 < d_2$) and calculating :

$$\beta = \frac{(2\lambda - \nu_3)}{3\rho_0}, \tag{2.35}$$

one can estimate the 5 parameters with the following equations :

$$\hat{\sigma}_k^2 = d_k\beta + \nu_2 - d_k^2, \tag{2.36}$$

$$\hat{\varpi} = d_2/(d_1 - d_2), \tag{2.37}$$

$$\hat{\mu}_k = d_k + \nu_1, \tag{2.38}$$

where $k = 1, 2$.

Cohen (1967) demonstrated that if the variances of each distribution are equal, the nonic equation can be reduced to a cubic, and ρ_0 can be determined analytically (Beyer 1982, p. 9). He recommended starting with this equal variance assumption and iterating to the solution of the 5 equations rather than solving the nonic directly.

Unfortunately, parameter estimates may be not be feasible since the quadratic equation may have no real roots (d_1 and d_2) and the nonic equation may have more than one negative real root, or none at all. Generally, at least one feasible solution can be determined if $\hat{\sigma}_k > 0$ and $0 < \hat{\varpi} < 1$ (Titterington et al. 1985).

Furthermore, this technique is limited to the bimodal case, and thus is not generally applicable to the decomposition of mixtures of distributions in applied geochemistry, because these mixtures may be comprised of more than two distributions. However, examples of applications of this technique in the geological sciences do appear in Martin (1936), Ghose (1970) and Everitt and Hand (1981).

2.2.4 Bayesian Methods

Bayes theorem allows beliefs about a parameter vector Ψ *prior* to observing x to be updated into beliefs about Ψ *posterior* to observing x through the relation :

$$p(\Psi|x) = \frac{L(\Psi)p(\Psi)}{\int_{\Psi} L(\Psi)p(\Psi)d(\Psi)}, \tag{2.39}$$

where $L(\Psi)$ is the likelihood, $p(\Psi)$ is the probability density before observation of x and $p(\Psi|x)$ is the probability density after observing x.

This method for determining estimates of the parameters is not straightforward, unless Ψ consists of a small number of unknown parameters (Titterington et al. 1985). If a large number of parameters are unknown, the problem centers on efficiently integrating Ψ in multiple dimensions (Smith and Makov 1982, Smith et al. 1985). Since distributions with a large number of parameters (> 2) are common in geochemical applications, this approach is not considered further.

2.2.5 Maximum Likelihood

Probably the most popular approach to parameter estimation of mixtures of distributions is the maximum likelihood approach (Hasselblad 1966; Day 1969; Gregor 1969; Sahu 1973; Dick and Bowden 1973). In the geological sciences, this approach has been applied not only to mixtures of normal distributions but to mixtures of Von Mises distributions as well (Jones 1968; James and Jones 1969). Parameter estimates obtained by this method are consistent and asymptotically normally distributed. The raw data maximum likelihood function (*RDML*) of observations from a mixture of distributions :

$$L(\Psi) = \prod_{i=1}^{n} \left[\sum_{k=1}^{v} \frac{\varpi_k}{\sigma_k} \phi(z_{i,k}) \right], \tag{2.40}$$

is maximized to produce the maximum likelihood estimates of the parameters Ψ, where :

$$z_{i,k} = \left(\frac{x_i - \mu_k}{\sigma_k} \right), \tag{2.41}$$

and :

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}. \tag{2.42}$$

The natural logarithm of this function is more often maximized, producing an identical set of estimates because the transformation is monotonic :

$$\ell(\Psi) = \ln L(\Psi) = \sum_{i=1}^{n} \ln \left[\sum_{k=1}^{v} \frac{\varpi_k}{\sigma_k} \phi(z_{i,k}) \right]. \tag{2.43}$$

Maximization, if possible, occurs through simultaneous evaluation of the roots of the partial derivatives of the log-likelihood function (the likelihood equation) :

$$0 = \frac{\partial \ell(\Psi)}{\partial \Psi_k}. \tag{2.44}$$

Because procedures to determine the roots of these partial derivatives are neither rapid nor straightforward, distribution models with large numbers of parameters and data sets with large n may generally take unacceptable amounts of time to iterate to a solution. As a result, a more expedient approximation is commonly utilized to determine the parameter estimates. The raw data are cumulated into class intervals and the class interval frequencies are used to calculate estimates of the parameters. The class interval data maximum likelihood (*CIDML*) function for this approach is :

$$L(\Psi) = \prod_{j=1}^{m} \left[\sum_{k=1}^{v} \varpi_k \left(\Phi(b_{j-1}|\theta_k) - \Phi(b_j|\theta_k) \right) \right]^{n_j}, \tag{2.45}$$

where b_j and b_{j+1} are the lower and upper class interval limits, respectively, Φ_k is the cumulative distribution function of each component population, n_j are the number of observations which fall in the j^{th} class interval and m is the total number of class intervals. The natural logarithm of this function :

$$\ell(\Psi) = \ln L(\Psi) = \sum_{j=1}^{m} n_j \ln \left[\sum_{k=1}^{v} \varpi_k \left(\Phi(b_{j-1}|\theta_k) - \Phi(b_j|\theta_k) \right) \right], \tag{2.46}$$

is likewise maximized, and the result generally requires less computation and fewer iterations than maximization of the *RDML* function.

Unfortunately, both of the above likelihood functions may have numerous local maxima. There exists $n \times v$ unbounded (non-stationary) maxima along the exterior of the parameter space (where $\sigma_k = 0$ and $\mu_k = x_i$) because σ_k occurs in the denominator of the normal density function. The likelihood function $\rightarrow \infty$ with each of these parameter sets as the denominator $(\sigma_k) \rightarrow 0$, thus these solutions to the likelihood equation are pathological (Everitt and Hand 1981). Similarly, maximization can occur if the $0 \leq \varpi_k \leq 1$ or $\sum_{k=1}^{v} \varpi_k = 1$ constraints are not satisfied. Roots of the likelihood equation producing solutions of this type may not be feasible. As a result, the best parameter estimates are found at a local (stationary) maximum in the interior of the parameter space, where all constraints are satisfied ($\sigma_k > 0$, $\sum_{k=1}^{v} \varpi_k = 1$ and $0 \leq \varpi_k \leq 1$). Unfortunately, several local (stationary) maxima of this type may exist, so care must be exercised to ensure that all have been located and the root with the highest likelihood be used to determine the final parameter estimates. For mixtures of normal distributions, this root will produce the true maximum likelihood parameter estimates (McLachlan and Basford 1988).

Numerous algorithms have been proposed to maximize the above likelihood functions. These include :

- the EM Algorithm,
- the Newton Raphson Algorithm, and
- Direct Search Methods.

These will be discussed in terms of the advantages and disadvantages each offers over the other techniques.

2.2.5.1 EM Algorithm

The Expectation-Maximization (EM) algorithm originally was described by Dempster, Laird and Rubin (1977) as a method of treating problems with incomplete data sets. This approach can be used to estimate the parameters of mixtures of normal distributions by assuming that, for each x_i, there exists a column vector $\vec{Y}_i$ of length v, composed of an unknown (missing) indicator variable defining the distribution to which the x_i belongs (all 0's except for a single 1 in the k^{th} entry denoting membership in the k^{th} distribution). The corresponding logarithmic *RDML* function then takes the form :

$$\ell(\Psi) = \ln L(\Psi) = \sum_{i=1}^{n} \vec{Y}_i^T \vec{V}(\varpi) + \sum_{i=1}^{n} \vec{Y}_i^T \vec{U}(\theta), \tag{2.47}$$

where $\vec{V}(\varpi)$ is a column vector of the $\ln \varpi_k$ terms, $\vec{U}(\theta)$ is a column vector of the $\ln \frac{1}{\sigma_k}\phi(z_{i,k})$ and $\vec{Y}_i^T$ is the row vector which is the transpose of $\vec{Y}_i$.

The EM algorithm works by generating, from some initial approximation of Ψ^0, a sequence of Ψ^r estimates such that, by Jensen's inequality :

$$\ell(\Psi^{r+1}) = \ln L(\Psi^{r+1}) \geq \ln L(\Psi^r) = \ell(\Psi^r). \tag{2.48}$$

Thus, successive likelihoods in this series monotonically increase, and optimization is accomplished through an iterative two-step process of, first, estimation and, second, maximization.

The expectation stage involves estimating the log-likelihood value at Ψ^r :

$$\ell(\Psi^r) = \sum_{i=1}^{n} \vec{W}_i(\Psi^r)^T \vec{V}(\varpi) + \sum_{i=1}^{n} \vec{W}_i(\Psi^r)^T \vec{U}(\theta), \tag{2.49}$$

where $\vec{W}_i(\Psi^r)$ is a column weighting (indicator) vector of length k, such that :

$$w_{i,k}(\Psi^r) = \frac{\varpi_k^r \phi(x_i|\theta_k^r)}{p(x_i|\Psi^r)}. \tag{2.50}$$

The weights are thus probabilities of k^{th} distribution membership for the i^{th} observation, given x_i and the parameters Ψ^r.

The maximization step, a consequence of the estimation step, generally requires an iterative approach to solve, but for mixtures of normal distributions, it can be done analytically (McLachlan and Basford 1988). It consists of the following formulae to determine estimates of the parameters :

$$\varpi_k^{r+1} = \frac{\sum_{k=1}^{v} w_{i,k}(\Psi^r)}{n} = \frac{n_k^r}{n}, \tag{2.51}$$

$$\mu_k^{r+1} = \frac{\sum_{k=1}^{v} w_{i,k}(\Psi^r)x_i}{n_k^r}, \tag{2.52}$$

$$\sigma_k^{r+1} = \sqrt{\frac{\sum_{k=1}^{v} w_{ik}(\Psi^r)(x_i - \mu_k^{r+1})^2}{n_k^r}}. \tag{2.53}$$

This two-step process is repeated until the log-likelihood $\ell(\Psi^r)$ does not change significantly, and (hopefully) a maximum has been reached.

Application of the EM algorithm to mixtures of more than two distributions is straightforward. Although convergence can be excruciatingly slow (Redner and Walker 1984), the algorithm will converge to a solution, unless it gets trapped at some saddle point (Dempster, Laird and Rubin 1977; McLachlan and Basford 1988). However, there is no guarantee that this procedure will converge to the correct parameter values. Constraining the procedure to ensure iteration to a global maximum where $0 \leq \varpi_k \leq 1$, $\sum_{k=1}^{v} \varpi_k = 1$ and $\sigma_k = 0$ (with $\mu_k = x_i$) may aid in both speeding up convergence and preventing convergence to a non-feasible solution (Hathaway 1985). Louis (1982) has developed a method for speeding up convergence, as well as extracting the observed Fisher information matrix to estimate the asymptotic standard errors of the maximum likelihood parameter estimates.

The EM approach can also be used for the cumulated data likelihood function because the indicator (weighting) vectors can be assumed to correspond to each of the

individual data values within each class interval, provided that the class intervals are small enough. Errors may occur if this function is applied to cases where the data have been cumulated into large class intervals for more rapid parameter estimation, because the foregoing assumption may not hold.

2.2.5.2 Newton-Raphson

An alternative approach for determining the parameter estimates of a mixture of normal distributions is through use of the Newton-Raphson algorithm. This consists of a gradient search of the log-likelihood surface to locate a maximum (subject to $0 \leq \varpi_k \leq 1$, $\sum_{k=1}^{v} \varpi_k = 1$ and $\sigma_k > 0$) using the following equation :

$$\vec{\Psi}^{r+1} = \vec{\Psi}^{r} - \tilde{\ell}''(\Psi^{r})^{-1}\vec{\ell}'(\Psi^{r}). \tag{2.54}$$

This procedure involves determination of the first and second partial derivatives of the log-likelihood function. These can be determined analytically for both the *RDML* and *CIDML* functions (Appendix B) and lead directly to calculation of the observed Fisher information matrix and estimates of the asymptotic variances of the maximum likelihood parameter estimates.

This numerical approach suffers from shortcomings similar to those of the EM algorithm in that no easily implementable procedure has been devised which prevents iteration toward a maximum where $\sigma_k = 0$ (with $\mu_k = x_i$), $\sum_{k=1}^{v} \varpi_k \neq 1$ or $0 \not\leq \varpi_k \not\leq 1$. Convergence to non-feasible solutions is common, especially where initial parameter estimates differ substantially from their true values. Convergence can be rapid (quadratic) if close to the solution; however, if a large number of component distributions are present (and thus a large number of parameters need to be estimated) a large, time-consuming matrix inversion is required.

2.2.5.3 Direct Search

Direct search algorithms systematically evaluate the objective function to locate its optimum. One popular approach is the SIMPLEX procedure (Nash 1979; Caceci and Cacheris 1984). This approach can be applied to either the *RDML* or *CIDML* functions and can be used to determine parameter estimates of any number of mixtures of distributions.

The procedure begins by defining $3v$ sets (one more than the number of parameters to be estimated) of parameter estimates to define a $(3v-1)$-dimensional polygon or SIMPLEX. These estimates (vertices) must span the full parameter space to be searched. New vertices are selected such that they produce objective function values which are higher than the current lowest-valued vertex, until a solution is reached.

Initially, the objective function is evaluated at each vertex and the the lowest- and highest-valued vertices are determined. The centroid of the vertices which do not have the lowest objective function value is also calculated. Then a series of reflections is made from the lowest-valued vertex through the centroid of the other vertices (see Figure 2.7 for graphical example of minimization). This reflection may be positive, producing a new vertex some multiple (generally one or two) times the 'centroid-to-lowest-valued vertex distance' past the centroid, or negative, such that the new vertex lies between the highest-valued vertex and the centroid.

The objective function values for each of these possible new vertices are evaluated and the highest one is then chosen. If a positive one-fold reflection is chosen, the SIMPLEX merely flips through itself toward the parameter solution. If a positive greater than one-fold reflection is chosen, the SIMPLEX flips through itself toward the parameter solution, becoming larger in the process. If a negative reflection is selected, the SIMPLEX collapses in on itself toward the parameter solution, becoming smaller

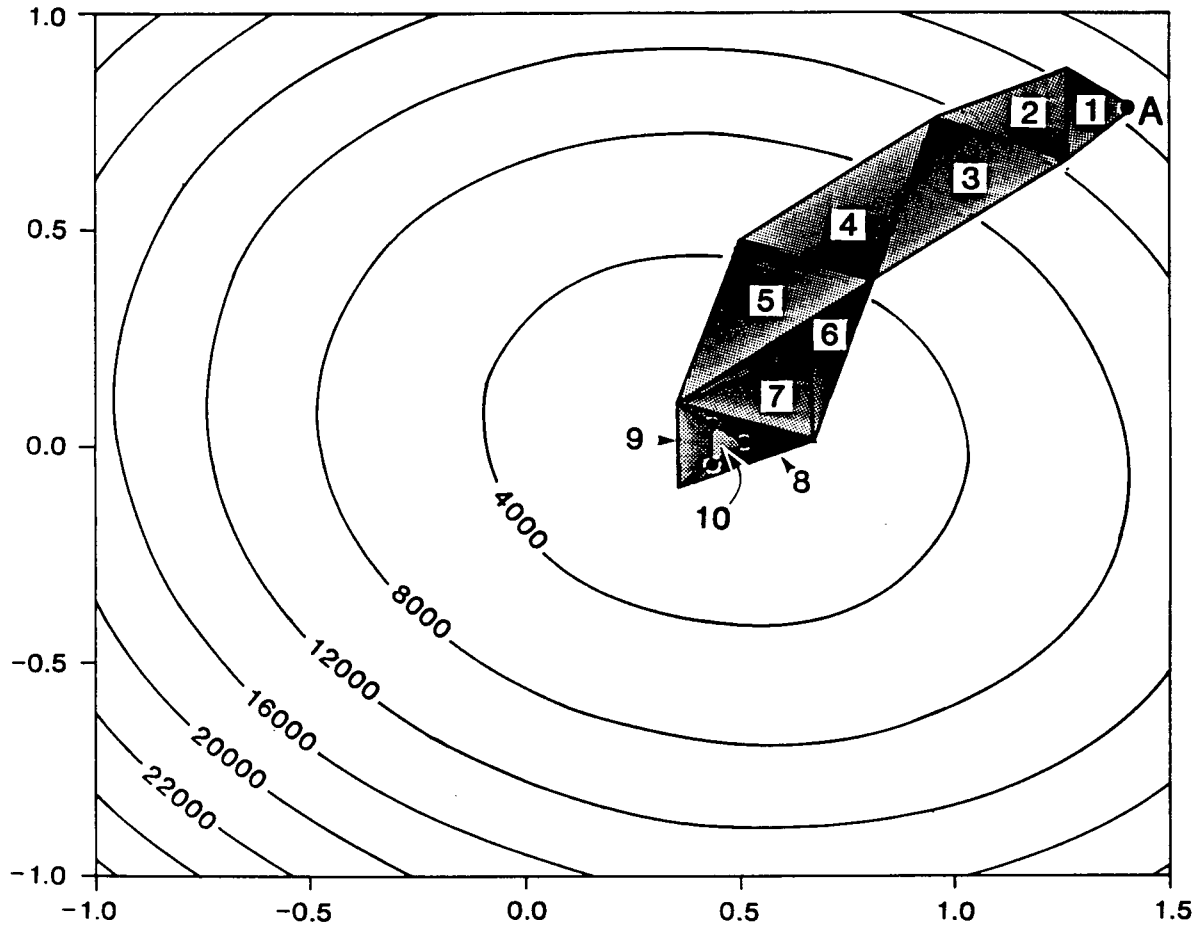


Figure 2.7: Two Dimensional Schematic Representation of SIMPLEX Algorithm Searching for Optimum Set of Parameter Values on an Objective Function Surface

The simplex has 3 vertices because two parameters are to be optimized. The right vertex of triangle (simplex) # 1 (A) represents the initial guess of the parameter values. Triangles # 2 and # 3 are double reflections, enlarging the simplex to speed optimization where the optimization surface is steep. Triangles # 4, 5 and 6 are single reflections, moving the parameter estimates toward the optimal solution. Triangle # 7 is a half reflection to reduce the size of the simplex as the optimation surface becomes shallower (close to the solution). Triangle # 8 is another single reflection and triangle # 9 is another half reflection. Triangle # 10 has been 'shrunk' to further reduce the size of the simplex and allow convergence to a set of optimal parameter values.

in the process. If none of these new vertices is higher than the current lowest-valued vertex, the entire SIMPLEX is shrunk by moving all of the lower-valued vertices one half the distance toward the highest-valued vertex.

In this way, the SIMPLEX will expand in size where the slope of the objective function is smooth (regular), shrink where the slope is irregular, and collapse on the solution where the SIMPLEX bounds the optimal parameter estimates. This algorithm thus reacts appropriately to the local objective function surface, avoiding numerical instability problems. A representation of this procedure is presented in Figure 2.7 for a two parameter case to demonstrate the various decision points and features of the SIMPLEX method.

The SIMPLEX method can be constrained easily to satisfy the limiting conditions of $\sigma_k > 0$ (with $\mu_k = x_i$), $\sum_{k=1}^{v} \varpi_k = 1$ or $0 \leq \varpi_k \leq 1$. Its convergence, although slower than the Newton-Raphson algorithm, is slightly faster than the EM algorithm because it has the desirable property of bounding and then collapsing onto the solution instead of creeping up towards it monotonically. Thus, early termination of the iterative process does not by necessity produce wildly inaccurate (biased) estimates. Unfortunately, no direct calculation of the observed Fisher information matrix is obtainable using this technique, so estimates of the asymptotic standard errors of the maximum likelihood parameter estimates may be calculated at the solution.

2.3 Threshold Selection and Classification

Once the parameters of the mixture of distributions model have been estimated, thresholds are chosen for classification of the data into groups corresponding to the multiple component distributions present in the data set. Two thresholds may be chosen at the mean plus and minus two standard deviations for each component distribution. This

produces $2v$ thresholds and defines upper and lower limits for each component distribution within which, on average, 95 % of the data from the component distribution occur.

The frequency distributions of adjacent component distributions may overlap to varying degrees (Figure 2.8). Those component distribution pairs where $\mu_B + 2\sigma_B > \mu_A - 2\sigma_A$ are considered to be **overlapping** cases (where B represents the distribution with the lower mean and A represents the distribution with the higher mean). In these cases, a total of more than 5 % of data will be misclassified using any threshold, assuming the theoretical *PDF* model is valid. Those component distribution pairs where the $\mu_B + 2\sigma_B \leq \mu_A - 2\sigma_A$ are considered to be **non-overlapping** cases, and less than 5 % of the data theoretically will be misclassified using either of these thresholds (or an intermediate one).

Obviously, the application for which the classification procedure is used will define the type of criteria used for threshold selection. In some cases, a minimum of one type of classification error (either of omission or inclusion) will be the optimal condition. These require the assignment of weights (ω- subjective criteria) to indicate the relative importance of the errors of omission and inclusion. In other applications, the total number of classification errors is minimized, and the 'error' weights are equal. In this study, the equal 'error' weights case will be used because of its generality and objectivity.

Figure 2.9 demonstrates how the use of different thresholds can result in varying amounts of misclassification. Using the parameters of a mixture of normal distributions, one can calculate the theoretical amount of misclassification of data from the distribution with the lower mean (B) in the distribution with the higher mean (A; due

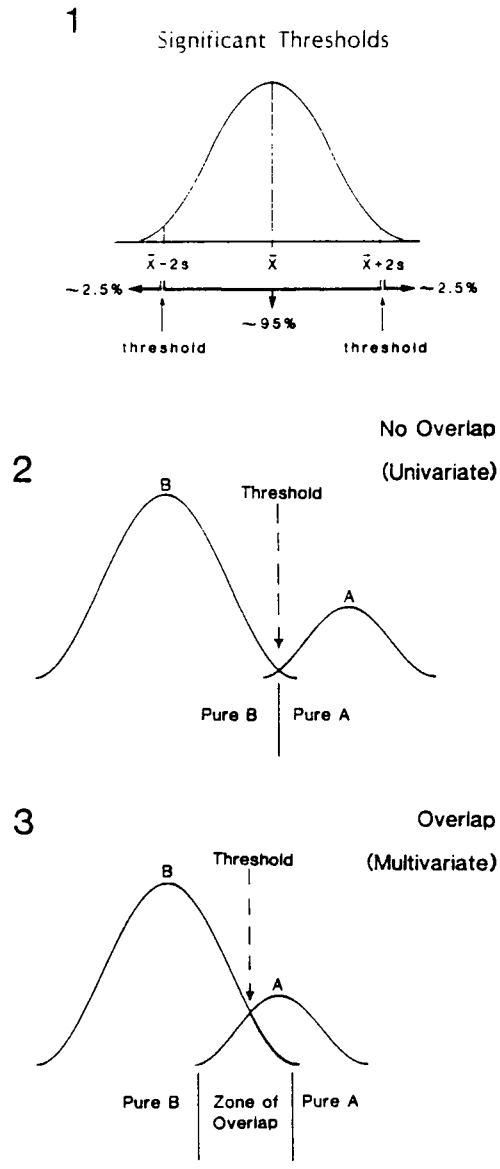


Figure 2.8: Examples of Overlapping and Non-Overlapping Mixtures of Normal Distributions

Thresholds have been chosen at the $\bar{x} \pm 2s$ for each component distribution (1).

Non-overlapping mixtures of distributions (2) generally do not require additional analysis to classify the data. The threshold separates essentially 'pure' (un-polluted) data ranges composed of data from only population A or population B.

Overlapping mixtures of distributions (3) may require additional information (other variables) because a single threshold cannot classify the data adequately. In this case, two thresholds define two 'pure' data ranges (for population A and B) and a zone of overlap.

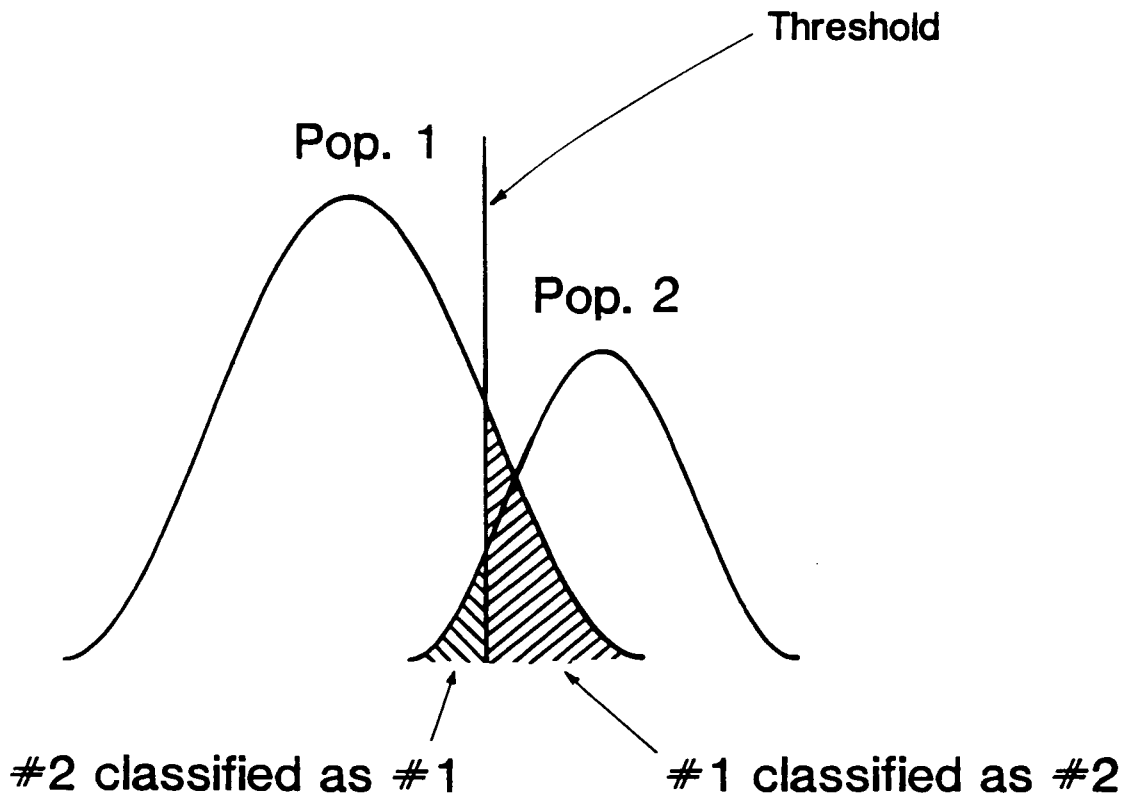


Figure 2.9: Total Probability of Classification Error Defined by a Threshold is Equal to the Area Bounded by the Tails of Each Component Distribution and the Threshold

The ruled area defines the amount of data misclassification for the threshold. The NW-SE ruled area represents the errors of inclusion (# 2 classified as # 1) while the NE-SW ruled area represents the errors of omission (# 1 classified as # 2).

to omission) :

$$e_B = \frac{\varpi_B}{\sigma_B}\Big(1 - \int_{-\infty}^{t_B} \phi(x|\theta_B)dx\Big), \tag{2.55}$$

and of data from distribution A in distribution B (due to inclusion) :

$$e_A = \frac{\varpi_A}{\sigma_A} \int_{-\infty}^{t_A} \phi(x|\theta_A)dx, \tag{2.56}$$

where t_A and t_B values corresponding to the threshold which has been standardized for each component distribution and e_A and e_B are the errors of inclusion and omission, respectively. Since the total number of observations misclassified is the weighted sum of these errors :

$$e_T = \omega_1 e_B + \omega_2 e_A, \tag{2.57}$$

where ω_1 and ω_2 are the weights assigned to the different types of errors (for purposes in this study, they are considered to be unity), then differentiating this equation with respect to the threshold (t) and equating to zero gives :

$$\frac{\partial e_T}{\partial t} = \frac{\omega_2 \varpi_A}{\sigma_A}\phi(t|\theta_A) - \frac{\omega_1 \varpi_B}{\sigma_B}\phi(t|\theta_B) = 0, \tag{2.58}$$

the total amount of misclassification is minimized where :

$$\frac{\omega_2 \varpi_A}{\sigma_A}\phi(t|\theta_A) = \frac{\omega_1 \varpi_B}{\sigma_B}\phi(t|\theta_B). \tag{2.59}$$

This result can be determined easily and rapidly through interval-halving the region between the two means of the distributions until the desired precision is attained. Visual estimates of the optimal threshold using histograms are possible only if a true anti-mode exists; thus, the above numerical computation is preferred.

2.4 Probability Graph Software

The PROBPLOT program is an interactive, graphical computer program designed and written to perform the above decomposition and classification (parameter estimation

and threshold selection) on a mixture of normal distributions (Stanley 1987). It is similar to several mainframe programs which also perform multi-modal frequency distribution decompositions. Programs which have been used in geological and geochemical applications to decompose mixtures of normal distributions include ROKE (I. Clark 1977), GETHEN (M.W. Clark 1977a, 1977b) and DISCRIM (Bridges and McCammon 1980).

2.4.1 Hardware Requirements

Version 1.1 of PROBPLOT is written in TURBO Pascal (version 3.02A) using the TURBO Pascal Graphics Toolbox (version 1.07A) for the IBM-PC (and compatible) family of micro-computers. Graphical and tabular output is designed to be printed on an Epson-compatible printer. The CGA (640 × 200), EGA (640 × 350) and Hercules (720×350) graphics card formats are all supported and separate versions of the program support high precision floating-point operations in hardware (with the 8087, 80287 or 80387 numerical co-processor chip) or software.

The three major sections of the PROBPLOT program allow the user to perform the multi-modal frequency decomposition by :

- determining the general form of the theoretical frequency distribution model (generally done by previewing histograms and probability plots displaying only the raw data, choosing the number of component distributions, and estimating the parameters of each),

- optimizing this frequency distribution model and decomposing the data distribution into its component populations (done by either maximizing the *RDML* or *CIDML* function, minimizing the χ^2 function, or selecting parameters visually to produce estimates of the parameters of a distribution model which fit the

frequency distribution of the data), and

- selecting thresholds to partition the data into groups representative of these component populations (done by minimizing the total misclassification error (e_T), assigning thresholds to some multiple of standard deviations away from the mean of each component distribution, or selecting thresholds visually).

The program has been written in a general way to allow its use in any field where frequency data analysis is required. Although this study emphasizes the analysis of geochemical data, the output from this program is non-specific, and any frequency data may be evaluated with it.

2.4.2 Program Capabilities

The PROBPLOT program can analyze data files containing up to 45 variables and 3500 observations. Data values must be real numbers, but those coded as '0' or '0.0' are considered as missing values and are not considered. Summary statistics are produced which report the number of missing observations, the number of true observations, the number of class intervals, the time and date, the mean and standard deviation of the data (and their anti-logarithmic equivalents, if the data have been transformed logarithmically), the coefficient of variation, skewness, minimum, maximum, 1^{st} quartile, median (2^{nd} quartile) and 3^{rd} quartiles. The variable evaluated, its units, the data transform (arithmetic [none] or logarithmic), the number of component distributions ($1 \leq v \leq 5$) fitted to the data, the estimates of the means, standard deviations and percentages of each of the component distributions, and the threshold values (or their anti-logarithmic equivalents if the data are transformed logarithmically) may also be output.

2.4.2.1 Histograms and Probability Plots

Data may be screened on input to filter outlying values. The default acceptable data limits consist of a minimum value of 0.0001 and maximum value of 99999.9999; however, these may be changed by the user. Values outside the acceptable range may exist in the data file, but will not be input.

The PROBPLOT program automatically determines the number of class intervals (or bins) to be used, based on the number of observations in the data set. The default number of class intervals is determined by the following formula :

$$m = 10 \times \log_{10} n, \tag{2.60}$$

where m is the number of class intervals and n is the number of observations (compare Burgess's Rule where $m = 1 + (3.3 \times \log_{10} n)$,(Garrett 1984) and the maximum entropy approach (Perillo and Marone 1986a, 1986b)). The formula above results in a maximum of 36 class intervals for any data set less than 3500 observations, restricting the size of all histograms to one page. The user may modify the number of class intervals manually, either to increase the resolution of the cumulative data, to create class intervals with integer boundaries, to remove class intervals containing multiple reporting values, or to eliminate vacant class intervals. The maximum and minimum number of class intervals possible is 36 and 5, respectively. If a 'user defined' number of class intervals is used, new histograms and probability plots may be generated with this different number of class intervals.

Class intervals are of equal size and distributed evenly across the entire range of the data considered (compare the maximum entropy approach of Full et al. 1984). The lower limit of the lowest class interval is one half a class interval length less than the minimum value in the data set, and the upper limit of the highest class interval is one half a class interval length greater than the maximum value in the data set.

The PROBPLOT program cumulates the data values into class intervals for use in the histograms and probability plots. This reduces the number of points on these plots, making underlying frequency distribution trends easier to recognize. It also allows application of the *CIDML* or χ^2 functions, which reduces the number of calculations required to obtain optimal estimates of the parameters which describe the 'best fitting' cumulative frequency distribution model. An observation is included in a class interval if its value is greater than or equal to the lower class interval limit and less than the upper class interval limit. Relative frequencies for each class interval are calculated by dividing the number of observations in the class interval by the total number of observations (n).

Cumulative relative frequency percentage data are not simply the sum of all class interval frequencies up to and including the current class interval. This is because the resulting cumulative frequency percentages will be different if the cumulation is made from the lowest to highest class interval, or vice versa. As a result, a slightly different computational alternative is used resulting in unique cumulative relative frequency percentages which satisfy all of the desirable characteristics of cumulative relative frequency data.

The following formula is used to calculate the cumulative relative frequency percentages :

$$c_j = 100 \times \left(\frac{\sum_{j=1}^{c} n_j + 0.5}{n+1}\right) \tag{2.61}$$

where c_j is the cumulative relative frequency percentage, $\sum_{i=1}^{c} n_j$ is the total number of observations up to (or down to) and including the current class interval (c), and n is the total number of observations in the data set (compare other formulae of Garrett 1984; Hoffman 1986, p. 23).

This formula results in the same cumulative frequency values whether the data are

cumulated from low to high or high to low, and the resulting distribution has its median value at the 50^{th} percentile. In effect, this formula compresses (biases) the frequency data away from the tails of the distribution (0^{th} and 100^{th} percentiles). This bias may be quite large for small data sets ($n = 20$ has a maximum bias of 2.38 %), but for larger data sets of the size where a probability plot analysis may be most useful, the resulting bias is insignificant ($n = 200$ has a maximum bias of 0.25 % and $n = 1000$ has a maximum bias of 0.05 %). Since the *CIDML* and χ^2 functions are based on the individual class interval frequencies, and not the cumulative frequencies, use of this formula is merely to aid in visual parameter approximation and has no effect on any of the numerical parameter estimation procedures.

Histograms (Figure 2.10) display the upper and lower limit of each class interval (the upper limit is on the same line, the lower limit is on the line immediately above), the frequency and cumulative frequency percentages, and stars to represent the number of observations in the class interval. With larger data sets (≥ 200) the number of observations represented by each star is increased to prevent the number of stars from running off the page. The formula to calculate the number of observations per star is :

$$n_{star} = \frac{m}{\sqrt{((n/100) + 1)}}, \tag{2.62}$$

where n_{star} is the number of observations per star and m is the number of class intervals. The number of observations represented by each star is indicated below the histogram.

```
15:57:13                                                    05/29/87
                          Daisy Creek Soil Grid

################################################################################
SUMMARY STATISTICS and HISTOGRAM                      LOGARITHMIC VALUES

 Variable =      CU          Unit =      PPM                N =   247

     Mean =    1.6188         Min =    1.0000   1st Quartile =    1.3273
Std. Dev. =    0.4017         Max =    2.9881         Median =    1.5051
     CV % =   24.8135    Skewness =    1.2436   3rd Quartile =    1.7924

================================================================================
  %    cum %    antilog   cls int   (# of bins = 24  -  bin size =   0.0864)
-----  -----  ---------  -------  ----------------------------------------
 0.00   0.20      9.053   0.9568
 0.81   1.01     11.046   1.0432  *
 2.83   3.83     13.479   1.1297  ****
 5.26   9.07     16.447   1.2161  ********
12.96  21.98     20.070   1.3025  ******************
 9.31  31.25     24.489   1.3890  *************
15.38  46.57     29.883   1.4754  ***********************
10.93  57.46     36.463   1.5619  ****************
 9.31  66.73     44.494   1.6483  *************
 5.67  72.38     54.292   1.7347  ********
 4.86  77.22     66.249   1.8212  *******
 3.24  80.44     80.838   1.9076  *****
 2.83  83.27     98.641   1.9941  ****
 3.64  86.90    120.364   2.0805  *****
 2.02  88.91    146.871   2.1669  ***
 0.81  89.72    179.215   2.2534  *
 1.62  91.33    218.683   2.3398  **
 1.21  92.54    266.842   2.4263  **
 2.02  94.56    325.608   2.5127  ***
 1,21  95.77    397.314   2.5991  **
 1.62  97.38    484.813   2.6856  **
 1.62  98.99    591.580   2.7720  **
 0.00  98.99    721.861   2.8585
 0.40  99.40    880.832   2.9449  *
 0.40  99.80   1074.812   3.0313  *
--------------------------------------------------------------------------------
                                0         1         2         3         4

                    Each "*" represents approximately 1.7 observations.

################################################################################
```

Figure 2.10: Example of Histogram Output from the PROBPLOT Program (From a soil survey data set of 247 geochemical samples described in Stanley 1984).

Probability plots are generated on the graphics screen which has a probability scale ranging from +2.5 to -2.5 standard deviations (approximately the 99.5 and 0.5 percentiles) on the abcissa, and the variable (or log-variable) scale on the ordinate. Rounded maximum and minimum ordinate limits bound the maximum and minimum values of the variable by ± approximately 10 % of the range.

2.4.2.2 Distribution Model Selection and Optimization

The PROBPLOT program, after allowing the user to preview arithmetic and logarithmic histograms and probability plots, allows the user to define a theoretical cumulative frequency distribution model to match the cumulative frequency distribution of the data. This is done by selecting a data transformation (arithmetic [none] or logarithmic) which creates a cumulative frequency distribution which can be fitted by a mixture of normal distributions model. Sometimes, the user may wish to evaluate data using both data transforms, comparing the implications and conclusions of the results before deciding which transform is the most appropriate.

In addition to data transformation, the amount of data truncation can be accounted for in the PROBPLOT program (Sinclair 1976). Both upper and lower truncation can be accommodated, but not both simultaneously. If a data truncation correction is made, the program will not allow use of the numerical parameter optimization procedures described below to obtain the 'best fit' cumulative frequency distribution. This is because truncation correction is made on the cumulative frequency data, not the relative frequency data. Thus, because the numerical parameter optimization procedures use the relative frequency data, numerical optimization should not be employed. A visual fitting option is still available in these cases.

The PROBPLOT program limits the user to a maximum of 5 component populations in the theoretical multi-modal normal distribution model (the minimum is 1

component population – a normal or lognormal distribution). This limitation generally does not restrict the user, because real data sets with a larger number of modes are rare, and if present, probably should be broken into subsets of smaller size, based on categorical criteria. Furthermore, a larger number of component populations in the theoretical distribution model would significantly increase the time required to determine the theoretical distribution model and to perform the numerical parameter optimization procedures.

The user, after selecting a data transform and possibly correcting for truncation, is asked to determine how many component populations exist in the data set. This decision should be based on the previously examined histogram and probability plots. The program then prompts for the percentiles where the appropriate inflection points occur. These define the relative amounts of the component populations, and allow the program to determine initial, provisional estimates of the mean and standard deviations of each component population.

These parameters are calculated in the program by partitioning the data into subsets bounded by the percentiles where the inflection points occur. The mean and standard deviation of each of the subsets are used to approximate the parameters of each component population. If component populations are not significantly overlapping, these estimates generally define a reasonably acceptable cumulative frequency distribution fit of the raw data; however, if the populations overlap substantially, these estimates may be biased. In either case, these parameter estimates are generally close enough to the true parameter values to act as initial estimates for any subsequent optimization procedures. Several authors (Hasselblad 1966; Everitt and Hand 1981; Titterington et al. 1985) have suggested using the 'truncated normal distribution technique' described by Hald (1949) and Cohen (1950, 1957, 1959, 1961) to determine initial parameter estimates (see Chapter 5); however, the initial parameter estimation technique described

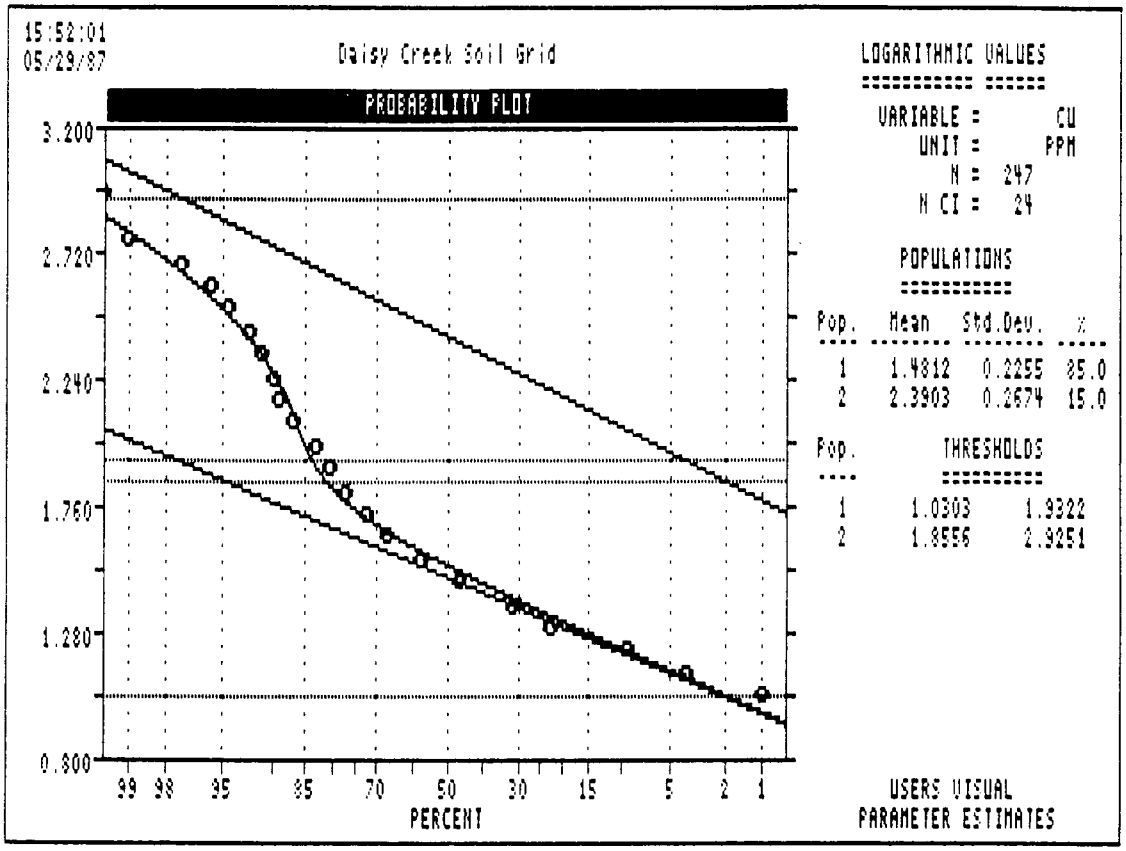


Figure 2.11: Example of Probability Graph Output from the PROBPLOT Program (From soil survey data set of 247 geochemical samples described in Stanley 1984).

above has been found to be adequate for most situations.

The PROBPLOT program then plots the cumulative frequency points on the probability plot and draws the curve defined by the current population parameters. It also draws the straight lines which are defined by the individual component populations. Then, the program allows the user to modify these provisional parameter estimates to obtain a new cumulative frequency distribution fit (Figure 2.11).

After allowing the user to estimate the form of a cumulative frequency distribution model, the PROBPLOT program offers three methods by which to optimize this distribution model through numerical iteration. The three functions are :

- Minimum χ^2 Optimization on Class Interval Data,

- Maximum Likelihood Optimization on Class Interval Data,
- Maximum Likelihood Optimization on Raw Data.

These procedures offer various advantages and disadvantages, basically differing in the way they trade speed for accuracy.

The χ^2 function is minimized to obtain 'optimal estimates' and is quite rapid. Unfortunately, if one of the component populations does not comprise a substantial proportion of the data set, or if there is a large number of class intervals with a very small number of observations in each, the mean and standard deviation estimates will be very biased. Although rapid, the behavior of the objective function iteration path can be unstable, giving spurious results or taking varying amounts of time to determine the optimal solution if initiated from different parameter 'seed' values. This is generally the fastest optimization technique (about 25 % faster than the *CIDML* function optimization) but, in general, provides the least visually acceptable 'best fit'. Tests to determine the significance level of the model fit (χ^2 tests) are not supported (the χ^2 value is not output) because of the general insensitivity (lack of power) of the test to this application (Everitt and Hand 1981; McLachlan and Basford 1988).

The logarithmic *CIDML* function optimization is also a fairly rapid technique, but generally produces a more visually acceptable 'best fit' than the χ^2 technique. Iteration toward a solution is much more stable than iteration with the χ^2 technique. The maximum log-likelihood value is output if the function is maximized to allow the application of a likelihood ratio test statistic (Johnson and Wichern 1982). This allows comparison of distribution models with different numbers of component populations to determine if an additional component distribution in the *PDF* model has statistical significance. It cannot be used to determine if a logarithmic transform model should be favored over an arithmetic one and must be performed with distribution models defined with the

same number of class intervals.

This likelihood ratio test, is similar to most statistical tests in that it compares the null hypothesis against an alternative. In this test, the significance of an additional mode or modes (γ) in the distribution model is determined by subtracting the log-likelihood value of the $(v - \gamma)$-modes distribution fit from the log-likelihood value of the v-modes distribution fit. If this difference is multiplied by :

$$\frac{-2(n - 2 - \frac{v}{2})}{n}, \tag{2.63}$$

(given a sufficiently large number of observations), the result is distributed approximately χ^2 (Wolfe 1971). Thus, the approximate significance of the additional mode(s) can be ascertained by comparison with critical values on a χ^2 table, although the power of this test is reduced if there are small Mahalanobis Distances ($\Delta^2 < 2$) between the component distributions (Everitt and Hand 1981; McLachlan and Basford 1988). Degrees of freedom for this test are equal to the absolute difference in parameterized degrees of freedom (the number of parameters minus one for each model), which is equal to two times the difference in the number of modes in each distribution model (γ).

The logarithmic *RDML* function (Titterington et al. 1985) is by a large degree, the slowest optimization technique (100 to > 500 % slower than the other two techniques depending on the size of the data set); however, its resulting parameter estimates are, by definition, the 'best fit' estimates. It also records the log-likelihood value of the best fit parameters, allowing the application of a likelihood ratio test to determine model significance. This technique produces the 'best' fit of the model to the data, if the *PDF* model is appropriate (McLachlan and Basford 1988).

In general, the basic nature of all of these fitting procedures is a weighted optimization. The functional form of the resulting cumulative frequency distributions are

governed more by the value ranges where more data are present. These are also, by consequence, the value ranges where additional, randomly obtained data will have the least effect in altering the form of the distribution. As a result, the tails of the distribution, as well as the anti-modes, will have a lower influence (weight) in controlling the form of the cumulative frequency distribution, and the curve will be less likely to go precisely through the cumulative frequency points in these regions. This is appropriate because the tails comprise regions on the probability plots where the most deviation (scatter) from a smooth curve will occur because of the relative sparsity of data.

In all three optimization functions, the successive iterations used to obtain optimal parameter estimates are made using the SIMPLEX algorithm, because it is easily adapted to both minimization and maximization. In addition to the use of the SIMPLEX iteration algorithm, a set of nested loops is used which significantly speeds up iteration, preventing the result from temporarily 'moving' away from the solution. This is done by allowing the parameters to vary in two sets, optimizing one set while holding the other set constant, then fixing the first set and allowing the second to vary. The means and standard deviations (the normal parameters) comprise the first set, whereas the component population percentages (the mixture parameters) comprise the second set.

Iteration proceeds for a series of loops, until a solution is reached, or until terminated by the user (whichever comes first) and then the program plots the curve defined by the current parameter estimates and the straight lines which are defined by each component population. This allows the user to make sure that, if only intermediate parameter estimates have been produced, irregular objective functions haven't led the algorithm off toward a unfeasible solution. After this preview, the user may modify the parameters by a variety of techniques, if necessary, and then restart the iteration, possibly with a different optimization function, continuing until a satisfactory solution

is obtained.

Stopping criteria for each optimization function is achieved when all SIMPLEX vertices have both spatial separation and objective function values within 0.5 % of each other. This precision level may be modified by the user. The elaborate iteration procedure described above has been found to be the most reliable route to a set of acceptable parameter estimates based on both experiential and theoretical grounds (the two groups of parameters result in stable iteration paths within each group, but do not iterate in a stable manner if optimized collectively).

2.4.2.3 Threshold Selection

After the user is satisfied with a cumulative frequency distribution model, an option to choose thresholds that separate the groups defined largely by population affinity is presented. The user may modify these thresholds by three methods. These include :

- imposing user defined 'custom' thresholds, perhaps to obtain integer or rounded number thresholds,
- choosing a different standard deviation multiplier for the mean plus or minus some multiple of standard deviations criteria (default = 2), or
- choosing an option which selects a single threshold between two component distributions which defines the theoretical minimum number of classification errors (the sum of the errors of omission and inclusion).

This last technique will create 1 threshold for a two population model, 3 thresholds for a three population model, 6 thresholds for a four population model, and 10 thresholds for a five population model. Not all of these thresholds may be significant in that thresholds separating non-adjacent populations may be meaningless if an intermediate population exists between them. Once these thresholds have been determined, they can be used to classify the frequency data according to the multi-modal distribution model which has been fit to the data.

After parameter optimization and threshold selection, the PROBPLOT program outputs the summary statistics describing the 'best fit' parameters and thresholds (Figure 2.12).

```
15:53:08                  Daisy Creek Soil Grid                  05/29/87

##########################################################################

      PARAMETER SUMMARY STATISTICS FOR PROBABILITY PLOT ANALYSIS

                    Data File Name = DAISY.DAT

   Variable =       CU       Unit =     PPM              N =   247
                                                      N CI =    24

      Transform = Logarithmic          Number of Populations = 2

   # of Missing Observations = 0.

==========================================================================

   Users Visual Parameter Estimates

          Population      Mean       Std Dev     Percentage
          ----------  ----------  ----------  ----------

              1         30.284  -    18.019        85.00
                                +    50.897
              2        245.661  -   132.729        15.00
                                +   454.683


==========================================================================

   Default Thresholds

   Standard Deviation Multiplier = 2.0

                Pop.        Thresholds
                ----        ----------

                  1     10.722     85.539
                  2     71.712    841.551

##########################################################################
```

Figure 2.12: Example of Summary Statistic Output from the PROBPLOT Program (From soil survey data set of 247 geochemical samples described in Stanley 1984).

Chapter 3

Likelihood Function Comparison

"Scratch the surface, and if you are really lucky, you'll find more surface."

Richard Avedon (1975)

"The aims of scientific thought are to see the general in the particular and the eternal in the transitory."

Alfred North Whitehead (1956)

Use of the logarithmic *CIDML* function instead of the logarithmic *RDML* function by the PROBPLOT program expedites numerical optimization. Each iteration of the logarithmic *CIDML* function, with m class intervals, involves $2m$ integrations of the mixture of normals *PDF*, m subtractions, m logarithm calculations, m multiplications, and $m-1$ additions. Alternatively, each iteration using the logarithmic *RDML* function, with n observations, involves n mixture of normals density calculations, n logarithm calculations and $n-1$ additions of these log-densities. If $m \ll n$, a condition common in geochemical applications where n is generally large and sampling, preparation and analytical errors have magnitudes which preclude use of small class intervals (large m), optimization with the *CIDML* function instead of the *RDML* function can result in a substantial reduction of calculation time.

However, the use of the *CIDML* function does have related costs. These costs can potentially take the form of a bias in the resulting parameter estimates. Although the

RDML function uses the raw data, the *CIDML* function does not. Rather, information is lost when the raw data are placed into class intervals because knowledge of where the raw data occur within the class intervals is not used. Potentially, this can result in a parameter estimation bias.

Clearly, as the width of the class interval is reduced, the amount of information lost is reduced because each class interval can more precisely represent the raw data within it. In fact, as the size of the class interval $\rightarrow 0$, the *CIDML* function converges to the *RDML* function, and no bias results.

Most applications in applied geochemistry involve large data sets and thus can benefit substantially from use of the *CIDML* function to determine the parameter estimates of the distribution. However, in this study, data classification techniques involve only the *RDML* function because of its optimal characteristics. As a result, a quantitative comparison of the results of both the *RDML* and *CIDML* functions must be made to determine the extent of the bias produced through use of the *CIDML* function and allow the application of these results to routine applied geochemistry data analysis.

Any comparison of the *RDML* and *CIDML* functions must involve multi-modal data sets with *known* parameter values. Thus, **although actual data sets from applied geochemistry case histories may be used, any conclusions resulting from analysis of these data will necessarily be less conclusive because the parameters defining the component distributions are not known.** Instead, stochastically generated data sets should be used with parameters which are known, in order to allow quantitative determination of the amount of bias produced by both the *RDML* and *CIDML* functions. Furthermore, a variety of *PDF* structures must be tested in order to evaluate how the bias changes with different parameter values.

Although this comparison considers only mixtures of two univariate normal distributions, other evaluations in this study consider multivariate distributions (Chapter 5). Therefore, a general procedure for generating multivariate normal distributions is presented below.

3.1 Stochastic Data Set Generation

Data sets comprised of mixtures of normal distributions were generated using a linear congruential random number generator. Uniform random integers were produced using the algorithm described by Ahrens and Deiter (1973) and Sedgewick (1983), whereby a vector of integers is generated using the following formula :

$$a_i = (a_{i-1}b - 1) \bmod g, \tag{3.64}$$

where $g = 16384$, the largest power of 2 which can be represented as an integer (16 bits) in Turbo Pascal, Version 3, and $b = 7821$ (one power of 10 less than g ending in $w21$, where w is even; Knuth 1981). Starting with some seed value a_0, this produces a sequence of uniform pseudo-random variates on the integers between 0 and 16383, inclusive. These are then each divided by 16383 to produce an approximately uniformly distributed variate with values between 0 and 1, inclusive ($U(0,1) \sim U$ on $\frac{0}{16383}$, $\frac{1}{16383}$, $\frac{2}{16383}$, ..., $\frac{16383}{16383}$ because 16383 is large). The uniform random variates ($\frac{a_1}{16383}$, $\frac{a_2}{16383}$, ..., $\frac{a_n}{16383}$) are stored in a vector with dimension equal to the smallest even integer greater than the number of variables to be generated. This vector of uniform random variates $\{u_j\}_{j=1}^{2p}$ can then be transformed into a vector of normal random scores using the following formulae (Box and Muller 1958; Ahrens and Deiter 1973) :

$$z_1 = \cos(2\pi u_2)\sqrt{-2\ln u_1}, \tag{3.65}$$

$$z_2 = \sin(2\pi u_2)\sqrt{-2\ln u_1}. \tag{3.66}$$

If u_1 and u_2 are two independent uniform random scores, then z_1 and z_2 are independent normal scores which are distributed $N(0,1)$.

A mean vector and covariance matrix for the population must then be specified. The statistical sample means and covariance matrix of the resulting normal random variables will approximate these population statistics. Production of the multivariate normal distributions is accomplished using a multivariate analog of the univariate z-score formula :

$$r = \mu + \sigma z, \tag{3.67}$$

where z is a z-score, μ and σ are the mean and standard deviation of the population, and r is a normal score.

Any σ can be thought of as the square root of the univariate covariance matrix. To obtain the multivariate equivalent statistic, the multivariate covariance matrix ($\tilde{\Sigma}$) must be decomposed using a Choleski decomposition (analogous to taking the square root of $\tilde{\Sigma}$; Burden and Faires 1985) to produce a lower-diagonal matrix ($\tilde{L}$) such that :

$$\tilde{\Sigma} = \tilde{L}\tilde{L}^T. \tag{3.68}$$

$\tilde{L}$ is thus the multivariate equivalent of the univariate standard deviation in the multivariate z-score formula (in matrix notation) :

$$\vec{R} = \vec{M} + \tilde{L}\vec{Z}. \tag{3.69}$$

By pre-multiplying the lower-diagonal $p \times p$ matrix ($\tilde{L}$) by the $p \times 1$ vector of standardized scores ($\vec{Z}$) and adding the $p \times 1$ vector of means ($\vec{M}$), a $p \times 1$ vector of multivariate normal scores ($\vec{R}$) is produced (Ghose and Pinnaduwa 1987; for alternative methods see : Alabert 1987; Davis 1987a, 1987b; Mantaglou 1987). This procedure can be repeated numerous times to produce a statistical sample comprised of n cases of p variables which will have a mean vector and covariance matrix which approximates that

of the population. Thus, data sets can be generated which consist of mixtures of multivariate normal distributions with different means, standard deviations, correlations and component proportions.

Bimodal mixtures of normally-distributed data sets were generated with an identifying number indicating the population membership for each observation, which was stored with the data so that subsequent classification procedures could be evaluated quantitatively. All stochastically generated data sets, as well as all statistics derived from them in the course of this study, were stored with 5 places to the right of the decimal point to prevent round-off errors and other inaccuracies.

3.2 Univariate Data Set Structures

A variety of data set structures was specified for use in this comparative evaluation. All were comprised of a mixture of two normal distributions; thus, extrapolation of the results of this study to data sets comprised of more than two distributions is possible. For each data set structure, ten different realizations were produced. The variations in data set structure consisted of modifying the various parameters that describe the distributions. These include variations in the total number of observations (n : Test # 1), variations in the component distribution percentages (ϖ : Test # 2), variations in the square root of the Mahalanobis distance between the two component distributions by changing the means and standard deviations (μ_2 : Test # 3, and σ_2 : Test # 4). Only one parameter was modified at any one time, and in all cases, μ_1 and σ_1 remained constant. A summary of the parameter values for each of the data set structures is presented in Table 3.2.

The Mahalanobis distance (Δ^2) was calculated using the following formula :

$$\Delta^2 = \frac{(\mu_1 - \mu_2)^2}{\sigma_p}, \tag{3.70}$$

Table 3.2: Parameter Values for the Different Data Set Structures Used to Generate the Stochastic Realizations

Structure Label	μ_1	σ_1	μ_2	σ_2	$\varpi(\%)$	n	Test #	Δ
(Datum) 1	20	5	40	5	50	200	1,2,3,4	4
16	20	5	40	5	50	50	1	4
2	20	5	40	5	50	100	1	4
3	20	5	40	5	50	300	1	4
4	20	5	40	5	50	400	1	4
5	20	5	40	5	50	500	1	4
6	20	5	40	5	70	200	2	4
7	20	5	40	5	85	200	2	4
8	20	5	40	5	95	200	2	4
9	20	5	25	5	50	200	3	1
10	20	5	30	5	50	200	3	2
11	20	5	35	5	50	200	3	3
12	20	5	45	5	50	200	3	5
13	20	5	50	5	50	200	3	6
14	20	5	40	10	50	200	4	2.53
15	20	5	40	15	50	200	4	1.79

(Listed According to Which Parameter is Different from the Datum, and Labelled According to their Relation to the Parameter Variation Tests)

where σ_p is the pooled standard deviation :

$$\sigma_p = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}}. \tag{3.71}$$

3.3 Procedure

Comparison of the *RDML* and *CIDML* parameter estimates was made using a version of the PROBPLOT program which does not constrain the two standard deviation parameter values to be greater than zero and the component proportion parameter to be between zero and one, inclusive. This allowed evaluation of the relative stability of the maximization algorithm using both likelihood function surfaces. For each of the 15 data set structures, the 10 data set realizations were evaluated using both likelihood functions.

Initially different seed values were used to determine whether the algorithm iterated to the same maximum. In each case, different inflection points were used to calculate different sets of parameter seed values (see Chapter 2). Results demonstrate that, for data set structure # 1 (Datum), 5 different sets of seed values on each of the 10 data sets all iterated to the same maximum for each data set. Variation in the results can be explained by the amount of round-off error produced by the iteration stopping criteria, which in all cases in this study was set at 0.01 % relative deviation between the vertices and between the likelihood function values.

As a result, for all data sets, the known inflection point (component population percentage) was used to determine the parameter seed values. This ensured a relatively rapid determination of the maximum likelihood solution. On one data set in each data set structure, two other inflection point values were used, where possible, which were 5 percentage points greater and less than the known value (only the lower could be used for data set structure # 8). In all cases, these final parameter estimates agreed well

with the estimates determined using the known inflection point, given round-off error. Furthermore, the variation in the resulting parameter estimates using different 'seed' parameter values on one data set structure realization was comparable to the variation between different realizations of the same data set structure.

Parameter estimates calculated directly from the data sets were tabulated. In addition, the means and standard deviations of the ten sets of *RDML* parameter estimates were determined using the PROBPLOT program. Multiple estimates of the parameters using the *CIDML* function were also derived with different numbers of equal sized class intervals. The numbers of class intervals ranged from 10 to 50, in increments of 5. The means and standard deviations of these 9 additional sets of parameter estimates were also calculated. The χ^2 function was used to determine the mean (and standard deviation) parameter estimates for the 10 realizations with the 'datum' data set structure (# 1) with the same class intervals used for the *CIDML* parameter estimates. This allowed comparison of the χ^2 and *CIDML* function parameter estimates.

3.4 Results

Although all of the *RDML* parameter estimates and χ^2 parameter estimates produced were feasible, the *CIDML* function commonly produced estimates outside of the parameter constraints. Several data sets iterated to conditions where $0 \not\leq \varpi$, $\varpi \not\leq 1$ or $\sigma_{1,2} \rightarrow 0$, and these produced 'computation overflow' errors. These unconstrained cases were considered failures, even though utilization of the parameter constraining version of the PROBPLOT program would normally have produced acceptable estimates. Where unfeasible estimates were produced, a new data set of similar structure was generated and estimates of its parameters were determined. In several cases, numerous additional data sets were required because of the high level of instability of the

Table 3.3: Number of Additional Data Set Realizations Required to Produce 10 Feasible Sets of Parameter Estimates for Different Data Structures

Structure Label	Variable	Number of Class Intervals								
Test # 1	n	10	15	20	25	30	35	40	45	50
2	100	3	2	0	0	0	0	0	0	0
1	200	1	0	0	0	0	0	0	0	0
3	300	0	0	0	0	0	0	0	0	0
4	400	0	0	0	0	0	0	0	0	0
5	500	0	0	0	0	0	0	0	0	0
Test # 2	$\varpi(\%)$									
1	50	1	0	0	0	0	0	0	0	0
6	70	0	0	0	0	0	0	0	0	0
7	85	1	0	0	0	0	0	0	0	0
8	95	0	0	0	0	0	0	0	0	0
Test # 3	μ_2									
9	25	12	3	1	0	0	0	0	0	0
10	30	6	1	0	0	0	0	0	0	0
11	35	2	0	0	0	0	0	0	0	0
1	40	1	0	0	0	0	0	0	0	0
12	45	0	0	0	0	0	0	0	0	0
13	50	0	0	0	0	0	0	0	0	0
Test # 4	σ_2									
1	5	1	0	0	0	0	0	0	0	0
14	10	1	0	0	0	0	0	0	0	0
15	15	7	2	0	0	0	0	0	0	0

optimization algorithm on the likelihood surface.

Table 3.3 presents the number of extra data sets which had to be generated to produce 10 feasible sets of *CIDML* parameter estimates. This table gives a semi-quantitative estimate of the relative stability of the *CIDML* function applied to different data set structures.

Table 3.4: Key for the Table of Differences Between the *RDML* Parameter Estimates, *CIDML* Parameter Estimates, χ^2 Parameter Estimates and Stochastically Generated Sample Parameter Estimates and the Population Parameter Values

(for Table 3.5 and Appendix A.1)

Stochastic	=	the means and standard deviations of the differences between the actual population parameter values and the calculated statistical sample parameter estimates for the stochastically generated data sets
$\ell\ \infty$	=	the means and standard deviations of the differences between the actual population parameter values and the *RDML* parameter estimates for the 10 stochastically generated data sets
$\ell\ \#$	=	the means and standard deviations of the differences between the actual population parameter values and the *CIDML* parameter estimates for the 10 stochastically generated data sets
$\chi^2\ \#$	=	the means and standard deviations of the differences between the actual population parameter values and the minimum χ^2 parameter estimates for the 10 stochastically generated data sets
$\bar{d}_{\hat{\mu}_1}$	=	mean of 10 differences for the lower mean estimates
$\bar{s}_{\hat{\mu}_1}$	=	standard deviation of 10 differences for the lower mean estimates
$\bar{d}_{\hat{\mu}_2}$	=	mean of 10 differences for the upper mean estimates
$\bar{s}_{\hat{\mu}_2}$	=	standard deviation of 10 differences for the upper mean estimates
$\bar{d}_{\hat{\sigma}_1}$	=	mean of 10 differences for the lower standard deviation estimates
$\bar{s}_{\hat{\sigma}_1}$	=	standard deviation of 10 differences for the lower standard deviation estimates
$\bar{d}_{\hat{\sigma}_2}$	=	mean of 10 differences for the upper upper standard deviation estimates
$\bar{s}_{\hat{\sigma}_2}$	=	standard deviation of 10 differences for the upper standard deviation estimates
$\bar{d}_{\hat{w}(\%)}$	=	mean of 10 differences for the component proportion estimates
$\bar{s}_{\hat{w}(\%)}$	=	standard deviation of 10 differences for the component proportion estimates

Table 3.3 indicates that the *CIDML* function does not generally produce reasonable estimates where a small number of class intervals are used. Unfortunately, if a small number of class intervals are used, the *CIDML* function offers the greatest advantage over the *RDML* function in terms of calculation time. Additionally, instability also appears to be the result of both the size of the data set (unstable if small) and the proximity of the two distributions (unstable if a substantial amount of population overlap exists, corresponding to small Mahalanobis distances, caused both by the proximity of the means (Test # 3) and the magnitude of the standard deviations relative to the means (Test # 4)). Variations in the relative percentages of the two populations do not appear to significantly affect the stability of the *CIDML* function or the performance of the SIMPLEX algorithm in finding its maximum.

Results from the comparison of the *RDML* and the *CIDML* function performance described above are listed in Table 3.5 for the 'datum' data set (# 1) structure only. Plots demonstrating the relative bias of the *CIDML* parameter estimates for this data set structure are presented in Figure 3.13. Tables listing results for all other data set structures (# 2 through # 15) are presented in Appendix A.1 and corresponding plots of the results of these analyses are included in Appendix A.2.

3.5 Discussion

Although the number of realizations of each data set structure described above clearly are inadequate to precisely estimate the amount of bias produced through use of the various optimization functions, several important trends are recognizable. These are related to the type of data set structure used, the type of optimization function used to obtain the parameter estimates, and, in the case of the *CIDML* function and the χ^2 function, the number of class intervals used.

Table 3.5: Differences Between the *RDML* Parameter Estimates, *CIDML* Parameter Estimates, χ^2 Parameter Estimates and Stochastically Generated Sample Parameter Estimates and the Population Parameter Values for Data Set Structure # 1

(see Table 3.4 for key)

$\mu_1 = 20$, $\mu_2 = 40$, $\sigma_1 = 5$, $\sigma_2 = 5$, $\varpi(\%) = 50$, and $n = 200$

	$\bar{d}_{\hat{\mu}_1}$	$\bar{s}_{\hat{\mu}_1}$	$\bar{d}_{\hat{\mu}_2}$	$\bar{s}_{\hat{\mu}_2}$	$\bar{d}_{\hat{\sigma}_1}$	$\bar{s}_{\hat{\sigma}_1}$	$\bar{d}_{\hat{\sigma}_2}$	$\bar{s}_{\hat{\sigma}_2}$	$\bar{d}_{\hat{\varpi}(\%)}$	$\bar{s}_{\hat{\varpi}(\%)}$
Stochastic	-0.20	0.51	0.07	0.77	0.04	0.40	0.16	0.40	0.00	0.00
$\ell\ \infty$	-0.04	0.67	-0.26	0.57	-0.11	0.42	-0.14	0.49	-0.24	1.93
ℓ 10	0.08	0.51	-0.46	0.62	-0.25	0.49	-1.55	0.73	2.75	2.79
ℓ 15	0.20	0.44	-0.25	0.55	-0.04	0.43	-0.81	0.43	1.89	1.91
ℓ 20	0.18	0.38	-0.25	0.50	-0.00	0.45	-0.62	0.37	1.39	1.94
ℓ 25	0.19	0.58	-0.23	0.46	-0.03	0.44	-0.45	0.35	0.84	2.02
ℓ 30	0.03	0.53	-0.16	0.46	-0.07	0.47	-0.46	0.44	0.78	2.32
ℓ 35	0.08	0.47	-0.11	0.39	-0.01	0.39	-0.36	0.36	0.76	1.93
ℓ 40	0.11	0.67	-0.04	0.43	-0.03	0.49	-0.43	0.42	0.96	2.29
ℓ 45	0.20	0.53	-0.18	0.49	0.07	0.43	-0.35	0.46	0.98	2.26
ℓ 50	0.12	0.31	-0.05	0.29	0.04	0.24	-0.26	0.49	0.45	2.00
χ^2 10	0.26	0.95	-0.01	0.94	-0.08	0.39	-1.48	0.63	2.82	3.76
χ^2 15	0.22	0.68	-0.05	0.65	0.16	0.28	-0.75	0.38	1.88	1.42
χ^2 20	0.18	0.72	0.05	0.72	0.27	0.36	-0.72	0.45	2.36	1.78
χ^2 25	-0.05	1.05	0.09	0.74	0.40	0.29	-0.50	0.63	1.94	2.27
χ^2 30	-0.04	0.55	0.04	0.52	0.41	0.39	-0.55	0.61	2.28	2.01
χ^2 35	0.16	0.72	-0.03	0.55	0.62	0.33	-0.43	0.60	1.80	1.80
χ^2 40	-0.13	0.52	-0.16	0.58	0.28	0.68	-0.02	1.08	0.62	3.76
χ^2 45	-0.15	0.64	0.05	0.49	0.67	0.43	-0.63	0.51	2.45	3.11
χ^2 50	-0.40	0.54	0.09	0.52	0.60	0.33	-0.42	0.52	1.90	2.82

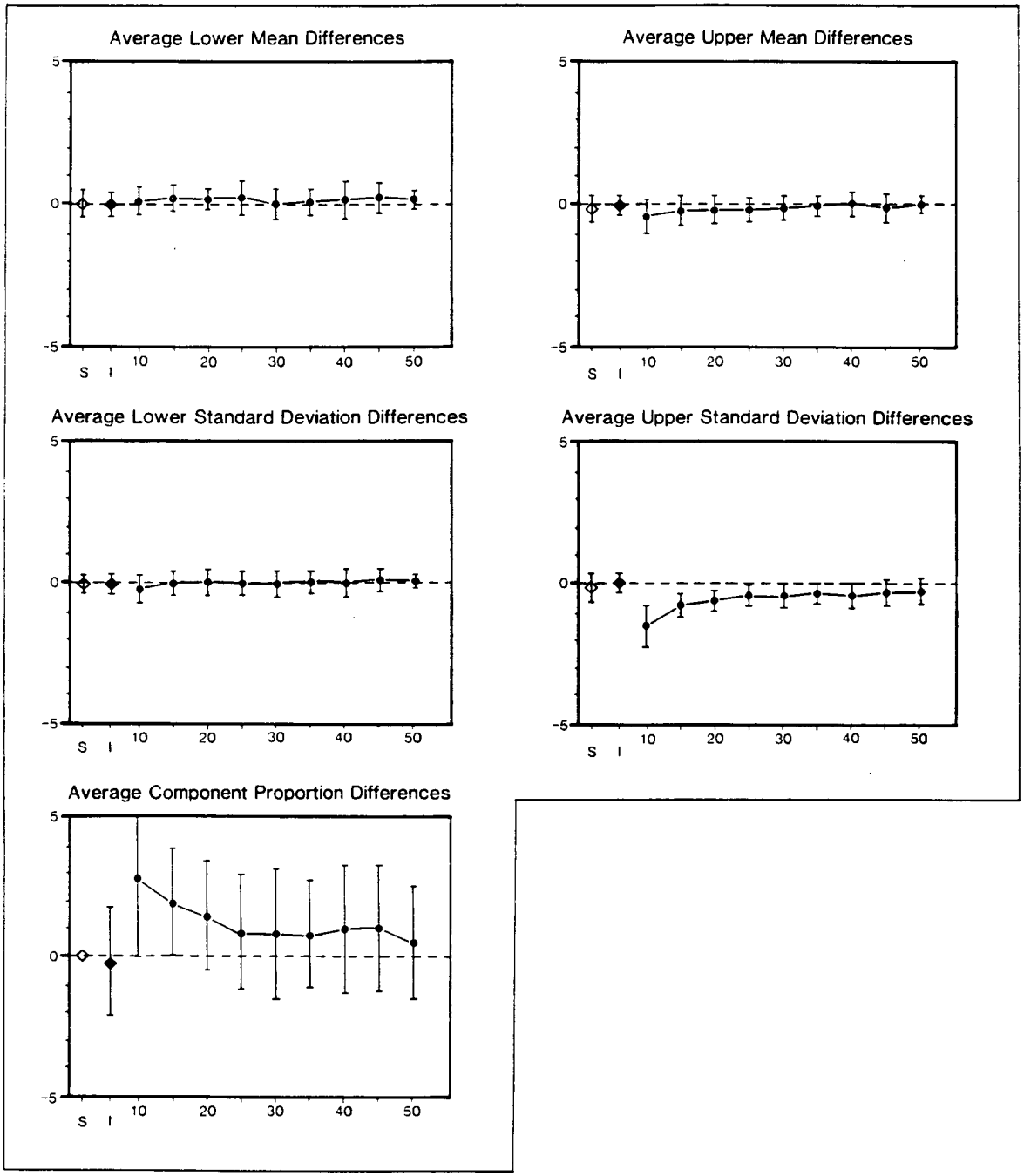


Figure 3.13: Parameter Bias of the *RDML* Parameter Estimates, *CIDML* Parameter Estimates and the Stochastically Generated Sample Parameter Estimates Relative to the True Population Parameter Estimates for Data Set Structure # 1 (Datum)

3.5.1 Raw Data Likelihood Function Bias

Comparison of the actual parameter values, the calculated statistical sample parameter estimates for the stochastically generated data sets and the *RDML* parameter estimates made using the stochastically generated data sets reveals several important trends. These are described in point form relative to the true population parameter values used to produce the stochastically generated bimodal normal data sets.

3.5.1.1 Effects of Variations in n

The mean calculated statistical sample and *RDML* parameter estimates of data set structures which differ only in the number of observations per data set (Test # 1) are neither different from the actual parameter values nor different from each other. Data set structures with larger numbers of observations exhibit lower but roughly equal amounts of variance for both the *RDML* and the mean calculated statistical sample parameter estimates.

The variances of both the statistical sample estimates and the *RDML* parameter estimates should be proportional to $1/n$, where n is the number of observations in the data set. Results presented in Table 3.6 indicate that the product of the parameter estimate standard deviations $\times\sqrt{n}$ exhibits no systematic change with increasing n. This product does, however, show variation. This observed non-proportionality (variation) may occur because :

- the small number of realizations for each data set structure (10), and
- the small size of each data set (100 to 500 observations)

prevent stable statistical estimates of the variances of the parameter estimates. No effort has been made to test the significance of any possible non-proportionality.

Table 3.6: Comparison of Proportionality of *RDML* Parameter Estimate Variances With the Number of Observations in the Data Sets Used to Estimate that Variance

n	$\hat{s}_{\mu_1}\sqrt{n}$	$\hat{s}_{\mu_2}\sqrt{n}$	$\hat{s}_{\sigma_1}\sqrt{n}$	$\hat{s}_{\sigma_2}\sqrt{n}$	$\hat{s}_{\varpi(\%)}\sqrt{n}$
Stochastic					
100	5.100	7.700	4.400	4.000	0.000
200	7.212	10.889	5.657	5.657	0.000
300	4.330	3.811	4.157	5.023	0.000
400	8.400	6.400	2.400	5.200	0.000
500	5.814	7.603	4.472	4.025	0.000
$\ell\,\infty$					
100	8.800	11.100	6.300	5.500	26.100
200	9.475	8.061	5.940	6.930	27.294
300	7.448	6.235	6.582	3.811	22.517
400	11.600	9.800	6.400	7.400	27.600
500	7.603	8.497	4.025	5.367	14.982

3.5.1.2 Effects of Variations in ϖ

Data set structures with large disparity in the proportions of the component distributions (Test # 2) have, with respect to data set structures with roughly equal component distribution proportions :

- a negative bias for the mean *RDML* parameter estimates of the mean of the less abundant component distribution (μ_2), relative to both the actual parameter value and the mean statstical sample parameter estimates;
- a positive bias for the mean *RDML* parameter estimates of the standard deviation of the less abundant component distribution (σ_2), relative to both the actual parameter value and the mean statistical sample parameter estimates;
- no bias in the mean *RDML* parameter estimates of the mean and standard deviation of the more abundant component distribution, relative to both the actual parameter value and the mean statistical sample parameter estimates;
- a negative bias for the mean *RDML* parameter estimates of the distribution percentage (ϖ) of the more abundant component distribution relative to both the actual parameter value; and
- lower variances for the mean *RDML* and mean statistical sample parameter estimates of the more abundant distribution, but larger variances for the mean *RDML* and mean statistical sample parameter estimates of the less abundant distribution.

3.5.1.3 Effects of Variations in μ_2

Data set structures with small Mahalanobis distances between the two component distributions (due to proximity of means; Test # 3) have, with respect to data set

structures with large Mahalanobis distances between the two component distributions :

- no bias in both the mean *RDML* and mean statistical sample parameter estimates of the mean and standard deviation of both component distributions relative to the actual parameter values;
- larger variances in the mean *RDML* parameter estimates of the mean and standard deviation for both component distributions;
- larger variances in the mean *RDML* parameter estimates of the mean and standard deviation for both component distributions, relative to the mean statistical sample parameter estimates, where the Δ between the component distributions is less than 2.0;
- smaller variances in the mean *RDML* parameter estimates of the mean and standard deviation for both component distributions, relative to the mean statistical sample parameter estimates, where the Δ between the component distributions is greater than 2.0; and
- a non-systematic bias and larger variance in the mean *RDML* component percentage parameter estimates (ϖ), relative to the actual parameter value, where the Δ is less than 2.0.

The non-systematic bias and larger variance of the mean *RDML* parameter estimate of the component percentage (ϖ) suggests that the SIMPLEX optimization algorithm behaves poorly on the (probably highly irregular) raw data likelihood surface in that parameter dimension for those data set structures (# 9 and # 10).

3.5.1.4 Effects of Variations in σ_2

Data set structures with small Mahalanobis distances between the two component distributions (due to the magnitude of standard deviations relative to the means; Test # 4) have, with respect to data set structures with large Δ between the two component distributions :

- a positive bias in the mean *RDML* parameter estimates of the mean of the distribution with the larger standard deviation, relative to the mean statistical sample parameter estimates and the actual parameter value;
- a negative bias in the mean *RDML* parameter estimates of the standard deviation of the distribution with the larger standard deviation, relative to the mean statistical sample parameter estimates and the actual parameter value;
- larger variances of the mean *RDML* and mean statistical sample parameter estimates of the mean and standard deviation of the distribution with the larger standard deviation;
- a positive bias in the mean *RDML* parameter estimates of the component percentage; and
- larger variances of the mean *RDML* parameter estimates of the component percentage.

3.5.2 Class Interval Data Likelihood Function Bias

Results from use of the *CIDML* function with different numbers of class intervals also exhibit several important trends. These are all described relative to the actual parameter

values to allow a quantitative evaluation of the bias associated with the technique. References are also made to the mean *RDML* parameter estimates and the mean statistical sample parameter estimates where a direct comparison reveals important information.

3.5.2.1 Effects of Variations in the Number of Class Intervals

With increasing numbers of class intervals, the variances of the *CIDML* parameter estimates decrease for all data set structures, a feature clearly related to the amount of information lost through categorization (cumulation) into class intervals of ever decreasing size. Thus, as the number of class intervals increases, the size of each class interval decreases and the *CIDML* function converges to the *RDML* function. Furthermore, for data set parameter values optimized with increasing numbers of class intervals :

- the variances of the *CIDML* mean and standard deviation estimates converge to the variance of the mean *RDML* parameter estimates of the mean and standard deviation estimates (and thus also converge on the actual parameter values);
- if bias exists, the mean *CIDML* parameter estimates of the standard deviations increase toward the mean *RDML* parameter estimates of the standard deviations;
- biased mean *CIDML* parameter estimates of the means approach the mean *RDML* parameter estimates of the corresponding means from both above and below; and
- where the mean *CIDML* component percentage estimates are positively biased, they converge with decreasing variance toward the *RDML* component percentage estimates from above.

Certain other trends are observed which may be related to both variations in data set structure and the interplay of these variations with the number of class intervals

used in the optimization procedure.

3.5.2.2 Effects of Variations in n

The primary effect that the size of the data set has on the optimal parameter estimates of the *CIDML* function consists of a reduction of the variation in the results. This 'stabilization' produces mean *CIDML* parameter estimates which better approximate mean *RDML* parameter estimates (and the actual parameter values) and their variances become smaller with larger data sets. Furthermore :

- the mean *CIDML* parameter estimates of the lower mean converge to the corresponding mean *RDML* parameter estimates (and the actual parameter values) from above with increasing numbers of class intervals;
- the mean *CIDML* parameter estimates of the upper mean converge to the corresponding mean *RDML* parameter estimate (and the actual parameter values) from below with increasing numbers of class intervals;
- the above positive and negative biases become less pronounced with larger data sets;
- the mean *CIDML* parameter estimates of the upper standard deviations are under-estimated with small numbers of class intervals and data sets of less than 300 observations; and
- mean *CIDML* parameter estimates of the component percentages are by far the most sensitive to the number of class intervals utilized, converging from above, but approximating the *RDML* parameter estimates only if greater than 30 class intervals are used with data sets of 100 observations (and greater than 15 class intervals with data sets of 500 observations).

The variances of each of the *CIDML* parameter estimates should also be proportional to $1/n$, where n is the number of observations in the data set. Results presented in Appendix A.3 indicate that the product of the parameter estimate standard deviations $\times\sqrt{n}$ show no systematic change with increasing n. The amount of variation among the data is smallest if a large number of class intervals was used to determine the parameter estimates, and if the size of the data set is large (> 300).

3.5.2.3 Effects of Variations in ϖ

The primary effect caused by increasing the difference between the component distribution percentages on the *CIDML* parameter estimates is to increase the accuracy and precision of the mean and standard deviation parameter estimates of the more abundant distribution, while decreasing the accuracy and precision of the mean and standard deviation of the less abundant distribution. Specifically, for data sets with disparate component percentages :

- the mean *CIDML* parameter estimates of the standard deviation of the distribution with a smaller component proportion are under-estimated with respect to the mean *RDML* parameter estimates and the actual parameter values;
- the mean *CIDML* parameter estimates of the mean of the distribution with the smaller component proportion are under-estimated with respect to the mean *RDML* parameter estimates and the actual parameter values, and this pattern is more pronounced with smaller numbers of class intervals;
- the mean *CIDML* parameter estimates of the component percentage (ϖ) are unbiased, but are over-estimated in data set structures which have **equal** component percentages and smaller numbers of class intervals.

3.5.2.4 Effects of Variations in μ_2

Data set structures with variations in the Mahalanobis distance between means of the two component distributions have similar mean *CIDML* parameter estimates of the means and standard deviations, but :

- the variances of *CIDML* parameter estimates of the component distribution means and standard deviations are smaller where the component distributions have large Δ; and
- where the difference between the means of the two distributions is small, the mean *CIDML* parameter estimates of the component percentage (ϖ) are imprecisely estimated with respect to the mean *RDML* parameter estimates and the actual parameter values.

This non-systematic bias of the mean *CIDML* parameter estimate of the component percentage (ϖ) is similar to the behavior of the mean *RDML* parameter estimates of the component percentage, and also suggests that the SIMPLEX optimization algorithm behaves poorly on the (probably highly irregular) class interval data likelihood surface in that parameter dimension for those data set structures (# 9 and # 10).

3.5.2.5 Effects of Variations in σ_2

Data set structures with variations in the standard deviation of one of the component distributions affects only the mean *CIDML* parameter estimates of the mean and standard deviation of the distribution with the larger standard deviation. Specifically, with increasing disparity in the standard deviations :

- the mean *CIDML* parameter estimates of the mean of the distribution with the larger standard deviation are over-estimated;

- the mean *CIDML* parameter estimates of the standard deviation of the distribution with the larger standard deviation are under-estimated;
- the variance of the *CIDML* parameter estimates of the mean and standard deviation of the distribution with the larger standard deviation increases; and
- increasing the difference between the standard deviations for the component distributions has no effect on the parameters of the component distribution with the smaller standard deviation.

3.5.3 χ^2 Function Bias

Results from use of the χ^2 function with different numbers of class intervals for data set structure # 1 also exhibit several important trends. These are similar to the trends observed in the *CIDML* parameter estimates for data set structure # 1. In general, χ^2 parameter estimates for all 5 parameters are, on average, all slightly larger than the *CIDML* parameter estimates. As a result, they approximate the true parameter values for μ_2 better than the *CIDML* parameter estimates. In contrast, the differences between the χ^2 parameter estimates and the true population values have equal to larger standard deviations (by up to 25 %) than the differences between the *CIDML* parameter estimates and the true population values. Thus, little difference exists between the quality of the parameter estimates produced by the χ^2 function and the class interval data likelihood function for this data set structure. Since other data set structures were not evaluated with the χ^2 function, extrapolation of these results and conclusions to other data set structures is not advised.

3.5.4 Asymptotic Variances of the Parameter Estimates

Estimates of the asymptotic variances (and covariances) of the parameter estimates at the maximum likelihood solution can be determined using the second derivative (Hessian) matrix of the likelihood function. These variances are approximated by the inverse of the observed Fisher information matrix, which is the negative of the Hessian matrix :

$$\tilde{\Sigma} \cong \tilde{I}^{-1} = (-\tilde{H})^{-1}, \tag{3.72}$$

where $\tilde{I}$ is the observed Fisher information matrix and $\tilde{H}$ is the Hessian matrix (Cox and Hinkley 1974). Derivation of the formulae for calculation of the Hessian matrix for both the *RDML* and *CIDML* functions are presented in Appendix B.

3.5.4.1 Raw Data Likelihood Function Estimates

The following tables present the standard deviations of the 10 *RDML* parameter estimates. The standard deviations determined from the variance of the 10 differences between the *RDML* parameter estimates and the actual parameter value are compared with the asymptotic standard deviations calculated from the inverse of the negative Hessian matrix (observed Fisher information matrix) evaluated at the *RDML* parameter estimates.

Tables 3.8 and 3.9 compare the empirically estimated parameter standard deviations with the mean asymptotic standard deviations approximated by the observed Fisher information matrix for the *RDML* function.

Table 3.7: Key for Tables Comparing the Empirical and Asymptotic Standard Deviations of the *RDML* Parameter Estimates and *CIDML* Parameter Estimates for all Data Set Structure Tests

(for Tables 3.8, 3.9 and 3.11)

$s_{\hat{\mu}_1}$ = standard deviation of the 10 lower mean parameter estimate differences

$\bar{\sigma}_{\hat{\mu}_1}$ = mean asymptotic standard deviation for the 10 lower mean parameter estimates

$s_{\hat{\sigma}_1}$ = standard deviation of the 10 lower standard deviation parameter estimate differences

$\bar{\sigma}_{\hat{\sigma}_1}$ = mean asymptotic standard deviation for the 10 lower standard deviation parameter estimates

$s_{\hat{\varpi}}$ = standard deviation of the 10 component proportion parameter estimate differences

$\bar{\sigma}_{\hat{\varpi}}$ = mean asymptotic standard deviation for the 10 component proportion parameter estimates

$s_{\hat{\mu}_2}$ = standard deviation of the 10 upper mean parameter estimate differences

$\bar{\sigma}_{\hat{\mu}_2}$ = mean asymptotic standard deviation for the 10 upper mean parameter estimates

$s_{\hat{\sigma}_2}$ = standard deviation of the 10 upper standard deviation parameter estimate differences

$\bar{\sigma}_{\hat{\sigma}_2}$ = mean asymptotic standard deviation for the 10 upper standard deviation parameter estimates

(Values in parentheses are the standard deviations of the asymptotic standard deviation estimates for the corresponding parameter.)

Table 3.8: Comparison of the Empirical and Asymptotic Standard Deviations of the *RDML* Parameter Estimates for Data Set Structure Tests # 1 and # 2

(see Table 3.7 for key)

Structure Label	$s_{\hat{\mu}_1}$	$\bar{\sigma}_{\hat{\mu}_1}$	$s_{\hat{\sigma}_1}$	$\bar{\sigma}_{\hat{\sigma}_1}$	$s_{\hat{\varpi}}$	$\bar{\sigma}_{\hat{\varpi}}$	$s_{\hat{\mu}_2}$	$\bar{\sigma}_{\hat{\mu}_2}$	$s_{\hat{\sigma}_2}$	$\bar{\sigma}_{\hat{\sigma}_2}$
Test # 1										
16	1.76	1.13	1.09	0.86	5.98	7.88	0.96	1.25	1.11	0.96
		(0.31)		(0.23)		(0.71)		(0.29)		(0.24)
2	0.88	0.88	0.63	0.68	2.60	5.85	1.11	0.98	0.55	0.75
		(0.24)		(0.18)		(0.75)		(0.19)		(0.12)
1	0.67	0.61	0.42	0.47	1.93	3.99	0.56	0.61	0.49	0.47
		(0.07)		(0.07)		(0.30)		(0.11)		(0.09)
3	0.43	0.48	0.37	0.38	1.29	3.17	0.36	0.48	0.22	0.37
		(0.05)		(0.04)		(0.10)		(0.04)		(0.03)
4	0.58	0.43	0.31	0.34	1.38	2.81	0.48	0.43	0.37	0.34
		(0.04)		(0.03)		(0.12)		(0.05)		(0.03)
5	0.34	0.36	0.18	0.28	0.66	2.45	0.38	0.38	0.23	0.29
		(0.03)		(0.03)		(0.10)		(0.02)		(0.03)
Test # 2										
1	0.67	0.61	0.42	0.47	1.93	3.99	0.56	0.61	0.49	0.47
		(0.07)		(0.07)		(0.30)		(0.11)		(0.09)
6	0.45	0.52	0.34	0.40	1.38	3.77	0.94	0.88	0.48	0.65
		(0.07)		(0.05)		(0.40)		(0.20)		(0.13)
7	0.48	0.45	0.39	0.34	2.02	3.75	1.57	1.56	0.82	1.19
		(0.11)		(0.07)		(1.36)		(0.76)		(0.34)
8	0.22	0.39	0.32	0.29	1.51	2.90	2.37	3.65	1.83	2.37
		(0.04)		(0.03)		(2.07)		(2.53)		(1.25)

Table 3.9: Comparison of the Empirical and Asymptotic Standard Deviations of the *RDML* Parameter Estimates for Data Set Structure Tests # 3 and # 4

(see Table 3.7 for key)

Structure Label	$s_{\hat{\mu}_1}$	$\bar{\sigma}_{\hat{\mu}_1}$	$s_{\hat{\sigma}_1}$	$\bar{\sigma}_{\hat{\sigma}_1}$	$s_{\hat{w}}$	$\bar{\sigma}_{\hat{w}}$	$s_{\hat{\mu}_2}$	$\bar{\sigma}_{\hat{\mu}_2}$	$s_{\hat{\sigma}_2}$	$\bar{\sigma}_{\hat{\sigma}_2}$
Test # 3										
9	1.33	2.82	0.90	1.55	1.48	78.48	2.24	2.21	0.58	1.54
		(1.63)		(0.66)		(39.79)		(0.91)		(0.86)
10	1.16	3.09	0.50	1.23	9.58	34.21	1.15	3.43	0.73	1.28
		(1.96)		(0.50)		(32.93)		(3.27)		(0.85)
11	1.21	0.78	0.60	0.53	9.50	5.71	1.69	0.84	0.86	0.58
		(0.18)		(0.09)		(1.31)		(0.21)		(0.11)
1	0.67	0.61	0.42	0.47	1.93	3.99	0.56	0.61	0.49	0.47
		(0.07)		(0.07)		(0.30)		(0.11)		(0.09)
12	0.50	0.53	0.36	0.40	3.24	3.59	0.36	0.54	0.60	0.42
		(0.04)		(0.04)		(0.01)		(0.07)		(0.06)
13	0.59	0.51	0.33	0.38	0.20	3.55	0.47	0.50	0.36	0.37
		(0.03)		(0.02)		(0.01)		(0.04)		(0.04)
Test # 4										
1	0.67	0.61	0.42	0.47	1.93	3.99	0.56	0.61	0.49	0.47
		(0.07)		(0.07)		(0.30)		(0.11)		(0.09)
14	0.83	0.75	0.37	0.54	5.32	6.80	1.91	2.14	0.76	1.37
		(0.08)		(0.05)		(9.23)		(0.25)		(0.12)
15	0.70	0.73	0.32	0.70	5.99	7.86	3.08	3.59	1.76	2.02
		(0.09)		(0.22)		(2.26)		(1.40)		(0.78)

In general, good agreement exists between each of these standard deviation estimates. Differences are more pronounced with the smaller data set structures (# 16) and those where inversion of the Hessian matrix was not possible in every case (realizations for data set structure # 9 could only have their Hessian matrices inverted 5 out of 10 times, because the half of these matrices were not positive definite). This was probably due to convergence of the SIMPLEX algorithm to a flat portion of the likelihood

Table 3.10: Average Estimated Asymptotic Correlation Matrix (Linear Correlation Coefficients) of the *RDML* Parameters for Data Set Structure # 1

(see Table 3.7 for key)

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.37 (0.02)	1.00 (0.00)			
ϖ	0.26 (0.02)	0.27 (0.01)	1.00 (0.00)		
μ_2	0.31 (0.02)	0.32 (0.02)	0.26 (0.02)	1.00 (0.00)	
σ_2	-0.33 (0.02)	-0.32 (0.01)	-0.28 (0.01)	-0.37 (0.02)	1.00 (0.00)

surface which was **not quite** a true maxima (see Table 3.3), and thus the parameter estimates were **not quite** the true *RDML* parameter estimates. Most empirically estimated parameter standard deviations lie within 2 standard deviations of the asymptotic parameter standard deviation estimates. The only significant variation between these estimates exists for the ϖ standard deviations, where the asymptotic standard deviation estimates are substantially greater than those estimated empirically.

Correlations between pairs of parameters can also be determined at the *RDML* parameter estimates. The mean estimated asymptotic correlation matrix for data set structure # 1 is presented in Table 3.10. The remainder of these correlation matrices are presented in Appendix C.1.

Several notable trends are observed within each of the matrices. First, positive correlations occur for all parameter pairs except for those involving σ_2. In addition, the highest absolute correlations occur between the mean and standard deviation parameters of each component distribution. Finally, the lowest absolute correlations occur between both the mean and standard deviation parameters and the component percentage parameter.

In general, all correlations have roughly the same absolute magnitude within each

matrix (except for those from data set structures which could not have their Hessian matrix inverted routinely; data set structure # 9; see Table 3.3). The notable trends between these matrices include an overall decreasing average correlation as the means of the two component distributions become disparate, and increasing variation in the correlation estimates as the number of observations in the data set structure decreases and as the standard deviations become more disparate.

3.5.4.2 Class Interval Data Likelihood Function Estimates

Asymptotic correlation estimates were also determined for the *CIDML* parameter estimates for data set structure # 1 (Datum) only. The standard deviations determined from the variance of the 10 differences between the *CIDML* parameter estimates and the true values are compared with the asymptotic standard deviations calculated from the inverse of the negative Hessian matrix (observed Fisher information matrix) evaluated at the *CIDML* parameter estimates.

Asymptotic covariance estimates for the *CIDML* function were calculated according to the formulae derived in Appendix B.2. Comparison of the asymptotic standard deviation estimates with those empirically estimated for different numbers of class intervals for data set structure # 1 (Datum) are presented in Table 3.11 (refer to Tables 3.7 for the key to the symbols used in this table). All asymptotic correlations for data set structure # 1 using the *CIDML* function are presented in Appendix C.2.

Table 3.11: Comparison of the Empirical and Asymptotic Standard Deviations of the *CIDML* Parameter Estimates for Data Set Structure # 1

# Class Intervals	$s_{\hat{\mu}_1}$	$\bar{\sigma}_{\hat{\mu}_1}$	$s_{\hat{\sigma}_1}$	$\bar{\sigma}_{\hat{\sigma}_1}$	$s_{\hat{\varpi}}$	$\bar{\sigma}_{\hat{\varpi}}$	$s_{\hat{\mu}_2}$	$\bar{\sigma}_{\hat{\mu}_2}$	$s_{\hat{\sigma}_2}$	$\bar{\sigma}_{\hat{\sigma}_2}$
10	0.51	0.56 (0.08)	0.62	0.46 (0.06)	2.79	1.20 (0.10)	0.49	0.60 (0.14)	0.73	0.35 (0.09)
15	0.44	0.61 (0.09)	0.55	0.49 (0.10)	1.91	0.86 (0.07)	0.43	0.60 (0.13)	0.43	0.36 (0.07)
20	0.38	0.59 (0.07)	0.50	0.48 (0.06)	1.94	0.66 (0.04)	0.45	0.60 (0.11)	0.37	0.39 (0.07)
25	0.58	0.61 (0.09)	0.46	0.49 (0.06)	2.02	0.54 (0.05)	0.44	0.65 (0.17)	0.35	0.41 (0.08)
30	0.53	0.58 (0.07)	0.46	0.47 (0.05)	2.32	0.44 (0.04)	0.47	0.61 (0.11)	0.44	0.40 (0.06)
35	0.47	0.60 (0.07)	0.39	0.49 (0.07)	1.93	0.39 (0.04)	0.39	0.63 (0.15)	0.36	0.41 (0.07)
40	0.67	0.62 (0.14)	0.43	0.52 (0.16)	2.29	0.36 (0.07)	0.49	0.66 (0.22)	0.42	0.40 (0.06)
45	0.53	0.63 (0.10)	0.49	0.52 (0.10)	2.26	0.32 (0.03)	0.43	0.67 (0.15)	0.46	0.42 (0.08)
50	0.31	0.66 (0.19)	0.29	0.54 (0.16)	2.00	0.28 (0.03)	0.24	0.67 (0.18)	0.49	0.45 (0.14)

Table 3.11 demonstrates that no change in the asymptotic or empirically determined standard deviations of the parameters at the maximum likelihood solution occurs with changing numbers of class intervals. The two standard deviation estimates are quantitatively similar (the empirically determined values are generally within 2 standard deviations of the asymptotic estimate), with the exception of the asymptotic standard deviation of the component percentage parameter (ϖ), which is substantially lower when estimated empirically. This difference is in **direct contrast** to the difference observed for the component percentage parameter standard deviation for the *RDML* function.

Asymptotic correlations between the parameters also show no change with different numbers of class intervals. The magnitude of the correlations are roughly the same for the *CIDML* function as the *RDML* function, with the exception that the correlation between ϖ and σ_2 is positive and close to zero. This differs from the asymptotic correlation estimates of the *RDML* function, which have decisively negative correlations between these two parameters.

3.6 Conclusions

The *RDML* parameter estimates are generally good estimates of the known parameter values using the estimation procedures described in Chapter 2. Estimation is accurate if the component distributions are widely separated (characterized by large Mahalanobis distances) and if a large number of observations are used. The *CIDML* parameter estimates are stably calculated and produce equally acceptable estimates of the known parameter values only where a large number of class intervals are used. This is equivalent to stipulating that the number of observations per class interval is small, and thus represents a situation where more accurate estimation of the population frequencies

for each class interval is obtained using the trapezoid approximation. Hence, either likelihood function can be used to produce good estimates of the parameters of a data set composed of a mixture of normal distributions, provided a large amount of data is available (generally greater than 50 observations per component distribution) and, in the case of the *CIDML* function, a large number of class intervals are used to cumulate the data (generally, on average greater than 10 observations per class interval).

If data set structures similar to the simulations which produced both biased and non-systematic results occur (the component proportion parameter estimate in Test #3 for both the *RDML* and the *CIDML* functions), application of either of the above procedures are ill-advised. The variable nature of the simulations suggest that the optimization algorithm performs poorly in searching for the maximum on both of these, probably, very irregular likelihood surfaces. In cases such as these, inconsistent results are likely and care should be taken in the application of the above procedures to data set structures of similar form.

Chapter 4

Univariate Technique Comparison

"**Power** : A probability of a possible outcome of a potential decision conditional upon an imaginable circumstance given a conceivable value of an algebraic embodiment of an abstract mathematical idea and the strict adherence to an extremely precise rule."

S.J. Senn (1988)

"You have to have some order in a disordered world."

Frank Lloyd Wright (1936)

Geochemical sample classification through threshold selection has been accomplished by a variety of univariate statistical techniques. Since classification in geochemical applications can involve both anomaly recognition and population discrimination, any approach used must be able to handle all possible conditions.

Traditional univariate parametric and non-parametric statistical approaches to anomaly recognition concerning normal distributions include Dixon, Grubbs, and Tietjen and Moore tests (Barnett and Lewis 1978). These include an entire range of tests involving situations where the parameters of the distribution (μ and σ) are both known and unknown, where outliers on each and both sides of the distribution can be tested and where single and multiple outliers can be considered. Most involve test statistics which are ratios of either the distances between ordered values from the statistical sample, or the variances calculated from the statistical sample with and without the

suspected outlying observations. All of these statistics test a null hypothesis H_0 : that all observations are drawn from a single normal distribution, against an alternative H_1 : that at least one observation is drawn from another normal distribution with a different mean and standard deviation.

Statistics which are calculated by ordering the data values and forming ratios of the distances between those values, of the form :

$$y(r,s,p,q) = \frac{x_{(r)} - x_{(s)}}{x_{(p)} - x_{(q)}}, \tag{4.73}$$

are called Dixon-type statistics (Dixon 1950, 1951, 1953), where r, s, p and q are the indices of the ordered data. Alternative tests involve Grubbs-type statistics (Grubbs 1950, 1969; Grubbs and Beck 1972; Sheesley 1977), where the ratios of the variances of the statistical sample calculated with and without the (up to 2) suspected outlying observations. If these statistics exceed the tabulated critical values at a specified confidence level, then the null hypothesis is rejected and the tested observations are considered to be outliers. Unfortunately, both of these statistic types are limited to tests of up to two outliers on one side of the distribution at a time, and thus, although they may be applied to a limited number of anomaly recognition situations, they cannot be used to discriminate populations.

A more general likelihood ratio test of similar form which allows evaluation of the significance of a number of outliers (h) on one tail was developed by Teitjen and Moore (1972). This test can accommodate the continuum of situations between and including anomaly recognition and population discrimination. The Teitjen and Moore (1972) test takes the following form :

$$L_h = \frac{\sum_{i=1}^{n-h}(y_i - \overline{y}_h)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \tag{4.74}$$

for outliers on the positive tail, and :

$$L_h = \frac{\sum_{i=h+1}^{n}(y_i - \overline{y}_h)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \tag{4.75}$$

for outliers on the negative tail, where n is the total number of observations, h is the number of observations suspected of being outliers (from another population), and $\overline{y}_h$ is the average of the $n - h$ observations not suspected of being outliers. L_h is tested against a table of critical values determined for the normal distribution to determine the significance level of the outliers considered. If L_h is **less** than the critical value at a certain confidence level, the h observations tested are considered outliers.

An alternative test by Teitjen and Moore (1972) tests a total of h outliers on both tails of the distribution simultaneously. This is done by first calculating $z_i = |y_i - \overline{y}|$ and sorting $\{z_i\}_{i=1}^{n}$ into ascending order. Then :

$$E_h = \frac{\sum_{i=1}^{n-h}(z_i - \overline{z}_h)^2}{\sum_{i=1}^{n}(z_i - \overline{z})^2}, \tag{4.76}$$

where :

$$\overline{z}_h = \frac{1}{n-h}\sum_{i=1}^{n-h} z_i. \tag{4.77}$$

E_h is also tested against a table of critical values to determine the significance of the h outliers. Again, if E_h is **less** than the tabulated critical value for a normal distribution, the h observations are considered to be outliers.

4.1 The Gap Statistic

The above statistical procedures **test for the existence** of outlying observations derived from other populations differing from the normal distribution of interest in their mean and standard deviation values. They are tests of discordancy (Barnett and Lewis 1978) determining whether a set of observations have a certain probability of being

derived from a single normal distribution. Although thresholds can be defined to separate anomalous from background observations using these techniques, they do not independently estimate the parameters of the distributions involved. None have been used in geochemical applications (Miesch 1981) and, in general, they perform poorly in recognizing anomalous observations where the populations significantly overlap (square rooit of Mahalanobis distances approximately less than 3; Teitjen and Moore 1972).

The gap statistic, developed by Miesch (1981), has been used in geochemical applications. It is used to select a value (threshold) which separates the outliers derived from an anomalous distribution from observations derived from a background distribution. The approach uses the order statistics of a uniformly distributed random variable to define the significance level of the standardized distance between two adjacent observations (David 1981).

This is accomplished by assuming the following :

- the data set $\{x_i\}_{i=1}^n$ is derived from a 3-parameter log-normal distribution (Aitchison and Brown 1957), defined by a mean, a standard deviation and α such that :

$$\{y_i\}_{i=1}^n = \{\ln(x_i - \alpha)\}_{i=1}^n \sim N(\mu_y, \sigma_y^2), \tag{4.78}$$

  if $\{x_i\}_{i=1}^n$ is positively skewed and :

$$\{y_i\}_{i=1}^n = \{\ln(\alpha - x_i)\}_{i=1}^n \sim N(\mu_y, \sigma_y^2), \tag{4.79}$$

  if $\{x_i\}_{i=1}^n$ is negatively skewed, and

- the parameters (μ_y, σ_y and α) of this data set are **known.**

It follows that if :

$$\{z_i\}_{i=1}^n = \left\{ \frac{ln(\pm(x_i - \alpha)) - \mu_y}{\sigma_y} \right\}_{i=1}^n, \tag{4.80}$$

then :

$$\{\Phi(z_i)\}_{i=1}^{n} \sim U(0,1). \tag{4.81}$$

If the z_i are re-labelled so that $z_{(1)} \leq z_{(2)} \leq \ldots \leq z_{(n)}$, then the gaps, or differences between adjacent values $(\Phi(z_{(1)}),\ \Phi(z_{(2)}) - \Phi(z_{(1)}),\ \ldots,\ \Phi(z_{(n)}) - \Phi(z_{(n-1)}),\ 1 - \Phi(z_{(n)}))$ are distributed like the gaps of a statistical sample from a $U(0,1)$ distribution.

Alternatively, the gaps in a statistical sample from a 3-parameter log-normal distribution have the same joint distribution as the gaps in a statistical sample from the $U(0,1)$ distribution. On average, gaps are large where the probability density is small and small where the probability density is large. The gaps are considered to be distributed like gaps from a statistical sample of a $U(0,1)$ distribution, provided corrections to the gaps are made to account for the different probability densities at each gap. Since the expected size of each gap is inversely proportional to the probability density at the mid-point of the corresponding gap, this correction can be accomplished by multiplying the size of the gap by the probability density at its mid-point (Miesch 1981).

A formal derivation of the gap statistic follows, where the largest gap is defined as :

$$\max_{1 \leq i \leq n-1} \Big(\Phi(z_{(i+1)}) - \Phi(z_{(i)})\Big), \tag{4.82}$$

because we ignore the first and last gap (they are not feasible threshold locations). Then, provided that $\Phi(z_{(i+1)}) - \Phi(z_{(i)})$ is small or, equivalently, that n is large (so that a trapezoidal approximation of the integral is accurate) :

$$\Phi(z_{(i+1)}) - \Phi(z_{(i)}) = \int_{z_{(i)}}^{z_{(i+1)}} \phi(u)du \doteq (z_{(i+1)} - z_{(i)})\phi(\frac{z_{(i+1)} + z_{(i)}}{2}), \tag{4.83}$$

and the largest gap can be approximated by :

$$\max_{1 \leq i \leq n-1} \Big((z_{(i+1)} - z_{(i)})\phi(\frac{z_{(i+1)} + z_{(i)}}{2})\Big). \tag{4.84}$$

This gap magnitude can then be compared with critical values, derived by Miesch (1981) from a *monte carlo* simulation for a statistical sample of size n to determine its level of significance.

Thus, the largest gap between adjacent values in a data set has a significance equal to the probability of the largest gap from a statistical sample of a $U(0,1)$ distribution. As a result, the largest gap can be tested to determine its significance level, and only when the largest gap can be considered a highly improbable occurrence in a statistical sample of observations derived from a $U(0,1)$ distribution is the largest gap considered as a possible threshold value. This general approach can be applied to any distribution (not just the 3-parameter log-normal distribution), provided that the parameters of the distribution are known and not estimated.

Several theoretical and practical problems and limitations exist in the use of this implementation of the gap statistic :

- as Miesch (1981) points out, analyses of geochemical samples must be independent (thus application to stream sediment geochemistry may not be valid because geochemical samples from the same stream may not be independent) and cannot contain variable bias;

- a large number of observations must be used so that the gaps are small and the trapezoidal approximation of the integral in the above equation is accurate (the trapezoidal approximation is not actually necessary, a more precise numerical evalution of the integral is possible, but the trapezoid technique is the method by which Miesch approximates the integral and thus it can contribute to errors in the analysis);

- data sets used for determination of the critical values were normal with **known** $\mu_y = 0$, $\sigma_y = 1$ and $\alpha = 0$ so that the 3-parameter log-normal transformation

was not required; the exact critical values will change if the 3-parameter log-normal transformation is required and μ_y, σ_y and α are **unknown** and must be estimated;

- the data distribution must be able to be transformed by a 3-parameter log-normal distribution such that the resulting standardized scores $\{z_i\}_{i=1}^n$ are unskewed and $N(0,1)$; unfortunately, even best estimates of the μ_y, σ_y and α parameters may produce significantly skewed scores;

- a 3-parameter log-normal distribution represents only an approximation of the theoretical distribution form of geochemical data (mixtures of binomial or Poisson distributions) and the accuracy of this approximation is not known; although no theoretical basis exists suggesting that 3-parameter log-normal distributions are an expected form for geochemical data (Aitchison and Brown 1957; Johnson and Kotz 1970), Miesch's choice of a 3-parameter log-transform was based on the "flexibility" which this family of distributions offers in transforming (normalizing) a variety of observed distributions which may exhibit positive, negative or no skewness (Miesch 1981); and

- it is not clear whether the location of the most significant gap corresponds to a threshold which best discriminates population observations from outliers; although case histories provided by Miesch (1981) appear to demonstrate that the largest gap does discriminate the outlying observations, no theoretical basis for this contention exists; thus, the gap statistic may be just another example of a discordancy test (**a test for the presence of outliers or for non-conformity to a 3-parameter log-normal distribution and not a test to determine which observations are the outliers**).

Unfortunately, if any of the above limitations or requirements are not satisfied, this implementation of the gap statistic test is rendered invalid.

4.2 Techniques Considered

In spite of the above problems, the gap statistic has been applied to geochemical and geological problems including both anomaly recognition and population discrimination (Miesch 1981). Results from the gap statistic appear to corroborate geologic hypotheses presented by Miesch, but this cannot be a justification of the validity of the technique on a theoretical or scientific basis. As a result, rigorous testing and comparison of the classification performance of the gap statistic with results from the probability plot technique was undertaken. These tests were made on data sets consistent with the postulated distribution model for geochemical data (see Chapter 1) to evaluate how well the gap statistic and probability plot techniques perform on mixtures of normal distributions.

The probability plot technique uses the *RDML* function to produce parameter estimates of a flexible distribution model, allowing application to a wide variety of distribution forms. The parameter estimates are used to select thresholds to classify data from the distributions present. These thresholds are selected so that a minimum amount of classification error is produced, based on the theoretical model (see Chapter 2). The distribution model implicit in the probability plot approach is completely consistent with the distribution model postulated for geochemical data in Chapter 1.

The gap statistic takes an alternative approach where the observed distribution is transformed using a 3-parameter logarithmic operator. The gap values are then calculated and compared with the order statistics from a uniform distribution. The locus of the largest significant deviation between the gap values and the order statistics

is defined as a threshold. In this way, the data are tested to see if they could be derived from a specific distribution, in this case a 3-parameter log-normal distribution. Since the accuracy of a 3-parameter log-normal distribution approximation of a normal (binomial or Poisson) distribution is unknown, this technique may be inconsistent with the postulated distribution model for geochemical data.

A brief discussion of the expected performance of two other threshold selection techniques (the 'mean plus 2 standard deviations' and '95^{th} percentile' procedures) will also be presented.

Although the 'mean plus 2 standard deviations', '95^{th} percentile', probability plot and gap statistic techniques assume different distribution models (a normal distribution, any distribution, a mixture of normal distributions and a 3-parameter log-normal distribution, respectively), all have been used in geochemical applications to discriminate populations and recognize anomalies. Since geochemical data theoretically can exhibit a variety of distribution forms, the choice of which technique to use in evaluating geochemical data becomes crucial. Bimodal distributions probably are best evaluated by the probability plot approach whereas unskewed unimodal distributions can be evaluated by either the 'mean plus 2 standard deviations' or gap statistic (without transformation) approaches. Unfortunately, with unimodal skewed distributions, the choice of which classification technique is most appropriate is not clear because these distributions may be produced by several distribution model mechanisms. With inadequate sampling, a poor approximation of a binomial or Poisson distribution by a normal distribution can produce a **single** distribution with positive (or negative) skewness. These situations might best be evaluated by the gap statistic (including transformation). Mixtures of more than one normal distribution can also exhibit positive (or negative) skewness, and these would best be evaluated with probability plots.

The important difference here is that, in the first case, we are left with a **single**

distribution, a situation which cannot provide additional geological insight because the data represent only one type of geological material. In the second case, we are left with **multiple** distributions which may be tested to ascertain if they represent different geological signatures. Obviously, with unimodal skewed distributions, decomposition of an hypothesized mixture of distributions is the only approach which allows for advances in the understanding of the geochemical data. Thresholds derived from this analysis can be then used to classify data so that a comparison with other categorical geologic variables may allow spatial or other relationships to be discovered (Popper 1968).

4.3 Previous Analysis

Miesch (1981) has already subjected the gap statistic to a *monte carlo* simulation in order to test whether it can correctly locate thresholds which discriminate mixtures of 3-parameter log-normal distributions. In Miesch's study, a **known** mean of 100 and standard deviation of 10 were used for a log-normal background distribution (α unknown). In all cases, 50 observations were used for each of the 1000 realizations for each data set structure. Miesch varied the distance between the two distributions (the higher mean was either 140 or 160 (square root of Mahalanobis distances of 4 and 6, respectively), with a standard deviation of 10) and the component percentage of the higher distribution (either 5 %, 10 % or 20 % of the data set). Results presented include the percentage of all anomalous data recognized as anomalous, the frequency that one or more background observations was recognized as anomalous in a realization, and the average number of false anomalies. A summary of the results at the 5 % confidence level are presented in Table 4.12 (Miesch 1981).

In addition, Miesch tested the threshold selection *power* of the gap statistic. In these tests, 1000 realizations of 50 observations drawn from a log-normal distribution

Table 4.12: Summary of the Performance of the Gap Statistic in a *Monte Carlo* Simulation of Threshold Selection

	Anomalous Mean	Anomaly Component Percentage		
		5	10	20
% of Anomalous	140	20.0	31.5	36.0
Values Detected	160	49.0	77.0	79.0
Frequency (%) of	140	11.0	18.0	15.5
False Anomalies	160	6.5	10.0	5.0
Average # of	140	3.55	2.40	1.25
False Anomalies	160	2.25	1.55	1.05

with **known** mean of 105 and standard deviation of 20 ($\alpha = 0$) were used to determine the frequency of anomaly recognition where none are present. Results indicate that, at the 5 % confidence level, an average of 5.2 % of the realizations were judged to have anomalous observations and an average of 13.5 observations per realization were judged to be anomalous.

Since an evaluation of the performance of the gap statistic has already been made for the 3-parameter log-normal distribution (Miesch 1981), albeit where the component distributions exhibit only marginal overlap ($\Delta = 4$ and 6), this study was concerned with how well the gap statistic performs in :

- rejecting the null hypothesis that the observations are derived from a single 3-parameter log-normal distribution where they have been derived from a mixture of two normal distributions, especially where the resulting distribution has a structure which is both positively skewed and unimodal, and
- evaluating how well the gap statistic selects thresholds which accurately discriminates the populations from mixtures of normal distributions.

If the gap statistic can be shown to be successful in discriminating mixtures of normal distributions, it may be applied to positively skewed distributions (which may be produced by a mixture of normal distributions or by inadequate sampling) to test whether they were derived from a 3-parameter log-normal (discrete) distribution, and thus a result of inadequate sampling.

4.4 Procedure

The different underlying assumptions of the above techniques can influence their classification performances when each is applied to different data set structures. To test their classification performance, in both an absolute and relative sense, both probability plot and gap statistic classification procedures were evaluated using *monte carlo* simulation.

The mixtures of normal distribution data set structures used in Chapter 3 were also used in this evaluation, and therefore a diverse range of distribution forms were represented. These data sets contained structural variation in the total number of observations (n), component percentages (ϖ), difference in means (μ_2) and difference in standard deviations (σ_2).

Classification of the data was made on all 160 data sets with each technique and for each data set. The number of observations misclassified in each case was recorded. The data sets were all mixtures of 2 distributions so only one threshold was defined in each technique and used to classify the data. Since misclassification can consist of observations from one population classified as being from the other population, and vice versa, both errors of omission and inclusion, and their sum, were recorded.

Threshold selection for the gap statistic utilized the midpoint of the maximum gap, regardless of its significance level. The entire range of values was considered, instead

Table 4.13: Average Mean, Standard Deviation and α Values for Positively Skewed Mixtures of Normal Distributions used to Transform the Stochastic Data in a Simulation of the Gap Statistic

Test # 2			
ϖ (%)	$\hat{\mu}_y$	$\hat{\sigma}_y$	$\hat{\alpha}$
70	3.386	0.339	-5.789
85	2.978	0.395	1.744
95	3.120	0.275	-2.999
Test # 4			
s_2	$\hat{\mu}_y$	$\hat{\sigma}_y$	$\hat{\alpha}$
10	3.521	0.360	-6.026
15	3.409	0.441	-3.535

of only the positive tail (Miesch 1981), because results of this evaluation are meant to be general; thus the existence of positive anomalies only could not be assumed. In all cases, the 3-parameter logarithmic transform (Aitchison and Brown 1957) of the data produced a new variable which, within the approximation stopping criterion of 0.0001, had a mean, standard deviation and skewness equal to 0, 1 and 0, respectively.

In the gap statistic simulation, stable estimates of μ, σ and α were not obtained for data set structures which, based on their population distribution, are un-skewed. This is because the corresponding distributions could be either slighty positively or negatively skewed and different transformation formulae (Equations 4.78 and 4.79) were required. The positively skewed data set structures which produced stable parameters for transform equation 4.78 are presented in Table 4.13 along with their average parameter values.

The significance of the largest gap also exceeded the critical value at the 95^{th} percentile for all data set structures except for some realizations from data set structures

Table 4.14: Comparison of the Average Maximum Gap Value (g) and the Corresponding Critical Value (c) for all Univariate Mixtures of Normal Distributions

Data Set Structure	$\bar{g}$	s_g	c
16	0.1484	0.0512	0.1240
2	0.0702	0.0215	0.7070
1 (Datum)	0.0466	0.0170	0.0395
3	0.0354	0.0005	0.0282
4	0.0359	0.0007	0.0231
5	0.0356	0.0006	0.0180
1 (Datum)	0.0466	0.0170	0.0395
6	0.0475	0.0088	0.0395
7	0.0569	0.0033	0.0395
8	0.0676	0.0042	0.0395
9	0.0708	0.0041	0.0395
10	0.0562	0.0020	0.0395
11	0.0454	0.0017	0.0395
1 (Datum)	0.0466	0.0170	0.0395
12	0.0683	0.0211	0.0395
13	0.1425	0.0337	0.0395
1 (Datum)	0.0466	0.0170	0.0395
14	0.0360	0.0021	0.0395
15	0.0421	0.0077	0.0395

1, # 2, # 14, # 15 and # 16. In these cases, only 3, 3, 1, 5 and 6 data set realizations (out of 10) had their largest gaps exceed the critical value at the 95^{th} percentile, respectively. A summary of the means and standard deviations of the largest gap value for each data set structure is presented in Table 4.14 along with the critical value corresponding to the 95^{th} percentile for the size of the data set.

Likewise, the probability plot technique involved the selection of only one threshold. *RDML* parameter estimates from the study of Chapter 3 were used. Determination of the optimal threshold using these parameter estimates was made using the minimum classification error approach described in Chapter 2.

Likelihood ratio tests were performed on the 10 data set realizations of each data set structure to determine the significance level of the parameters of the second mode in the distribution model (a test of the *power* of the probability plot technique). These produced average significance levels greater than the 99^{th} percentile for all data set structures except # 9 and # 10, which had average significance levels at the 50^{th} and 97^{th} percentiles, respectively. Similar tests to determine the significance level of the parameters of a third mode were made on one randomly selected data set realization from each data set structure. In no case did the significance level of the maximum likelihood estimates of the parameters of this third mode exceed the 85^{th} percentile.

4.5 Results

The average number and variance of the total number of misclassification errors, errors of omission and errors of inclusion were determined for each data set structure and each classification technique. These are presented in Tables 4.15 and 4.16 respectively. Figures 4.14, 4.15, 4.16 and 4.17 depict the relative efficiencies (performance) of the probability plot and gap statistic techniques with variation in data set structure.

Table 4.15: Classification Error Comparison for the Gap Statistic and Probability Plot Classification Techniques

Structure Label	Variable	Gap Statistic		Probability Plots	
Test # 1	n				
16	50	1.50	(2.80)	1.50	(2.80)
2	100	4.00	(1.87)	3.60	(1.18)
1 (Datum)	200	4.50	(2.07)	4.60	(2.22)
3	300	7.30	(2.67)	7.30	(2.36)
4	400	9.00	(1.76)	9.40	(2.91)
5	500	9.90	(3.70)	9.50	(3.54)
Test # 2	$\varpi(\%)$				
1 (Datum)	50	4.50	(2.07)	4.60	(2.22)
6	70	42.10	(26.47)	5.30	(2.41)
7	85	91.00	(31.85)	5.00	(3.09)
8	95	124.10	(18.97)	2.00	(1.05)
Test # 3	μ_2				
9	25	62.30	(6.00)	68.40	(8.28)
10	30	32.20	(3.33)	36.40	(4.99)
11	35	13.50	(2.92)	20.20	(12.26)
1 (Datum)	40	4.50	(2.07)	4.60	(2.22)
12	45	1.10	(0.88)	1.10	(0.74)
13	50	0.50	(0.71)	0.40	(0.70)
Test # 4	σ_2				
1 (Datum)	5	4.50	(2.07)	4.60	(2.22)
14	10	34.80	(19.05)	17.60	(3.57)
15	15	48.00	(20.59)	32.30	(4.83)

Values presented are the mean and standard deviation (in parentheses) of the number of observations which were misclassified for each realization.

Table 4.16: Errors of Omission Versus Inclusion for the Gap Statistic and Probability Plot Classification Techniques

Structure Label	Variable	Gap Statistic	Probability Plots
Test # 1	n		
16	50	86	86
2	100	39	19
1 (Datum)	200	64	43
3	300	45	41
4	400	50	55
5	500	41	40
Test # 2	$\varpi(\%)$		
1 (Datum)	50	64	43
6	70	1	54
7	85	0	18
8	95	0	25
Test # 3	μ_2		
9	25	52	40
10	30	52	51
11	35	42	16
1 (Datum)	40	64	43
12	45	27	36
13	50	80	50
Test # 4	σ_2		
1 (Datum)	5	64	43
14	10	23	80
15	15	50	92

Values are the number of population A data (with the higher mean) classified as population B data, presented as average % of the total classification error.

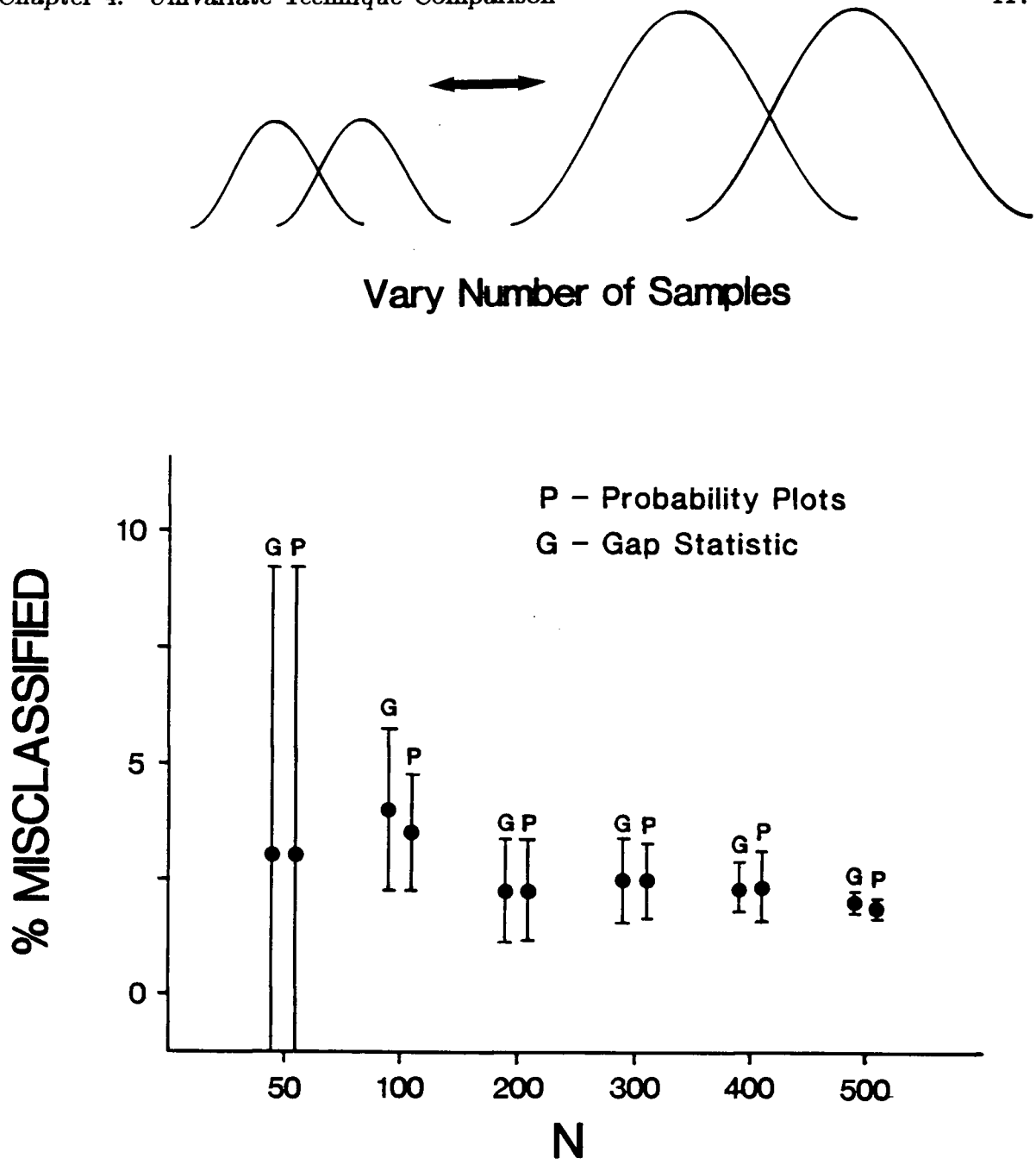
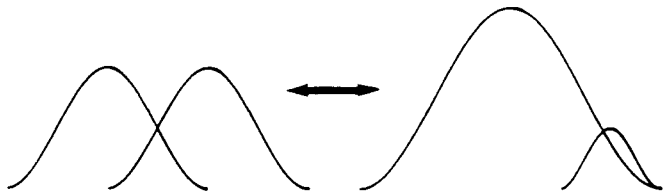


Figure 4.14: Plot Comparing the Gap Statistic and Probability Plot Classification Techniques with Data Sets Composed of Mixtures of Normal Distributions Containing Different Numbers of Observations

Dots represent mean of ten values and vertical bars represent ± 1 standard deviation.



Vary Relative Sizes of Populations

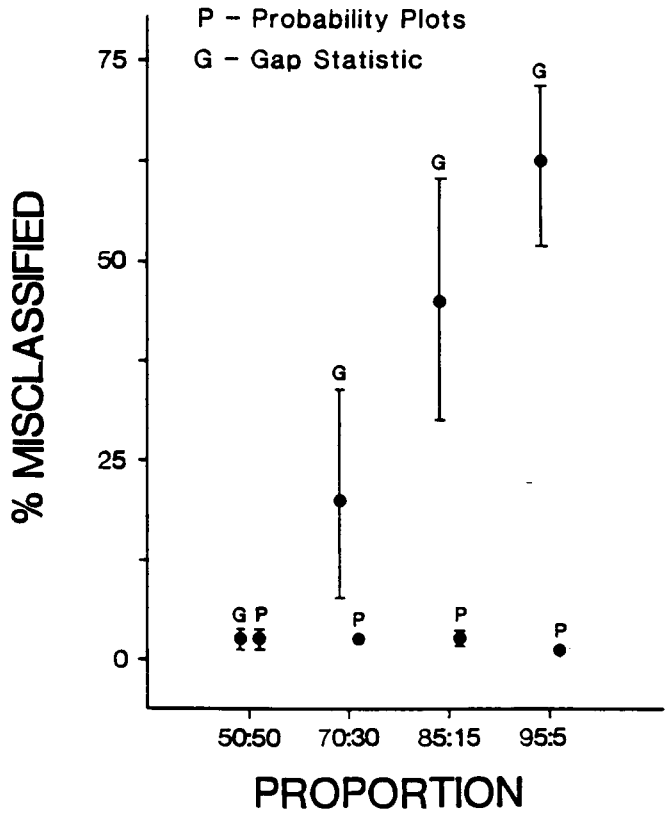


Figure 4.15: Plot Comparing the Gap Statistic and Probability Plot Classification Techniques with Data Sets Composed of Mixtures of Normal Distributions Containing Differing Proportions of Component Distributions

Dots represent mean of ten values and vertical bars represent $\pm$ 1 standard deviation.

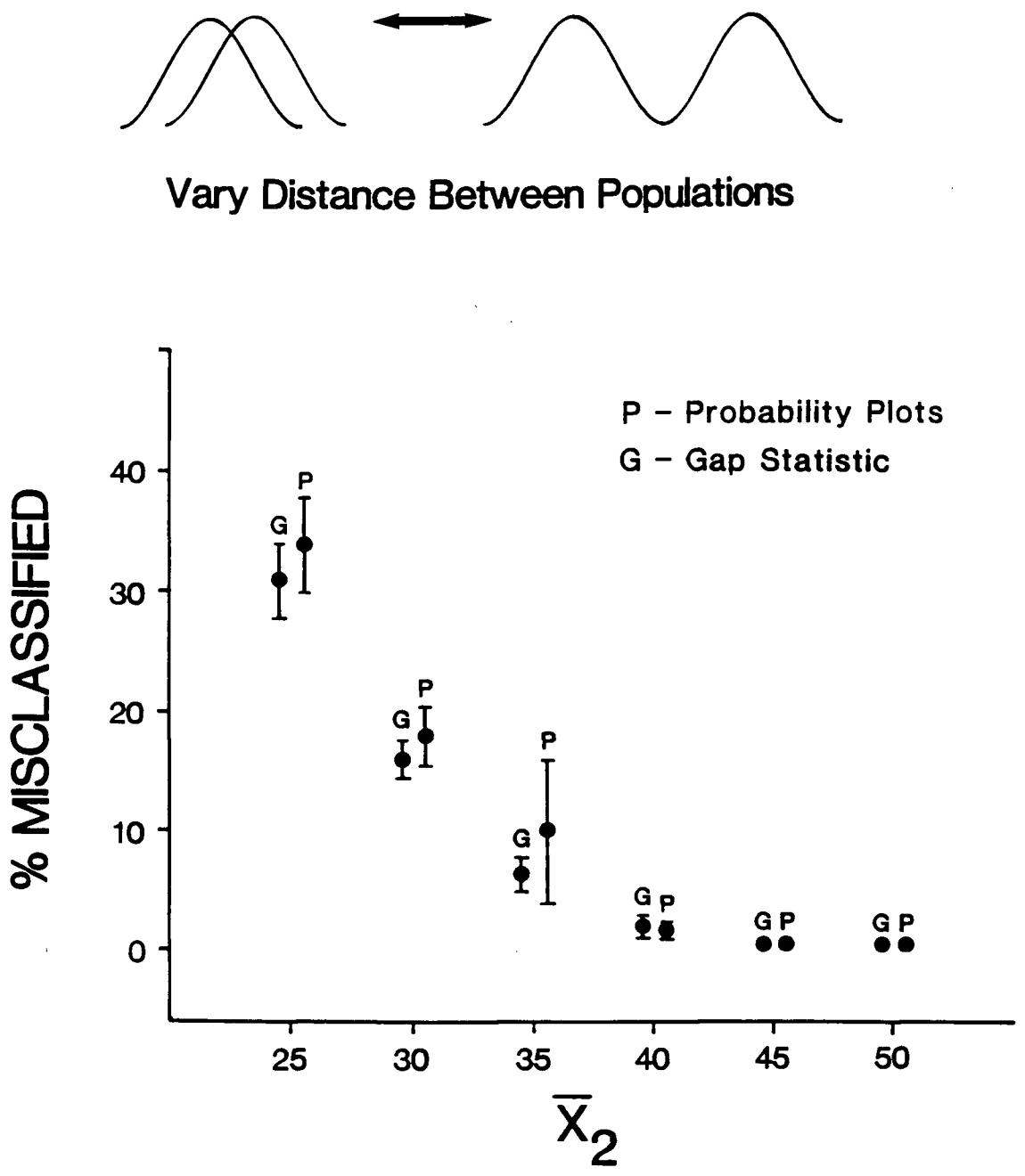
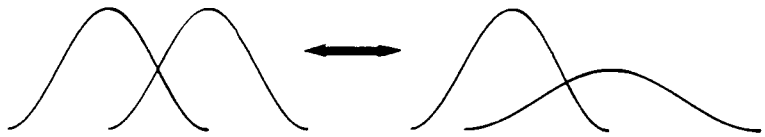


Figure 4.16: Plot Comparing the Gap Statistic and Probability Plot Classification Techniques with Data Sets Composed of Mixtures of Normal Distributions Containing Different Distances Between the Component Means

Dots represent mean of ten values and vertical bars represent $\pm$ 1 standard deviation.



Vary Population Standard Deviations

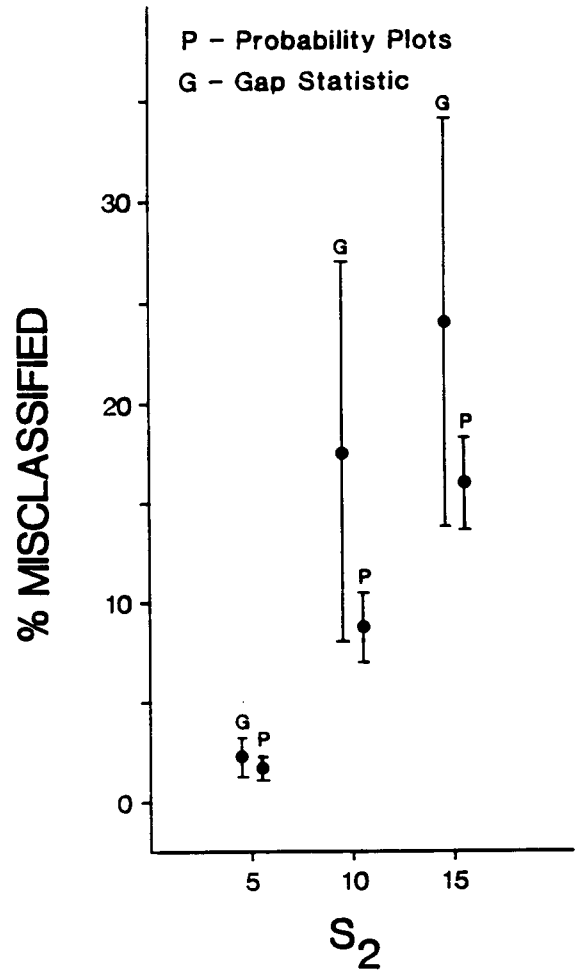


Figure 4.17: Plot Comparing the Gap Statistic and Probability Plot Classification Techniques with Data Sets Composed of Mixtures of Normal Distributions Containing Different Component Standard Deviations

Dots represent mean of ten values and vertical bars represent $\pm$ 1 standard deviation.

4.6 Discussion

In general, both classification techniques produced comparable results, with only a few exceptions. Trends observed in the total amount of classification error include :

- a decreasing average amount of total misclassification (but equal percentage) and decreasing amount of variance of total misclassification for both the probability plot and gap statistic techniques occurs with increasing number of observations per data set (n),
- a constant total amount of misclassification for the probability plot technique with increasing disparity in the percentages of the component distributions (ϖ); but, a rapidly increasing total amount of misclassification and increasing amount of variability for the gap statistic approach with increasing disparity in the percentages of the component distributions,
- a decreasing total amount of misclassification for both techniques with increasing distance between component distribution mean values (increasing $|\mu_1 - \mu_2|$ relative to σ_p, thus increasing Δ),
- an increasing total amount of misclassification with increasing disparity of the component distribution standard deviations (increasing σ_p relative to $|\mu_1 - \mu_2|$, thus decreasing Δ) for both techniques; the gap statistic technique exhibits a higher and more variable total amount of misclassification than the probability plot approach.

In terms of total misclassification, the probability plot technique performed substantially better then the gap statistic approach if the data set had an initial skewness substantially different from 0. Ironically, these were cases where the 3 parameters of

the transformation operator for the gap statistic were most stable, and where the gap statistic has the most utility in identifying the form of the distribution. The poor performance of the gap statistic in these cases may be due to the various theoretical considerations (limitations) described above.

In terms of the relative proportions of the different types of classification errors (errors of omission and errors of inclusion), these two techniques produced substantially different results. The theory of the probability plot technique requires that the minimum number of total classification errors occur where the number of errors of omission equals the number of errors of inclusion. The average percentage of total classification error where the observations from the distribution with the larger mean are classified as part of the distribution with the smaller mean is 46 % ($\pm$ 23 %). This cannot be considered to be substantially different from the expected 50 %.

The results from the gap statistic technique are very different. Where data sets have equal component percentage values, such as in tests # 1, # 3 and # 4, no appreciable difference between the number of errors of omission and inclusion can be detected (they have a combined mean of 50 % $\pm$ 18 % observations from the distribution with the larger mean classified as part of the distribution with the smaller mean). However, if the component proportions are not equal, observations from the distribution with the smaller mean are classified at least 99 % of the time as part of the distribution with the larger mean (Table 4.16). The threshold is chosen at a value located on the **lower tail of the lower distribution.** This clearly indicates that, for these data set structures, the gap statistic only tests for discordancy; it can not adequately define a threshold, even though it is capable of indicating the existence of outliers from another normal distribution.

4.7 Other Threshold Selection Techniques

Application of the 'mean plus 2 standard deviations' and '95^{th} percentile' threshold selection techniques to these data set structures will clearly classify the data inadequately. Using the '95^{th} percentile' technique, the upper 5 % of the data set will **always** be classified as anomalous, whether an anomalous distribution exists or not. Thus only data set structure # 8 will be classified optimally. Likewise, thresholds at the 'mean plus 2 standard deviations' will classify a variable percentage of the data as anomalous, depending on the nature (μ, σ and ϖ) of the component distributions. Since these mixtures of normal distributions are not, themselves, normal, this approach requires an assumption about the data set distribution which is not met. Clearly, both of these procedures assume the existence of a distribution which is not represented by the data and are not flexible enough to be applied to cases where mixtures of normal distributions are present.

4.8 Conclusions

The gap statistic procedure (including the 3-parameter logarithmic tranformation) applied to mixtures of normal distributions appears to successfully test for discordancy (for the existence of outliers from a second population). It also appears capable of classifying un-skewed mixtures of normal distributions with error which is comparable, if not equivalent, to the probability plot technique.

However, the gap statistic classifies observations from skewed mixtures of normal distributions at a significantly lower level of performance than the probability plot technique. This large discrepancy in performance between the gap statistic approach and the probability plot technique indicates that the probability plot technique is an overall superior data classification procedure for mixtures of normal distirbutions. In fact,

due to the optimal properties of maximum likelihood estimation, the probability plot technique represents the 'best' possible approach to classification (of the approaches considered).

Other classification procedures such as the 'mean plus 2 standard deviations' and the '95^{th} percentile' techniques are generally inappropriate for the data set structures expected for geochemical data.

Chapter 5

Multivariate Technique Comparison

"Basic research is when I'm doing what I don't know I'm doing."
Werner Von Braun (1964)

"To err is human, but to really foul things up requires a computer."
Paul Ehrlich (1978)

Although results from Chapter 4 demonstrate that the probability plot technique classifies the stochastically generated mixtures of normal distributions at an equal or superior level of performance than the gap statistic, several data set structures could not be classified at an acceptable precision level by either technique. If the Mahalanobis Distance between two component distributions is small ($\Delta < 2$), neither technique classifies the data successfully. For example, where the two means of the component distributions are 20 and 30 with a common standard deviation of 5 ($\Delta = 2$) and equal component distribution percentages, the smallest possible amount of data misclassification will be approximately one third of the data set (theoretically 32 %). Based on the average results of Chapter 4, the probability plot approach misclassified 32 % of the data, whereas the gap statistic approach misclassified 36 %. In cases such as these, where the component distributions overlap substantially, analysis of other geochemical variables may allow better classification.

Obviously, any systematic univariate analysis of geochemical data should be directed toward the discovery of those variables which manifest the least amount of population

overlap, and thus can be used to discriminate the populations with the least amount of classification error. Clearly, in the ideal case, the maximum level of classification success will be achieved using those geochemical variables which manifest the largest difference (Δ) between the component distributions. Although philosophical considerations require that at least one geochemical variable of this type exist (otherwise there would be no difference between the component distributions), practical limitations may prevent either the recording of these unknown geochemical variables or their subsequent analysis. Factors preventing data collection include budget constraints, lack of adequately sensitive analytical techniques, or lack of knowledge that a certain geochemical variable has the potential to discriminate populations. Conversely, an abundance of recorded geochemical variables may make it too costly for the geochemist to evaluate each variable independently and the discriminating variables may never be identified. In either case, the geochemist is left without a single geochemical variable with which to adequately discriminate the populations. As a result, multivariate analysis of the data set may be required to achieve an acceptably low amount of population misclassification.

5.1 Background Characterization Approach

One multivariate approach used to reduce the amount of population misclassification is the background characterization approach (*BCA*; Stanley and Sinclair 1987, Day et al. 1987). This involves the formulation of a linear background model to describe the variation observed in the background data. Any background model used to assist in the classification of geochemical data from overlapping mixtures of distributions must be consistent with the underlying relationships among the variables used in the model. As described in Chapter 1, two general situations exist, one involving a causal relationship between geochemical variables and one involving a common association

between geochemical variables.

The first situation involves a single variable, measured with error, which is **causally** related to other variables, also measured with error. Situations such as these can be examined appropriately with a regression approach, where one variable is cast as a linear function of other variables. For example, in background soil samples, Cu, Zn, Co and Ag may occur adsorbed to poorly crystalline Fe- and Mn-oxyhydroxides. The abundance of these oxyhydroxides could, thus, control the abundance of Cu, Zn, Co and Ag. If this hypothesis is correct, the abundances of Cu, Zn, Co and Ag may all be described (predicted) by functions of the measured Fe and Mn concentrations.

The alternative situation is where Fe and Mn concentrations have not been measured. In this case, the hypothesized Fe- and Mn-oxyhydroxide 'causative' factor cannot be observed, and thus is considered a *latent* variable. The *manifest* variables, those which have been measured (Cu, Zn, Co and Ag), are hypothesized to be (linearly) functionally related to this *latent* variable, but a regression model cannot be formulated because no estimates of this *latent* variable are available. A principal components (PC) analysis would be an appropriate examination method for situations such as these because the PC, themselves, would represent estimates of the un-observed *latent* variables.

In reality, because of the numerous chemical and physical controls on the abundance of elements in geological materials, the difference between these two situations may be obscure; however, a clear cut theoretical distinction exists between them. For example, an un-measured variable, such as pH, may control the abundances of Cu, Zn, Co and Ag as well as Fe and Mn. Thus, although Cu may be regressed against Fe and Mn, the combination of Cu, Fe and Mn may also be used to determine a PC representing pH. Clearly, the appropriateness of the background model will depend to a large extent on the way the model is designed and its relevance to specific geochemical processes.

After selection of a background model philosophy, the *BCA* involves selection of thresholds using the techniques described in Chapter 3, with the exception that, because of extreme overlap between adjacent populations, a single threshold cannot classify the data adequately. As a result, two thresholds are chosen (conventionally at 2 standard deviations above and below the respective means) which define 3 data ranges of two different types. Two of these data ranges are of one type, composed, predominantly, of observations from the same population (B or A; Figure 2.8). These can be thought of as 'pure' data ranges, although in reality a small proportion of the other population are included in each. The third data range would be of a different type, containing a significant number of observations from both populations (A and B). This data range would be bounded by the two chosen thresholds (Figure 1.2) and is called the 'range of population overlap'. Using the observations from the un-contaminated data range for population A or B, the parameters of the background model can be estimated and the resulting model used to 'characterize' the variation in the background (A or B) population. The parameters for this model are estimated using only, or at least predominantly, data from the 'background' population.

This background model can then be applied to observations from the range of population overlap, where accurate classification based on the variable of interest is not possible. Observations from the background population which occur in the range of population overlap can be expected to 'react' to the background model in a way similar to those which occur in the range of the 'pure' background population because they all are derived from the same population. Observations in the range of population overlap which are members of another (anomalous) population would be expected to 'react' differently to the background model than the background observations (Figure 1.2). This is because, derived from a different population, the relationships among these variables generally would be different.

In this way, observations in the range of population overlap for one variable can be classified using additional information made available through use of other variables in the background model. If both populations comprise a significant proportion of the data set (a population discrimination situation), two 'background' models can be developed (one for each population) to help classify those observations in the range of population overlap (Figures 1.1 and 1.2). For anomaly recognition situations, only one background model can be developed because of the paucity of observations derived from one of the populations. Clearly, application of a background model derived from a population with numerous observations has the potential to be more successful because the parameter estimates for that model will be more accurate.

Unfortunately, a background model determined using data from a 'pure' data range will not be the same as one determined using all the data from the associated population. This is because the distribution is **truncated** by the bounding threshold. A bias will exist which, if the amount of truncation is small, may not be significant, but if the amount of truncation is large, may totally invalidate the *BCA* to anomaly recognition or population discrimination. Fortunately, the effects of the truncation can be reduced, if not eliminated, by correcting for the truncation. This correction is made on the observed parameters of the truncated distribution, so that the corrected values are estimates of the parameters of the un-truncated population. Both regression and *PC* analysis can be performed using these 'truncation corrected' parameters instead of the truncated data (see below). A derivation of the procedure for correcting the observed truncated statistical sample parameters is presented in Appendix E.

Simulation of the *BCA* to anomaly recognition was undertaken to evaluate how well the procedure can produce accurate background models and residuals for the data from the range of population overlap which are similar (and predictable) to those from the 'pure' background observations. Obviously, any simulation of the mathematical

and statistical operations involved in application of the *BCA* must be performed on data sets which meet the assumptions required. The method of data set generation, described in Chapter 3, is consistent with the regression models and *PC* background model used in this analysis.

5.2 Multivariate Data Set Structures

Multivariate data sets were generated with specific means, standard deviations and correlation structures according to the linear congruential procedures described in Chapter 3. Twelve different trivariate (Y, X_1, X_2) data set structures were used to determine the performance (efficiency) of the *BCA* to classification of data for a range of the extent of overlap between populations. Each data set structure is multivariate normal with means of 20, 40 and 60 for the three variables. Standard deviations are 4, 6 and 8, respectively. Ten realizations of each data set structure were generated and all contained 200 observations.

The only parameters allowed to vary between data set structures are the off-diagonal terms of the correlation matrix. Presented below are the correlation structures (matrices) used for the generation of the multivariate normal data set realizations (# 17 through # 28), as well as those which were examined in the course of this study and which are discussed below :

$$\begin{pmatrix} 1.0 & 0.7 & 0.7 \\ 0.7 & 1.0 & 0.1 \\ 0.7 & 0.1 & 1.0 \end{pmatrix}, \qquad \text{Data Set Structure \# 17} \tag{5.85}$$

$$\begin{pmatrix} 1.0 & 0.5 & 0.5 \\ 0.5 & 1.0 & 0.1 \\ 0.5 & 0.1 & 1.0 \end{pmatrix}, \qquad \text{Data Set Structure \# 18} \tag{5.86}$$

$$\begin{pmatrix} 1.0 & 0.3 & 0.3 \\ 0.3 & 1.0 & 0.1 \\ 0.3 & 0.1 & 1.0 \end{pmatrix}, \quad \text{Data Set Structure \# 19} \tag{5.87}$$

$$\begin{pmatrix} 1.0 & 0.9 & 0.1 \\ 0.9 & 1.0 & 0.1 \\ 0.1 & 0.1 & 1.0 \end{pmatrix}, \quad \text{Data Set Structure \# 20} \tag{5.88}$$

$$\begin{pmatrix} 1.0 & 0.7 & 0.1 \\ 0.7 & 1.0 & 0.1 \\ 0.1 & 0.1 & 1.0 \end{pmatrix}, \quad \text{Data Set Structure \# 21} \tag{5.89}$$

$$\begin{pmatrix} 1.0 & 0.5 & 0.1 \\ 0.5 & 1.0 & 0.1 \\ 0.1 & 0.1 & 1.0 \end{pmatrix}, \quad \text{Data Set Structure \# 22} \tag{5.90}$$

$$\begin{pmatrix} 1.0 & 0.3 & 0.1 \\ 0.3 & 1.0 & 0.1 \\ 0.1 & 0.1 & 1.0 \end{pmatrix}, \quad \text{Data Set Structure \# 23} \tag{5.91}$$

$$\begin{pmatrix} 1.0 & 0.9 & 0.5 \\ 0.9 & 1.0 & 0.1 \\ 0.5 & 0.1 & 1.0 \end{pmatrix}, \quad \text{Data Set Structure \# 24} \tag{5.92}$$

$$\begin{pmatrix} 1.0 & 0.9 & 0.3 \\ 0.9 & 1.0 & 0.1 \\ 0.3 & 0.1 & 1.0 \end{pmatrix}, \quad \text{Data Set Structure \# 25} \tag{5.93}$$

$$\begin{pmatrix} 1.0 & 0.7 & 0.5 \\ 0.7 & 1.0 & 0.1 \\ 0.5 & 0.1 & 1.0 \end{pmatrix}, \quad \text{Data Set Structure \# 26} \tag{5.94}$$

$$\begin{pmatrix} 1.0 & 0.7 & 0.3 \\ 0.7 & 1.0 & 0.1 \\ 0.3 & 0.1 & 1.0 \end{pmatrix}, \quad \text{Data Set Structure \# 27} \tag{5.95}$$

$$\begin{pmatrix} 1.0 & 0.5 & 0.3 \\ 0.5 & 1.0 & 0.1 \\ 0.3 & 0.1 & 1.0 \end{pmatrix}, \qquad \text{Data Set Structure \# 28} \tag{5.96}$$

Realizations for correlation structures such as :

$$\begin{pmatrix} 1.0 & 0.9 & 0.9 \\ 0.9 & 1.0 & 0.1 \\ 0.9 & 0.1 & 1.0 \end{pmatrix}, \qquad \text{Data Set Structure \# 29} \tag{5.97}$$

and :

$$\begin{pmatrix} 1.0 & 0.9 & 0.7 \\ 0.9 & 1.0 & 0.1 \\ 0.7 & 0.1 & 1.0 \end{pmatrix}, \qquad \text{Data Set Structure \# 30} \tag{5.98}$$

could not be produced because these matrices are not positive definite and thus not legitimate correlation structures. Finally, five data set structures which were not used to generate stochastic data sets, but which are discussed below, are :

$$\begin{pmatrix} 1.0 & 0.1 & 0.1 \\ 0.1 & 1.0 & 0.7 \\ 0.1 & 0.7 & 1.0 \end{pmatrix}, \qquad \text{Data Set Structure \# 31} \tag{5.99}$$

$$\begin{pmatrix} 1.0 & 0.7 & 0.7 \\ 0.7 & 1.0 & 0.7 \\ 0.7 & 0.7 & 1.0 \end{pmatrix}, \qquad \text{Data Set Structure \# 32} \tag{5.100}$$

$$\begin{pmatrix} 1.0 & Y & Y \\ Y & 1.0 & X \\ Y & X & 1.0 \end{pmatrix}, \qquad \text{Data Set Structure \# 33} \tag{5.101}$$

$$\begin{pmatrix} 1.0 & Y & 0.1 \\ Y & 1.0 & X \\ 0.1 & X & 1.0 \end{pmatrix}, \qquad \text{Data Set Structure \# 34} \tag{5.102}$$

and :

$$\begin{pmatrix} 1.0 & 0.1 & 0.1 \\ 0.1 & 1.0 & X \\ 0.1 & X & 1.0 \end{pmatrix}, \qquad \text{Data Set Structure \# 35} \tag{5.103}$$

where X and Y equal 0.9, 0.7, 0.5 and 0.3, and $Y \leq X$.

The correlation matrices used to generate stochastic trivariate data sets (# 17 through # 28) all take the form of oblate ellipsoids. These are meant to encompass a reasonably complete spectrum of positive definite correlation structures. Correlation structures with high correlation among the independent variables (data set structures # 33, # 34 and # 35) were avoided because, as discussed in Appendix D, the reduced major axis regression solution is numerically unstable if there are high correlations among these variables, preventing a solution from being determined (see below).

5.3 Procedure

Statistical models used to represent the background data include *PC* as well as two *ML* regression methods. The first regression method assumes errors exist in all variables and the regression coefficients are determined by minimizing :

$$\sigma_\epsilon^2 = \frac{\sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij})^2}{n(1 + \sum_{j=1}^{p} \frac{b_j^2}{\lambda_j})}. \tag{5.104}$$

A formal derivation of this method is presented in Appendix D. In order to determine the regression coefficients, the assumption that :

$$\frac{s_y^2}{s_{x_j}^2} = \lambda_j = \frac{\sigma_\epsilon^2}{\sigma_{\eta_j}^2}, \tag{5.105}$$

was used, where s_y and s_{x_j} are the observed variances of the data and σ_ϵ and σ_{η_j} are the variances of the errors. This approach, with error assumed in each variable, is referred to as the 'reduced major axis' (*RMA*) regression and is based on the premise that a structural relation exists between the dependent and independent variables.

The other regression function evaluated is the 'ordinary least squares' (*OLS*) approach, where error is assumed to exist in only the dependent variable. Although this is known to be a poor assumption for geochemical data, the *OLS* regression coefficients were calculated in the process of producing the *RMA* regression coefficients, so no addition effort was required. Furthermore, *RMA* regression options are not yet generally included in commercially available statistical packages; thus, results from the *OLS* truncation analysis may be relevent. Both of these approaches are *ML* solutions for regression models involving different sets of assumptions about the placement of errors in the observations (Fuller 1987).

In order to evaluate the efficiency of the *BCA*, the parameters of the relevent background functions, as well as the residuals calculated from these functions, must be evaluated. This is because, although the residuals will allow an assessment of how well the background function represents the truncated background data and thus how well the procedure performs in recognizing background observations from the range of population overlap, the parameters of the background function will allow an assessment of the overall stability of the results.

5.3.1 Parameter Analysis

The parameters for each of the background functions based on the population, statistical sample, truncated statistical sample and truncation corrected statistical sample were evaluated to determine the stability of the parameter estimates of the truncated statistical sample and truncation corrected statistical sample parameter estimates.

5.3.1.1 Population Parameters

For each data set structure, a principal components (*PC*) analysis and both the *ML* regression analysis with error in only the dependent variable (*OLS* regression analysis)

and the *ML* regression analysis with error in every variable (*RMA* regression analysis) were performed on the **population.** The *PC* were derived using the population correlation matrix, but the regression coefficients had to be derived in a different manner.

By multiplying the population covariance matrix (Σ) by the number of observations in the statistical samples produced using that covariance matrix, the population corrected cross product matrix can be derived :

$$\begin{pmatrix} ss_{yy} & ss_{yx_1} & ss_{yx_2} \\ ss_{x_1y} & ss_{x_1x_1} & ss_{x_1x_2} \\ ss_{x_2y} & ss_{x_2x_1} & ss_{x_2x_2} \end{pmatrix}, \tag{5.106}$$

where ss_{yy} and $ss_{x_jx_j}$ are the sum of corrected squares of y and x_j, respectively :

$$ss_{yy} = E\left(\sum_{i=1}^{n}(y_i - \mu_y)^2\right), \tag{5.107}$$

$$ss_{x_jx_j} = E\left(\sum_{i=1}^{n}(x_{ij} - \mu_{x_j})^2\right), \tag{5.108}$$

and ss_{yx_j} and $ss_{x_jx_k}$ are the sum of corrected cross products of y and x_j, and x_j and x_k, respectively :

$$ss_{yx_j} = E\left(\sum_{i=1}^{n}(y_i - \mu_y)(x_{ij} - \mu_{x_j})\right), \tag{5.109}$$

$$ss_{x_jx_k} = E\left(\sum_{i=1}^{n}(x_{ij} - \mu_{x_j})(x_{ik} - \mu_{x_k})\right). \tag{5.110}$$

This matrix can be partitioned in terms of the dependent (Y) and independent (X_1 and X_2) variables :

$$\left(\begin{array}{c|cc} ss_{yy} & ss_{yx_1} & ss_{yx_2} \\ \hline ss_{x_1y} & ss_{x_1x_1} & ss_{x_1x_2} \\ ss_{x_2y} & ss_{x_2x_1} & ss_{x_2x_2} \end{array}\right) \tag{5.111}$$

and recast in matrix notation, giving :

$$\left(\begin{array}{c|c} \vec{Y}'\vec{Y} & \vec{Y}'\tilde{X} \\ \hline \tilde{X}'\vec{Y} & \tilde{X}'\tilde{X} \end{array}\right) \tag{5.112}$$

where :

$$\vec{Y} = \begin{pmatrix} \check{y}_1 \\ \check{y}_2 \\ \vdots \\ \check{y}_n \end{pmatrix}, \tag{5.113}$$

$$\tilde{X} = \begin{pmatrix} \check{x}_{11} & \check{x}_{12} \\ \check{x}_{21} & \check{x}_{22} \\ \vdots & \vdots \\ \check{x}_{n1} & \check{x}_{n2} \end{pmatrix}, \tag{5.114}$$

assuming :

$$\check{x}_{ij} = x_{ij} - \mu_{x_j}, \tag{5.115}$$

and :

$$\check{y}_i = y_i - \mu_y. \tag{5.116}$$

Since the b_j terms in *OLS* regression can be determined using (Draper and Smith 1981) :

$$\vec{B} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\vec{Y}, \tag{5.117}$$

determination of these coefficients can be achieved using the partitioned matrix notation terms of the population corrected cross product matrix.

Similarly, since the sum of squared residuals numerator in *RMA* regression :

$$\sum_{i=1}^{n}(\check{y}_i - \sum_{j=1}^{p} b_j\check{x}_{ij})^2, \tag{5.118}$$

can be recast in matrix notation as :

$$\vec{Y}'\vec{Y} - 2\vec{B}'\tilde{X}'\vec{Y} + \vec{B}'\tilde{X}'\tilde{X}\vec{B}, \tag{5.119}$$

the expected **population** sum of squared residuals numerator can be determined by multiplication of the partitioned matrix notation terms of the corrected cross product

matrix with :

$$\vec{B} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}. \tag{5.120}$$

The denominator term for the *RMA* regression sum of the squared residual formula :

$$n(1 + \sum_{j=1}^{p} \frac{b_j^2}{\lambda_j}), \tag{5.121}$$

can then be determined by using the population variances (the diagonal terms of Σ) to calculate the λ_j terms. The $\vec{B}$ solution which minimizes the *RMA* sum of squared residuals can then be determined through non-linear optimization (minimization) using the SIMPLEX method (Figure 2.7) of Nash (1979) and Caceci and Cacheris (1984). Inital estimates of the b_j terms were, in all cases in this analysis, the *OLS* estimates.

5.3.1.2 Sample Parameters

A similar *PC*, *OLS* and *RMA* regression analysis was also performed on the stochastic realizations. In these cases, solution of the *PC* was accomplished in a straightforward manner, using S, the statistical sample covariance matrix. Similarly, solution of both *OLS* and *RMA* regression analyses was accomplished using the actual data, and in the case of *RMA* regression, the statistical sample variances.

5.3.1.3 Truncated Sample Parameters

Each of the realizations was also subjected to four levels of truncation. Truncation was made at the 95^{th}, 85^{th}, 70^{th} and 50^{th} percentiles of Y's (the first variable - with a mean = 20 and standard deviation = 4) population distribution. These correspond to Y-values of 26.580, 24.148, 22.096 and 20.000, respectively, and to z-scores of 1.645, 1.037, 0.524 and 0.000. Truncation reduced the size of each realization to approximately 190, 170, 140 and 100 observations, respectively.

Each of these four truncated realizations was analyzed with standard *PC*, *OLS* and *RMA* multivariate regression techniques. The methods for solution of these analyses were identical to the analyses performed on the observed statistical sample without truncation (above).

5.3.1.4 Truncation Corrected Sample Parameters

The second set of analyses was performed on the same truncated realizations after 'correction' for truncation (see Appendix E). This truncation correction can only be made on the statistical parameter estimates of the truncated realizations, and not on the individual observations themselves. Those observations which were truncated from the realization were lost. As a result, the subsequent calculation of the *PC*, *OLS* and *RMA* regression coefficients was made using the truncation corrected statistical sample covariance matrix (S_c) in a manner similar to that described for the population derived *PC*, *OLS* and *RMA* regression coefficients which were made using Σ.

Thus, for each realization, 10 *PC*, *OLS* and *RMA* regression analyses were perfomed. These consist of analyses on the population, on the full realization, on the four levels of truncation of the realization, and the four levels of truncation on the 'truncation corrected' realization. This allows an evaluation of how well each analysis technique performs in approximating the true *PC* and regression coefficients with increasing amounts of truncation.

5.3.2 Residual and Score Analysis

After determination of the parameters of the three background functions (*OLS*, *RMA* and *PC*), the residuals and scores were calculated for both the observations used to determine the background function (the un-truncated observations) and those observations which were truncated. No residuals could be calculated using the population

parameters, because in this case, no actual observations exist. In the case of the two regression functions, the residuals :

$$r_i = y_i - \hat{y}_i, \tag{5.122}$$

were calculated, where $\hat{y}$ is the predicted value of y_i for the relevant regression function. Similarly, the second and third *PC* scores were calculated. In addition, the pythagorean distances between the first *PC* and the observations were calculated according to :

$$d_{i23} = \sqrt{\frac{p_{i2}^2}{\lambda_{e2}} + \frac{p_{i3}^2}{\lambda_{e3}}}, \tag{5.123}$$

where d_{i23} is the i^{th} distance, p_{i2} and p_{i3} are the i^{th} second and third *PC* scores and λ_{e2} and λ_{e3} are the eigenvalues for *PC* # 2 and # 3. This distance is essentially, the radial (elliptical) distance between the major axis (*PC* # 1) and the observation, and thus should be distributed χ^2 for the un-truncated data sets. Although not discussed in the text, all statistics for this distance calculation have been determined and are included in the appropriate tables and figures.

Clearly, the expected values of the residuals and scores for the background (un-truncated) data used to determine the coefficients for the background models are zero. This is not true for the expectations of the residuals and scores for the truncated data and may not be true for the background and truncated data if the truncation corrected parameters are used to calculate the residuals and scores. A comparison of these different residuals and scores will indicate how sensitive the different background functions are to different levels of population truncation, with and without truncation correction.

5.4 Results

Results (means and standard deviations of 10 realizations) from the analysis of the parameter estimates for the *BCA* regression and *PC* background models for data set structure # 17 and # 23 are presented in Tables F.58 through F.69 of Appendix F. Figures 5.18 through 5.21 graphically depict these results for data set structure # 17 only.

Results from the analysis of the residuals and scores of the truncated data for the *BCA* regression and *PC* background models for data set structure # 17 and # 23 are presented in Tables G.70 through G.89 of Appendix G. As above, Figures 5.22 through 5.23 graphically depict these results for data set structure # 17 only.

This subset of data set structures (# 17 and # 23) was found to represent a good cross-section of results from these analyses, ranging from highly correlated data sets (oblate ellipsoids – # 17) to poorly correlated data sets (approximate spheres – # 23), and is presented *in lieu* of the results from all data set structures because of space considerations.

5.5 Discussion

Results from the simulation of the *BCA* to anomaly recognition indicate that this approach has the potential to classify observations from the range of population overlap where adjacent component distributions overlap significantly.

5.5.1 Parameter Estimates of Background Models

Results for the truncation corrected and uncorrected statistical sample parameters for the four data set structures indicate that truncation significantly affects the values,

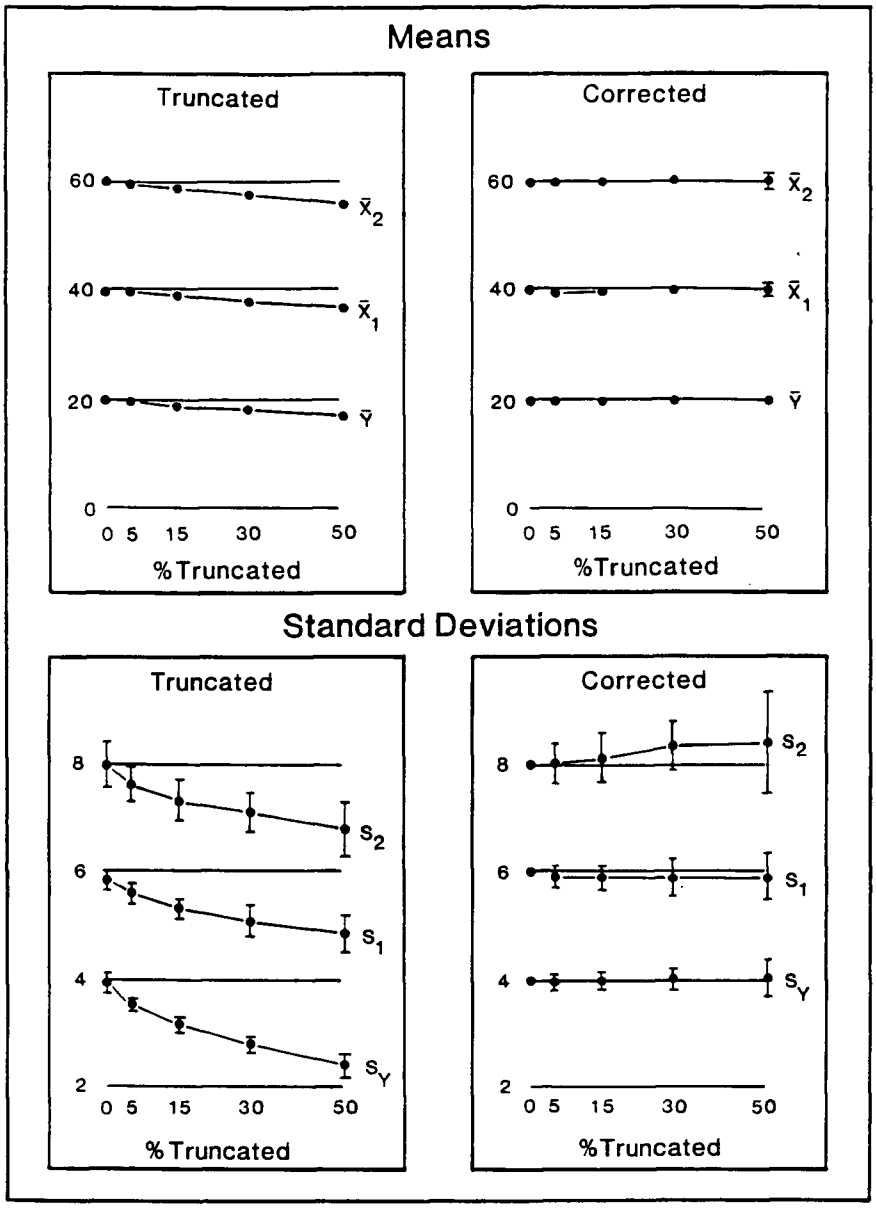


Figure 5.18: Plot Comparing the Population, Statistical Sample, Truncated Statistical Sample and Truncation Corrected Statistical Sample Parameter Estimates of the Means and Standard Deviations for Multivariate Data Set Structure # 17

Values in this figure are the means and (± 1) standard deviations of parameters for the 10 realizations used for each case. The population estimates are presented in the '0 % truncation / truncation corrected' location on the graph while the un-truncated statistical sample estimates are presented in the '0 % truncation / truncated' location on the graph.

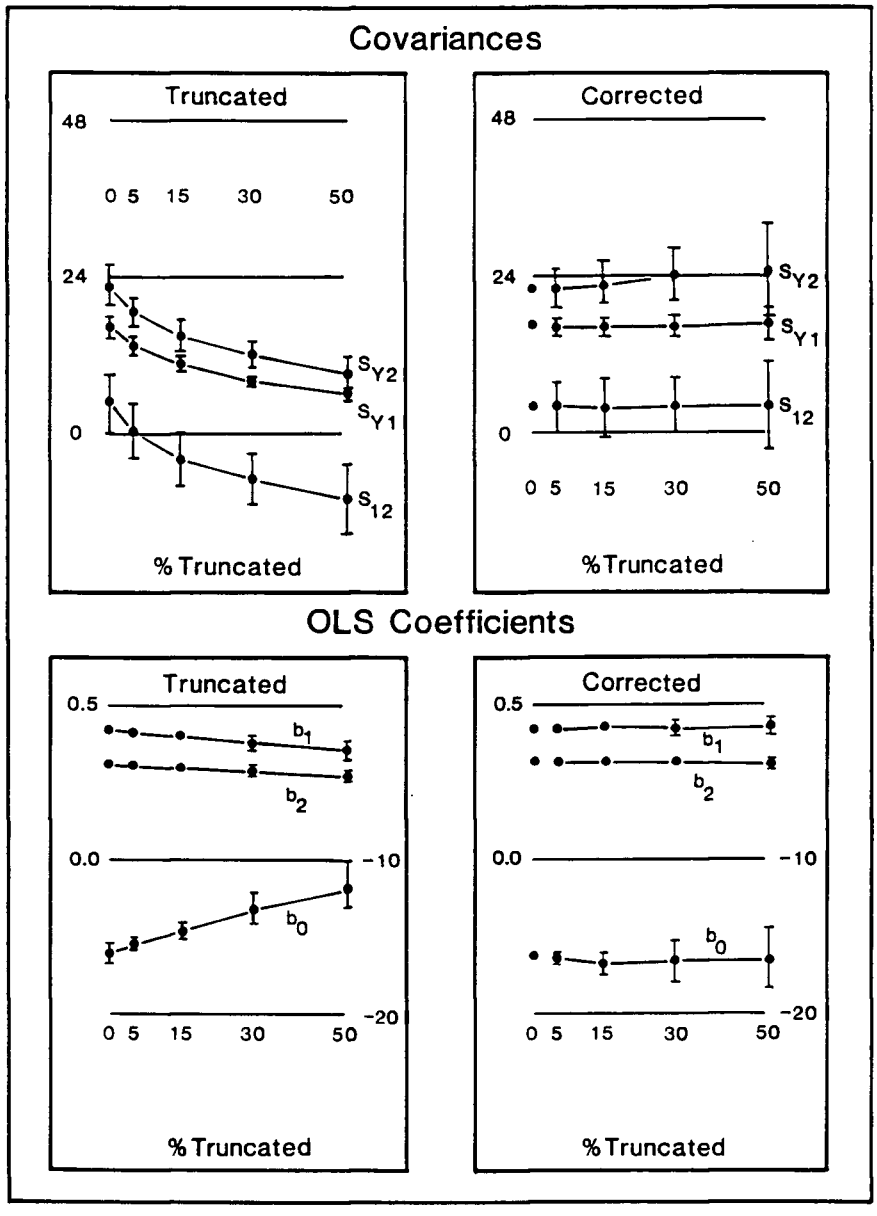


Figure 5.19: Plot Comparing the Population, Statistical Sample, Truncated Statistical Sample and Truncation Corrected Statistical Sample Parameter Estimates of the Covariances and *OLS* Coefficients for Multivariate Data Set Structure # 17

Values in this figure are the means and (± 1) standard deviations of parameters for the 10 realizations used for each case. The population estimates are presented in the '0 % truncation / truncation corrected' location on the graph while the un-truncated statistical sample estimates are presented in the '0 % truncation / truncated' location on the graph.

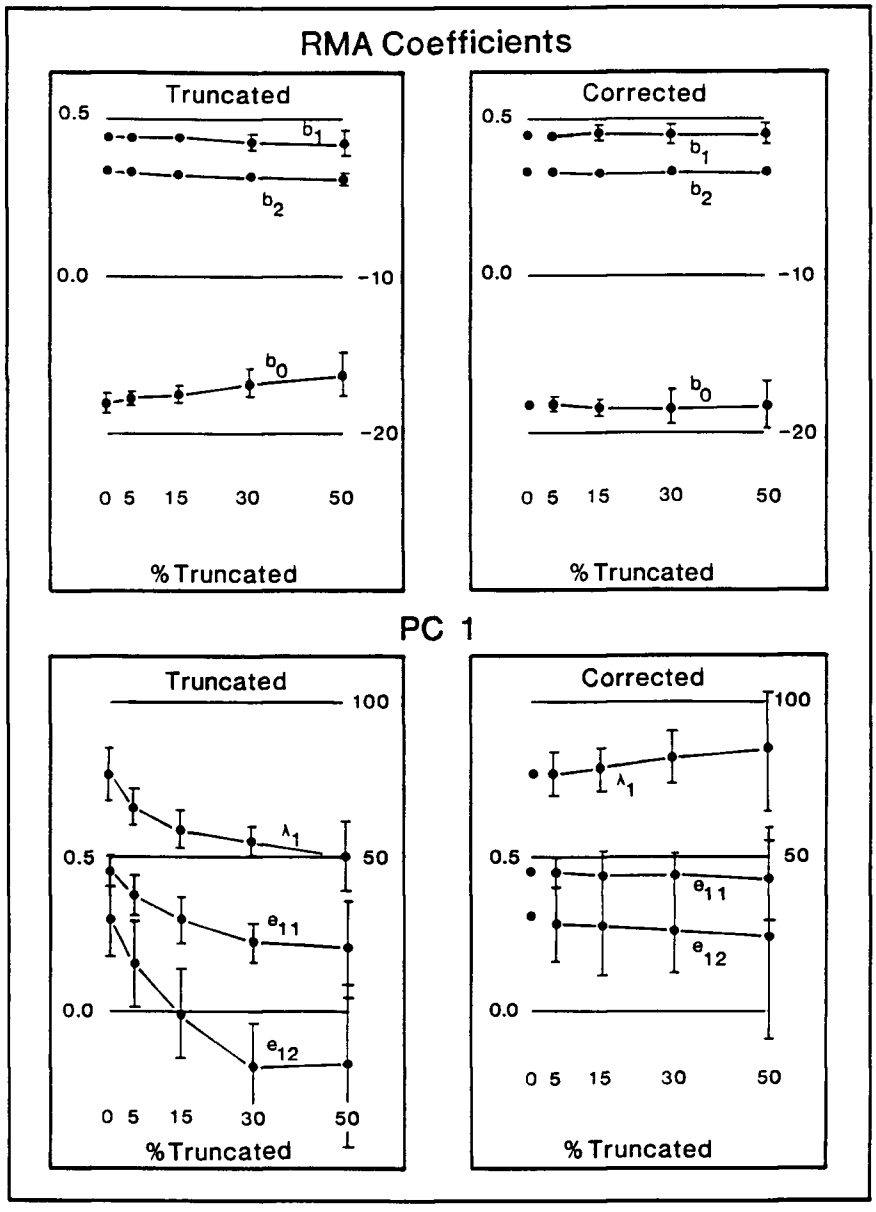


Figure 5.20: Plot Comparing the Population, Statistical Sample, Truncated Statistical Sample and Truncation Corrected Statistical Sample Parameter Estimates of the *RMA* and *PC* #1 Coefficients for Multivariate Data Set Structure # 17

Values in this figure are the means and (± 1) standard deviations of parameters for the 10 realizations used for each case. The population estimates are presented in the '0 % truncation / truncation corrected' location on the graph while the un-truncated statistical sample estimates are presented in the '0 % truncation / truncated' location on the graph.

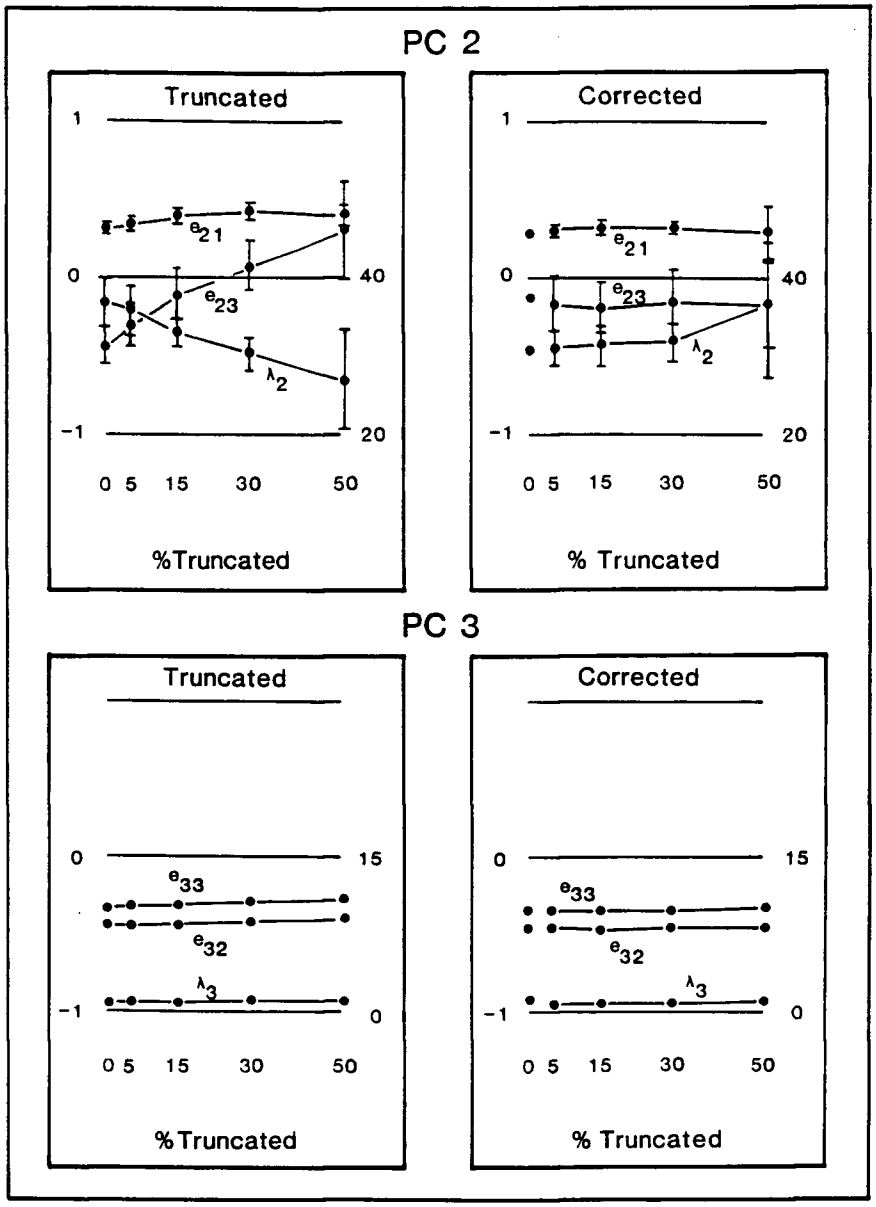


Figure 5.21: Plot Comparing the Population, Statistical Sample, Truncated Statistical Sample and Truncation Corrected Statistical Sample Parameter Estimates of the *PC* # 2 and # 3 Coefficients for Multivariate Data Set Structure # 17

Values in this figure are the means and (± 1) standard deviations of parameters for the 10 realizations used for each case. The population estimates are presented in the '0 % truncation / truncation corrected' location on the graph while the un-truncated statistical sample estimates are presented in the '0 % truncation / truncated' location on the graph.

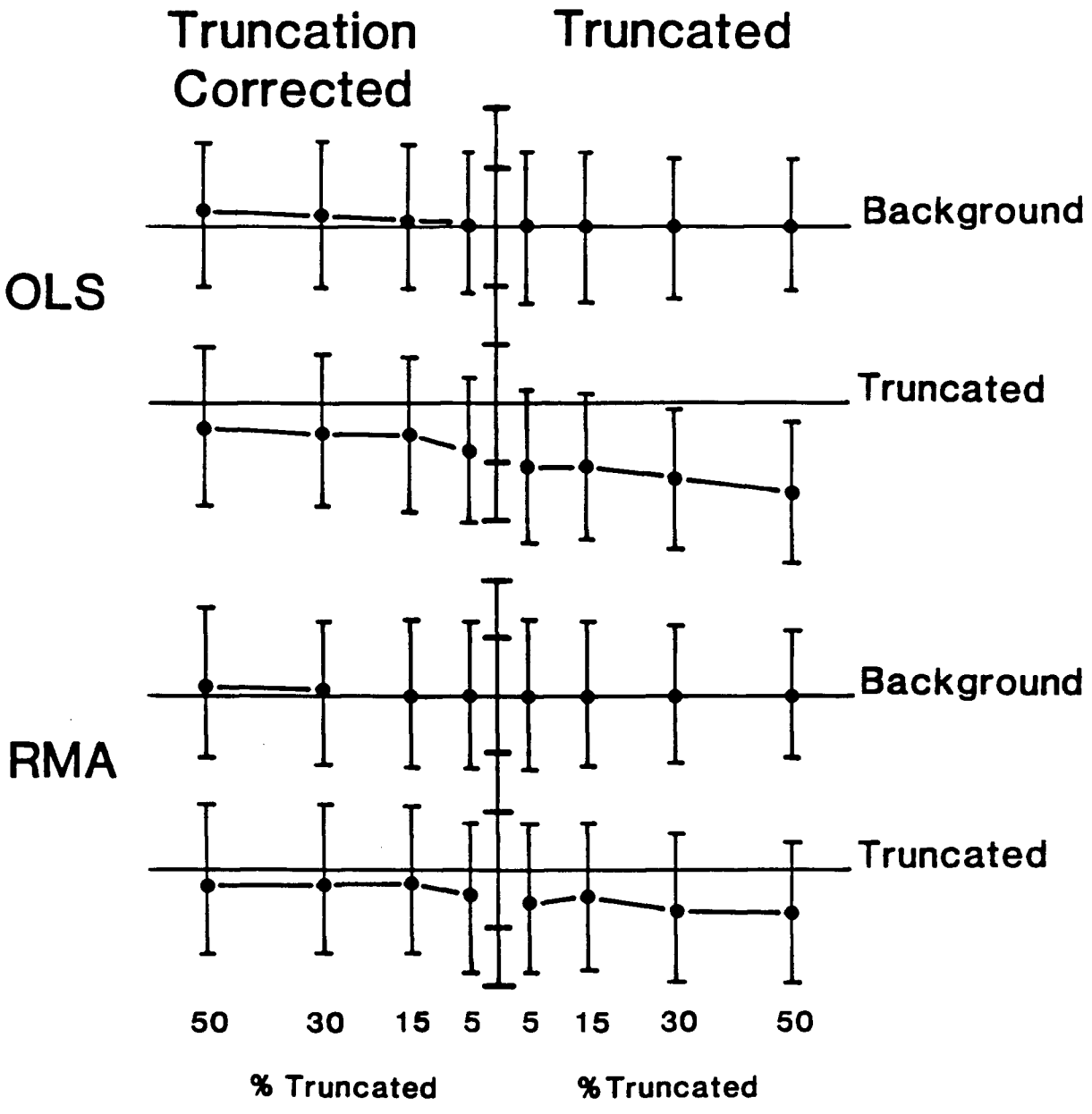


Figure 5.22: Plot of the Means and Standard Deviations of *OLS* and *RMA* Regression Residuals for Data Set Structure # 17

Tick marks on the central vertical axis represent a 1 unit value. Dots represent the mean of 10 values while the vertical bars represent $\pm$ 1 standard deviation.

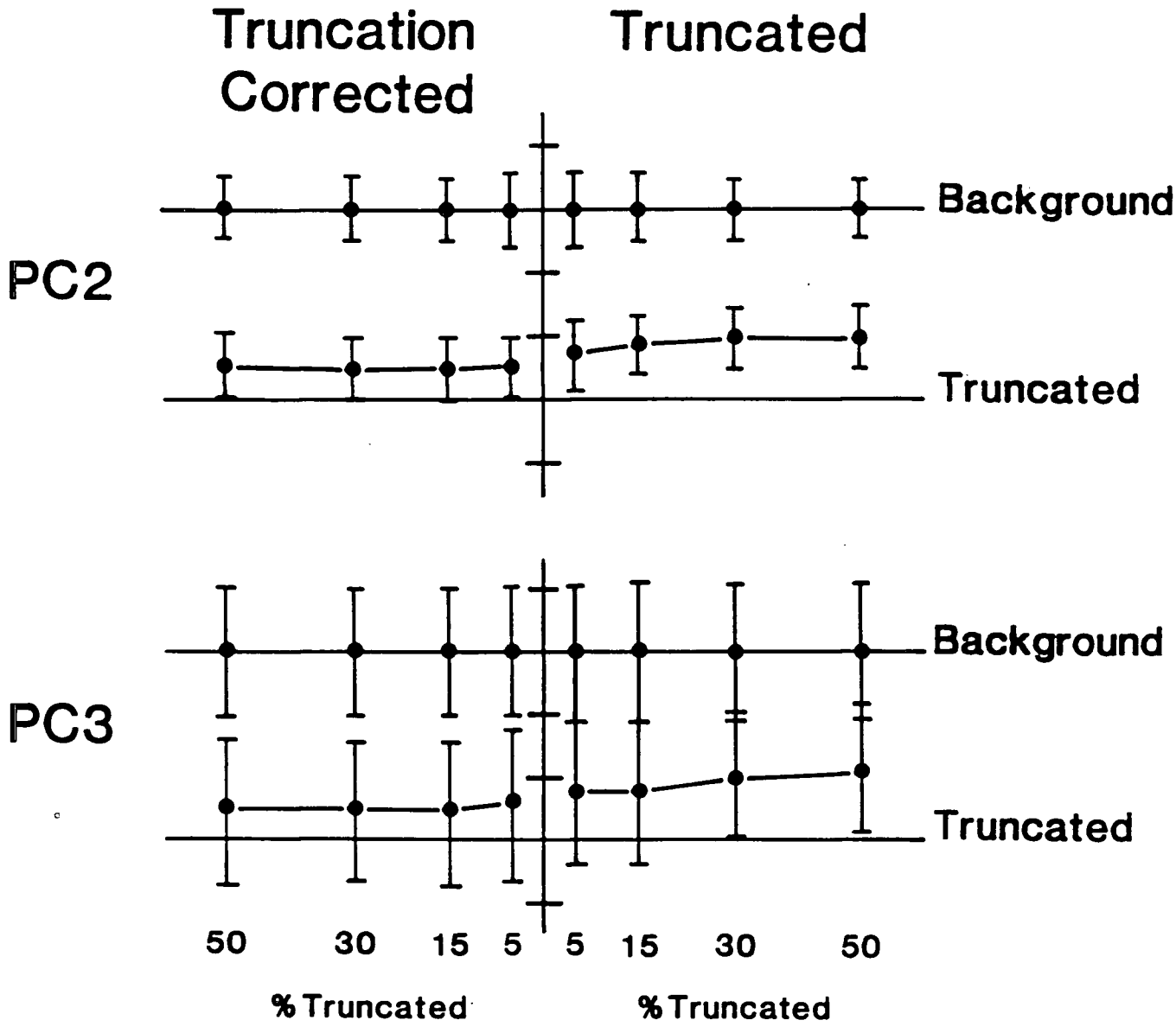


Figure 5.23: Plot of the Means and Standard Deviations of *PC* # 2 and # 3 Scores for Data Set Structure # 17

Tick marks on the central vertical axis represent a 10 unit value for *PC* # 2 and a 1 unit value for *PC* # 3. Dots represent the mean of 10 values while the vertical bars represent ± 1 standard deviation.

and that by correcting for truncation, the resulting estimates approximately reproduce the un-truncated parameters. With truncation, the descriptive statistics (means, standard deviations, covariances and correlations) of Y, X_1 and X_2 all decrease significantly; however, truncation corrected estimates of these parameters approximate the un-truncated statistical sample values without any significant bias (Tables F.58, F.60, F.59 and F.61; Figures 5.18 and 5.19).

The parameter estimates (b_0, b_1, b_2, R^2 and $\overline{\sum(\hat{y}_i - y_i)^2}$) of the background regression models (*OLS* and *RMA*) become biased with increasing amounts of truncation (the b_1, b_2, R^2 and $\overline{\sum(\hat{y}_i - y_i)^2}$ estimates decrease and the b_0 estimates increase). However, parameter estimates calculated using the truncation corrected descriptive statistics, above, adequately approximate the true population values (Tables F.62, F.64, F.63 and F.65; Figures 5.19 and 5.20).

Data set truncation also affects the calculated eigenvector and eigenvalue coefficients for the *PC* background model. The scaled eigenvector coefficients (those not scaled to 1) decrease or increase from an initial un-truncated statistical sample estimate, depending on the initial correlation structure. In all cases, both coefficients for eigenvector # 1 decrease with truncation. Conversely, the two coefficients for eigenvector # 2 mostly move in opposite directions whereas those for eigenvector # 3 generally increase with truncation. Increasing amounts of data set truncation also causes the calculated eigenvalues to decrease from initial un-truncated statistical sample estimates. Truncation correction of the descriptive statistics (covariances) used to calculate these eigenvectors also removes the bias induced through data set truncation (Tables F.66, F.68, F.67 and F.69; Figures 5.20 and 5.21).

Although bias in the background parameter estimates is small with small amounts of data set truncation (5 or 15 %), it becomes substantial if the amount of truncation is greater than 30 % of the data set. Unfortunately, if the thresholds of a mixture of

two normal distributions with a common standard deviation, equal component proportions and $\Delta = 2$ are chosen at the mean plus or minus two standard deviations of the corresponding component distributions, 32 % truncation is applied to each of the component distributions. Thus, since geochemical data bases comprised of mixtures of normal distributions with $\Delta \approx 2$ are not uncommon (e.g. Stanley and Sinclair 1988), significant bias in the truncated background model parameter estimates can be expected in geochemical applications.

As a result, the truncation correction described in Appendix E should be implemented to avoid background model parameter bias. Background models calculated using the truncation corrected parameters should produce residuals and scores which are similar to the residuals and scores calculated on the un-truncated data set (Figure 5.24).

5.5.2 Residual and Score Comparison

Although the above analysis of the parameters of the background model indicated that background corrections are required to produce unbiased estimates of the true parameter values with truncated data sets, this does not guarantee that the resulting residuals or scores will be unbiased. Furthermore, residuals or scores calculated from the background data, although they may exhibit predictable expected values and variances, will not necessarily be distributed similarly to the residuals or scores calculated from the truncated data.

Specifically, for all background functions (*OLS*, *RMA* regressions and the *PC*) calculated using the truncated data, the average mean values of the associated residuals and scores are zero. However, where the truncation corrected parameters are used to calculate these background residuals and scores, the *PC* scores have a mean value of zero, but the residuals from both regression functions have a slightly positive mean

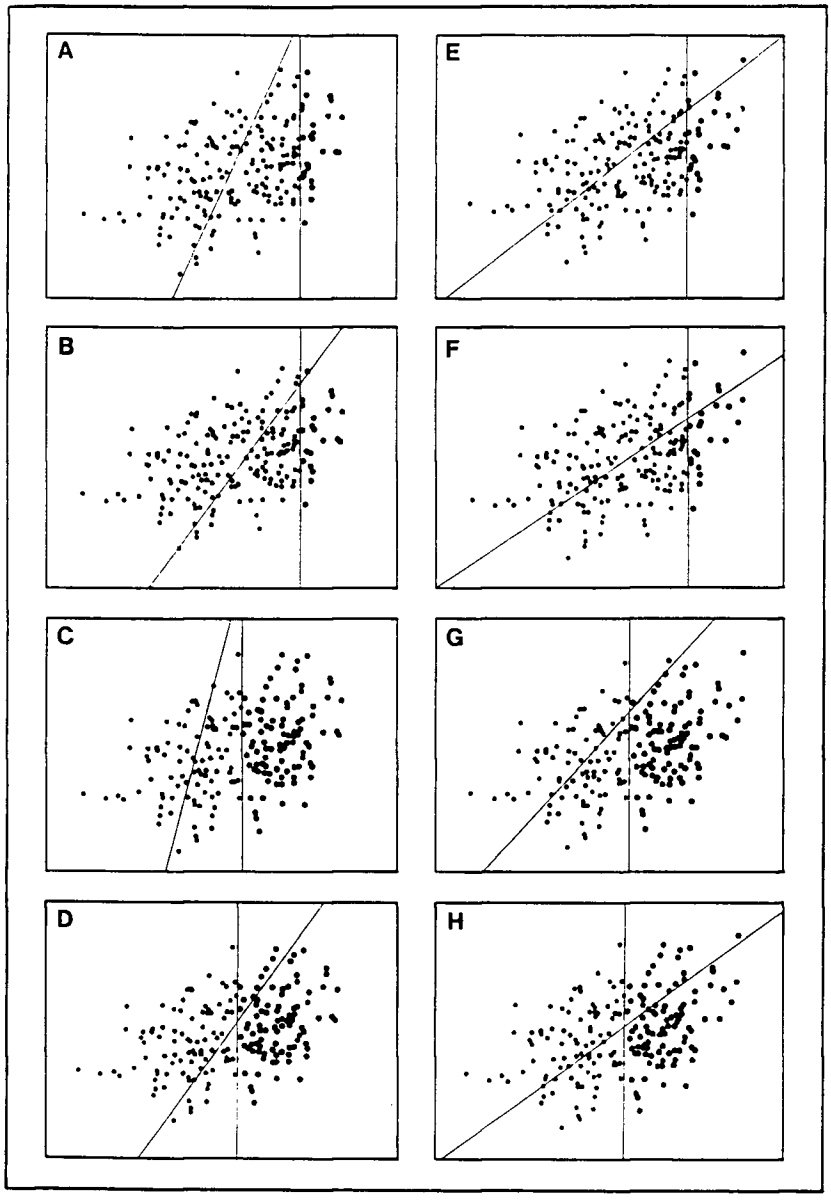


Figure 5.24: Comparison of Residuals of *OLS* and *RMA* Regression Background Models With and Without Truncation Correction for Realizations from Data Set Structures # 17 and # 23

All scatterplots have an abcissa of the observed Y value and an ordinate of the predicted Y value. Scatterplots on the left predict the Y value with the *OLS* regression function while those on the right predict the Y value with the *RMA* regression function. Scatterplots A, C, E and G have the predicted values calculated using the truncated data while scatterplots B, D, F and H have the predicted values calculated using the truncation corrected parameters. Scatterplots A, B, E and F are of a realization from data set structure # 17 while those of scatterplots C, D, G and H are of a realization from data set structure # 23.

value.

Similarly, the standard deviations of the residuals and scores calculated from the background model parameters have expected values which are predictable. The average standard deviations of the *PC* scores are equivalent to the average square roots of the associated eigenvalues; the average standard deviations of the residuals are equivalent to the square roots of the associated average sum of squared residuals. However, residuals calculated from the truncation corrected parameters using the regression functions are less than the associated average sum of squared residuals for the *OLS* regression function and greater than the associated sum of squared residuals for the *RMA* regression function. Scores for *PC* # 2 and # 3 calculated using the truncation corrected parameters have average standard deviations which are generally less than the average square root of the associated eigenvalues.

Finally, the background residuals and scores are not normally distributed. With truncation, the *OLS* regression residuals exhibit extremely positive average skewness whereas *RMA* regression residuals exhibit only marginally positive average skewness. Truncation correction appears to reduce the amount of skewness for both regression techniques; however, the skewnesses observed for the *OLS* truncation corrected regression residuals are still significantly positive. Average background scores calculated for *PC* # 2 and # 3 generally become negatively skewed with truncation, but this skewness appears to be removed if the truncation corrected *PC* parameters are used.

Identification of the observations which occur within the range of population overlap but which were derived from the background population will clearly have to be based on the residuals or scores of these observations and their distribution relative to the distribution of the residuals for the background observations (Figure 5.24). In all data set structures examined, average regression residuals for the truncated data are negative (residuals calculated using the *OLS* function are all more negative than those

calculated using the *RMA* function), whereas average *PC* scores for the truncated data are positive. These trends may reverse where there are negative correlations among pairs of variables in a data set. In addition, data set structure affects the magnitudes of these discrepancies from zero. Where $|R|$ (or equivalently, the multiple correlation coefficient or the percent variance explained by the first *PC*) is large, the discrepancies are small, and *vice versa*.

Similarly, average standard deviations of the residuals and scores for the truncated data are smaller than for the average background data for small amounts of truncation, becoming approximately equal to the average background values where truncation reaches 50 %. Likewise, average skewnesses of the residuals and scores for the truncated data tend to be more extreme with small amounts of truncation, becoming approximately equal where truncation reaches 50 %. The magnitudes of both the standard deviations and skewnesses of residuals and scores also vary inversely with $|R|$, the multiple correlation coefficient or the percent variance explained by the first *PC*.

5.6 Conclusions

OLS regression (despite not being consistent with a distribution model for geochemical data with error in every variable), *RMA* regression and *PC* can all be used as background functions to explain the variation among data from the background distribution. Residuals and scores calculated from these functions for the background data and (truncated) data from the range of population overlap are generally stable and predictable. However, *RMA* regression residuals are less biased away from zero and less skewed than their corresponding *OLS* regression residuals.

Truncation correction of the means, standard deviations and correlations of the truncated distribution can be used to calculate estimates of the background functions

for the un-truncated distribution. These truncation corrected background model parameters can then be used to determine truncation corrected residuals and scores which are even more stable and predictable, and less biased, than those calculated from the un-corrected parameters.

Although inclusion of data from a second distribution was not simulated, the stability and predictability of the residuals and scores calculated using the truncation corrected background model parameters should allow discrimination of data from other distributions from that of the background distribution which occurs in the range of population overlap, provided that the inclusion of these polluting observations has only a negligible effect on the background model parameter estimates.

Techniques for classifying the calculated residuals and scores include the very same techniques described in Chapter 1 and tested in Chapter 4 for selecting thresholds (histograms, probability plots and the gap statistic).

Chapter 6

Conclusions and Recommendations

"Enough research will tend to support your theory."

Murphy's Law of Research (1979)

"A conclusion is the place where you got tired of thinking."

Matz's Maxim (1980)

6.1 Conclusions

Several important conclusions from this study which relate to the approaches taken by geoscientists in the evaluation of geochemical data result from the *monte carlo* simulations of thresholds selection and classification procedures presented in Chapters 3, 4 and 5. Major conclusions resulting from the comparison of likelihood functions in the decomposition of mixtures of normal distributions and the calculation of optimal parameter estimates (Chapter 3) include :

- *ML* optimization can be used successfully to decompose a mixture of normal distributions, producing estimates of the means, standard deviations and component porportions of the component distributions, thus mixtures of geochemical populations in a data set can be described and discriminated;

- both *RDML* and *CIDML* approaches may be used to obtain reasonable estimates of these parameters; however, biased estimates can result using the *CIDML* approach where a small number of class intervals are used (corresponding to a poor

estimate of the frequency of the c3lass interval by the trapezoid approximation); and

- both *RDML* and *CIDML* approaches produce poor (and commonly variable) estimates of the population parameters if a small number of observations are used (< 50 observations per distribution) or if the component distributions overlap substantially ($\Delta \leq 2$).

This analysis demonstrates that *ML* estimation of the parameters of each distribution in a mixture of normal distributions can be done rapidly and accurately using the *CIDML* function. Thus, optimal parameter estimation applied to large data sets common in geochemical applications is not only possible, but can be accomplished easily and rapidly with inexpensive personal computers.

The comparison of classification results between the probability plot and gap statistic approaches of Chapter 4 indicates that :

- the probability plot approach can be used successfully to classify observations from the component distributions of a mixture of normal distributions;
- the performance of the probability plot approach is the 'best' of the techniques considered using a single threshold;
- classification with the probability plot approach may be inaccurate where a small number of observations is used (< 50 observations per distribution) or if the component distributions overlap considerably ($\Delta \leq 2$);
- the gap statistic can be used successfully to classify observations from component distributions of a mixture of normal distributions; however, the performance of the gap statistic, where the data distribution is substantially skewed, is inferior to that of the probability plot approach; and

- the gap statistic, at the very least, may be used as a test of discordancy, determining the plausibility of the observed distribution being derived from a single 3-parameter log-normal distribution.

The probability plot technique, thus, is recommended for use by geoscientists to decompose mixtures of normal distributions, select thresholds and classify geochemical samples. Software to accomplish this task is described in Chapter 2.

Finally, the simulation of the *BCA* to anomaly recognition using both regression and *PC* functions as background models (Chapter 5) indicates that :

- either *OLS* regression, *RMA* regression or *PC* can be used to predict the background model coefficients required by the *BCA*;

- these coefficients may then be used to calculate residuals or scores for background data and data from the range of population overlap;

- since the projected residuals and scores for background data from the range of population overlap are neither identically distributed nor generally known, all truncated residuals and scores should be compared to the background residuals and scores to ascertain which observations from the range of population overlap are truely anomalous;

- the *RMA* regression truncated residuals are less biased from 0 and have smaller skewness than the corresponding *OLS* regression truncated residuals; thus, the *RMA* reression function is prefered because the resulting residuals are more predictable; and

- the truncation correction procedures applied to truncated data sets appear to reduce the disparity between the background and truncated residuals and scores

in the *BCA*, allowing for a better chance of identifying observations which are not part of the background distribution.

Thus, the *BCA* can be used successfully to decrease the amount of misclassification of data from the range of population overlap where a large amount of overlap exists between the component distributions.

6.2 Summary

In summary, this study has presented a general systematic methodology for the classification of geochemical data. A mixture of distributions model is 'fit' to the geochemical data. If a small amount of population overlap exists, a single threshold can be selected using the parameters of this distribution model to classify the data. If substantial overlap exists, two thresholds can be selected to define background ranges and a range of population overlap. A background variation model can then be used to describe the background data and discriminate the background observations in the range of population overlap from the anomalous observations.

This methodology has been tested using stochastic data sets and *monte carlo* simulations to evaluate whether the techniques used are applicable to the variety of data set strctures which may be encountered in applied geochemistry. Results demonstrate that the techniques advocated for estimating the parameters of a mixture of distributions, the selection of thresholds, and the discrimination of observations from ranges of population overlap all perform adequately and allow classification of geochemical data which are distributed as a mixture of distributions.

6.3 Recommendations for Further Work

While a thorough comparison of the classification efficiency and performance of probability plots and the gap statistic has been presented above, the same cannot be said of the simulation of the *BCA* to anomaly recognition. Specifically, no attempt to simulate the effects of the presence of an anomalous population was undertaken. This is because the quality of classification using the *BCA* to anomaly recognition is not solely dependent upon the behavior of observations from the anomalous distributions, but rather is also dependent upon how well the observations from the truncated range of the background population can be predicted. If these are predicted well, provided there is enough difference between the anomalous and background populations, the anomalous observations should be able to be distinguished.

However, the effect of having a small number of anomalous observations in the background data range used for determining the parameters of the background model has not been assessed. These outliers, while few in number, can lead to inaccurate background model parameter estimates. Obviously, simulation of how a small number of these outlying observations affect the background model parameters and resulting residuals and scores should be undertaken. In the meantime, conservative thresholds should be selected which lead to a number of contaminating observations included in the background data range which is as small as possible. At the same time, the locations of these thresholds should be chosen such that the the amount of truncation is not too large and does not lead to inaccurate background model parameter estimates when the distribution parameters are corrected for truncation.

Obvious methods for reducing the effects of these polluting observations involve assigning weights to all observations which are inversely proportional to the residuals or scores in an iterative fashion. Unfortunately, this may only be done using the truncated

data set procedures because these procedures are the only ones which calculate the background model parameters using the actual data.

Since this study represents only a small proportion of the analysis required to ascertain the quality of all classification techniques which may be used in geochemistry, expansion of this study could lead to an evaluation of other 'true' multivariate approaches. Topics which may represent good subjects for further study in the future include multivariate *ML* parameter estimation of mixtures of multivariate normal distributions. This approach involves little subjective criteria, is completely consistent with the proposed paradigm for the distribution of geochemical data and represents the optimal approach to multivariate classification. However, the time-consuming numerical calculations required to determine the maximum likelihood parameter estimates of the distribution model :

$$f(\vec{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\tilde{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{x}_i-\vec{\mu}_x)^T\tilde{\Sigma}^{-1}(\vec{x}_i-\vec{\mu}_x))}, \tag{6.124}$$

(calculation of the determinant and inverse of $\tilde{\Sigma}$) limit the use of this technique to those with inexpensive main-frame computer accounts. Furthermore, the number of parameters which must be approximated becomes very large when a large number of variables are considered :

$$s = v\left[2p + \binom{p}{2} + 1\right] - 1, \tag{6.125}$$

where s is the number of parameters, v is the number of normal distributions and p is the number of variables. Thus, 19 independent parameters must be approximated for a simple mixture of two trivariate normal distributions. As a result, multivariate *ML* parameter estimation and classification is an approach which may be feasibly studied in the future, when high speed computing power becomes even cheaper.

A second topic not evaluated in this study represents not so much a new approach as a new application of an existing approach. Specifically, this study has addressed

positive mixtures of normal distibutions of the form :

$$f(x) = \sum_{k=1}^{v} \frac{\varpi_k}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}, \tag{6.126}$$

where $\sigma_k > 0$, $\sum_{k=1}^{v} \varpi_k = 1$ and $0 \leq \varpi_k \leq 1$. By relaxing one of these conditions, specifically $0 \leq \varpi_k \leq 1$, **negative** mixtures of normal distributions may be considered. These could be used to model geologic or geochemical processes such as sediment winnowing, where an initial distribution is acted upon by a process which removes specific components from that distribution through specific stochastic processes.

PDF's of the form :

$$f(x) = \frac{\varpi_1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} + \frac{\varpi_2}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2}, \tag{6.127}$$

where $\varpi_1 > 1$, $\varpi_2 < 0$ and $\varpi_1 + \varpi_2 = 1$, can be used to describe a frequency distribution which is the *difference* between two normal distributions. In order for this mathematical model to be consistent with reality, the parameters of this *PDF* must ensure that $f(x) \geq 0$, or more specifically (for a negative mixture of two normal distributions) :

$$|\varpi_1| \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} \geq |\varpi_2| \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2}. \tag{6.128}$$

This approach merely represents a reformulation of the constraints on the distribution model. Figure 6.25 depicts the expected probability graph of a negative mixture of normal distributions relative to additive mixtures of normal distributions described by Sinclair (1974, 1976). This negative mixture model could provide new insight into the study of winnowing and dissolution processes because it is completely consistent with the processes which lead to the observed distributions.

Both of these topics represent logical extentions of this study and may lead to substantial improvements in data classification and our understanding of geochemical processes.

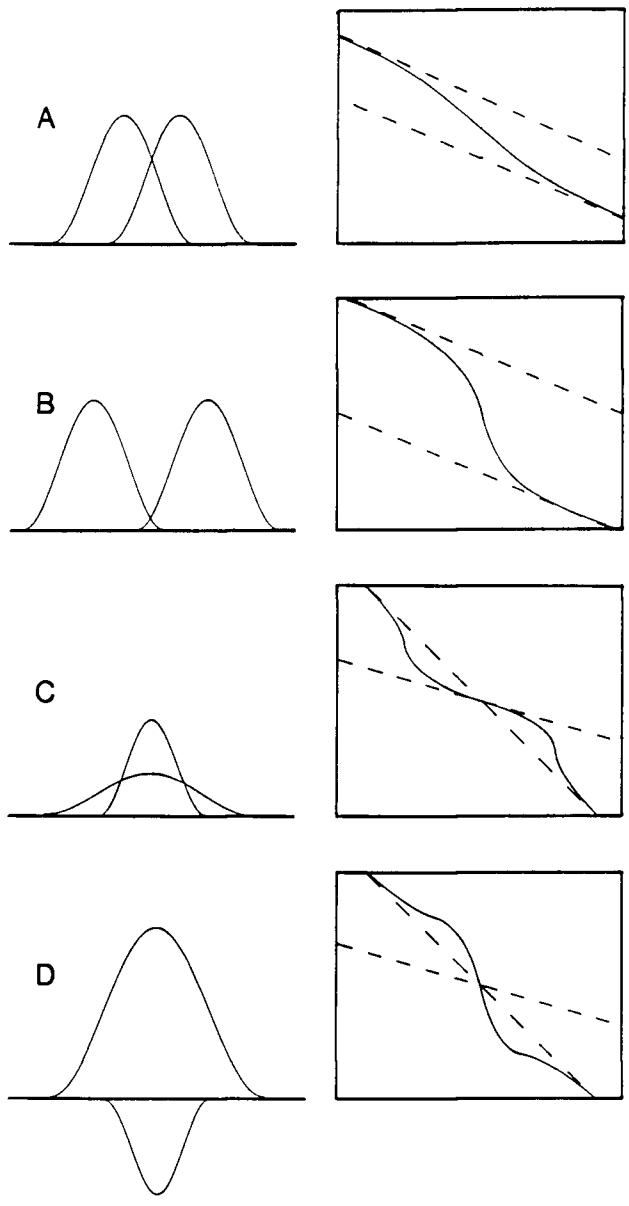


Figure 6.25: Probability Graph Comparison of Non-Overlapping, Overlapping, Intersecting and Negative Mixtures of Normal Distributions

(A) Probability graph and frequency distribution of Overlapping mixture of distributions.

(B) Probability graph and frequency distribution of Non-overlapping mixture of distributions.

(C) Probability graph and frequency distribution of Intersecting mixture of distributions.

(D) Probability graph and frequency distribution of Negative mixture of distributions.

The sigmoidal form for probability graphs A, B and D are all similar (flat, steep and then flat), but substantially different from C (steep, flat and then steep). Thus, it is easy to see how distributions produced by a negative mixture could easily by mistakenly modeled as mixtures of overlapping or non-overlapping distributions.

References

Ageno, M. and Frontali, C. (1963): *Analysis of Frequency Distribution Curves in Overlapping Gaussians.* Nature, Vol. 198, June 29, pp. 1294-1295.

Ahrens, J.H. and Deiter, U. (1973): *Non-Uniform Random Numbers.* Institut für Mathematik, Technische Hochschule, Graz, Austria, 136 p.

Ahrens, L.A. (1954): *The Lognormal Distribution of the Elements.* Geochimica et Cosmochimica Acta, Vol. 5, pp. 49-73.

Aitchison, J. and Brown, J.A.C. (1957): *The Lognormal Distribution with Special Reference to its Use in Economics.* Cambridge University Press, London, 176 p.

Alabert, F. (1987): *The Practice of Fast Conditional Simulations Throught the LU Decomposition of the Covariance Matrix.* Mathematical Geology, Vol. 19, No. 5, pp. 369-386.

Anderson, T.W. (1984): *Estimating Linear Statistical Relationships.* Annals of Statistics, Vol. 12, No. 1, pp. 1-45.

Barnett, V. and Lewis, T. (1978): *"Outliers in Statistical Data".* John Wiley and Sons, Toronto, 365 p.

Behboodian, J. (1970): *On a Mixture of Normal Distributions.* Biometrika, Vol. 57, pp. 215-216.

Behboodian, J. (1972): *Information Matrix for a Mixture of Two Normal Distributions.* Journal of Statistics and Computer Simulation, Vol. 1, pp. 295-314.

Beyer, W.J. - ed. (1982): *CRC Standard Mathematical Tables - 26th Edition.* CRC Press, Inc., Boca Raton, Fla.

Bhattacharya, C.G. (1967): *A Simple Method of Resolution of a Distribution into Gaussian Components.* Biometrics, Vol. 23, pp. 115-135.

Birch, M.W. (1964): *A Note on the Maximum Likelihood Estimation of a Linear Structural Relationship.* Journal of the American Statistical Association, Vol. 59, pp. 1175-1178.

Bölviken, B. (1971): *A Statistical Approach to the Problem of Interpretation in Geochemical Prospecting.* Canadian Institute of Mining and Metallurgy, Special Volume No. 11, pp. 564-567.

Box, G.E.P. and Muller, M.E. (1958): *A Note on the Generation of Random Normal Deviates.* Annals of Mathematical Statistics, Vol. 29, pp. 610-611.

Brazier, S., Sparks, R.S.J., Carey, S.N., Sigurdsson, H. and Westgate, J.A. (1983): *Bimodal Grain Size Distribution and Secondary Thickening in Air-Fall Ash Layers.* Nature, Vol. 301, Jan. 13, pp. 115-119.

Bridges, N.J. and McCammon, R.B. (1980): *DISCRIM: A Computer Program Using an Interactive Approach to Dissect a Mixture of Normal or Lognormal Distributions.* Computers and Geosciences, Vol. 6, No. 4, pp. 361-396.

Brooks, C., Hart, S.R. and Wendt, I. (1972): *Realistic Use of Two-Error Regression Treatments as Applied to Rubidium-Strontium Data.* Reviews in Geophysics and Space Physics, Vol. 10, pp. 551-577.

Brooks, C., Wendt, I. and Harre, W. (1968): *A Two-Error Regression Treatment and Its Application to Rb-Sr and Initial Sr^{87}/Sr^{86} Ratios of Younger Variscian Granitic Rocks From the Schwarzwald Massif, Southwest Germany.* Journal of Geophysical Research, Vol. 73, No. 18, pp. 6071-6084.

Burden, R.L. and Faires, J.D. (1985): *Numerical Analysis.* Prindle, Weber and Schmidt, Boston, pp. 349-352.

Caceci, M.S. and Cacheris, W.P. (1984): *Fitting Curves to Data: The SIMPLEX Algorithm is the Answer.* Byte Magazine, May 1984, pp. 340-362.

Cassie, R.M. (1954): *Some Uses of Probability Paper in the Analysis of Size Frequency Distributions.* Australian Journal of Marine and Freshwater Research, Vol. 5, pp. 513-522.

Charlier, C.V.L. and Wicksell, S.D. (1924): *On the Dissection of Frequency Functions.* Arkiv f. Matematik Astron. och Fysik., Bd. 18, No. 6.

Choi, K. (1969): *Estimators for the Parameters of a Finite Mixture of Distributions.* Annals of the Institute of Statistical Mathematics, Vol. 21, pp. 107-116.

Clark, I. (1977): *ROKE, A Computer Program for Nonlinear Least-Squares Decomposition of Mixtures of Distributions.* Computers and Geosciences, Vol. 3, pp. 245-256.

Clark, M.W. (1976): *Some Methods for Statistical Analysis of Multimodal Distributions and Their Application to Grain-Size Data.* Mathematical Geology, Vol. 8, No. 3, pp. 267-282.

Clark, M.W. (1977a): *GETHEN: A Computer Program for the Decomposition of Mixtures of Two Normal Distributions by the Method of Moments.* Computers and Geosciences, Vol. 3, pp. 257-267.

Clark, M.W. (1977b): *GETHEN, A Computer Program for the Decomposition of Mixtures of Two Normal Distributions by the Method of Moments.* Errata, Computers and Geosciences, Vol. 4, No. 4, pp. 373-374.

Clifton, H.E., Hunter, R.E., Swanson, F.J. and Phillips, R.L. (1969): *Sample Size and Meaningful Gold Analysis.* U.S. Geological Survey Professional Paper # 625-C, 17 p.

Cohen, A.C. Jr. (1950): *Estimating the Mean and Variance of Normal Populations from Singly Truncated and Doubly Truncated Samples.* Annals of Mathematical Statistics, Vol. 21, pp. 557-569.

Cohen, A.C. Jr. (1957): *On the Solution of Estimating Equations for Truncated and*

Censored Samples from Normal Populations. Biometrika, Vol. 44, pp. 225-236.

Cohen, A.C. Jr. (1959): *Simplified Estimators for the Normal Distribution When Samples Are Singly Censored or Truncated.* Technometrics, Vol. 1, No. 3, pp. 217-237.

Cohen, A.C. Jr. (1961): *Tables for Maximum Likelihood Estimates : Singly Truncated and Singly Censored Samples.* Technometrics, Vol. 3, No. 4, pp. 535-541.

Cohen, A.C. Jr. (1966): *Discussion of 'Estimation of Parameters for a Mixture of Normal Distributions' by Victor Hasselblad.* Technometrics, Vol. 8, No. 3, pp. 445-446.

Cohen, A.C. Jr. (1967): *Estimation in Mixtures of Two Normal Distributions.* Technometrics, Vol. 9, No. 1, pp. 15-28.

Court, A. (1949): *Separating Frequency Distributions into Two Normal Components.* Science, Vol. 110, Nov. 11, pp. 500-501.

Cox, D.R. and Hinkley, D.V. (1974): *Theoretical Statistics.* Chapman and Hall, London, 511 p.

David, H.A. (1981): *Order Statistics.* 2^{nd} Edition, John Wiley and Sons, New York, 360 p.

Davies, O.L. and Goldsmith, P.L. - eds. (1972): *Statistical Methods in Research and Production.* Oliver and Boyd, Edinburgh, 478 p.

Davis, M.W. (1987a): *Production of Conditional Simulations Via the LU Triangular Decomposition of the Covariance Matrix.* Mathematical Geology, Vol. 19, No. 2, pp. 91-98.

Davis, M.W. (1987b): *Generating Large Stochastic Simulations - The Matrix Polynomial Approximation Method.* Mathematical Geology, Vol. 19, No. 2, pp. 99-108.

Day, N.E. (1969): *Estimating the Components of a Mixture of Normal Distributions.* Biometrika, Vol. 59, No. 3, pp. 639-648.

Day, S.J., Broster, B.E. and Sinclair, A.J. (1987): *Sulphide Erratics Applied to Subglacial Exploration : St. Elias Mountains, British Columbia.* Canadian Journal of Earth Sciences, Vol. 24, pp. 723-730.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977): *Maximum Likelihood Estimation from Incomplete Data Via the EM Algorithm.* Journal of the Royal Statistical Society – B, Vol. 39, pp. 1-38.

Dick, N.P. and Bowden, D.C. (1973): *Maximum Likelihood Estimation for Mixtures of Two Normal Distributions.* Biometrics, Vol 29, pp. 781-790.

Dixon, W.J. (1950): *Analysis of Extreme Values.* Annals of Mathematical Statistics, Vol. 21, pp. 488-506.

Dixon, W.J. (1951): *Ratios Involving Extreme Values.* Annals of Mathematical Statistics, Vol. 22, pp. 68-78.

Dixon, W.J. (1953): *Processing Dat for Outliers.* Biometrika, Vol. 9, pp. 74-89.

Deiter, U. and Ahrens, J.H. (1973): *Uniform Random Numbers.* Institut für Mathematik, Technische Hochschule, Graz, Austria, 188 p.

Dolby, G.R. (1976a): *The Ultrastructural Relation : A Synthesis of the Functional and Structural Relations.* Biometrika, Vol. 63, pp. 39-50.

Dolby, G.R. (1976b): *A Note on the Linear Structural Relation When Both Residual Variances Are Known.* Journal of the American Statistical Association, Vol. 71, pp. 352-353.

Draper, N.R. and Smith, H. (1981): *Applied Regression Analysis.* John Wiley and Sons, New York, 709 p.

Eisenberger, I. (1964): *Genesis of Bimodal Distributions.* Technometrics, Vol. 6, pp. 357-363.

Everitt, B.S. and Hand, D.J. (1981): *Finite Mixture Distributions.* Chapman and Hall, New York, 143 p.

Folk, R.L. (1971): *Longitudinal Dunes of the Northwestern Edge of the Simpson Desert, Northern Territory, Australia, 1. Geomorphology and Grain Size Distribution.* Sedimentology, Vol. 16, pp. 5-54.

Fowlkes, E.B. (1979): *Some Methods for Studying the Mixture of Two Normal (Log-normal) Distributions.* Journal of the American Statistical Association, Vol. 74, No. 367, pp. 561-575.

Fryer, J.G. and Robertson, C.A. (1972): *A Comparison of Some Methods for Estimating Mixed Normal Distributions.* Biometrika, Vol. 59, No. 3, pp. 639-648.

Full, W.E., Ehrlich, R. and Kennedy, S.K. (1984): *Optimal Configuration and Information Content of Sets of Frequency Distributions.* Journal of Sedimentary Petrology, Vol. 54, No. 1, pp. 117-126.

Fuller, W.A. (1987): *Measurement Error Models.* John Wiley and Sons, New York, 440 p.

Garrett, R.G. (1984): *Workshop 5 : Thresholds and Anomaly Recognition.* Journal of Geochemical Exploration, Vol. 21, pp. 137-142.

Garrett, R.G. (1987): *The Chi-Square Plot : A Tool for Multivariate Outlier Recognition.* Programme and Abstracts, 12^{th} International Geochemical Exploration Symposium, Orléans, France, May, pp. 129-130.

Ghosh, A. and Pinnaduwa, H.S.W.K. (1987): *A FORTRAN Program for Generation of Multivariate Normally Distributed Random Variables.* Computers and Geosciences, Vol. 13, No. 3, pp. 221-234.

Ghose, B.K. (1970): *Statistical Analysis of Mixed Fossil Populations.* Mathematical Geology, Vol. 2, No. 3, pp. 265-276.

Gregor, J. (1969): *An Algorithm for the Decomposition of a Distribution into Gaussian Components.* Biometrics, Vol. 56, pp. 79-93.

Grubbs, F.E. (1950): *Sample Criteria for Testing Outlying Observations.* Annals of

Mathematical Statistics, Vol. 21, pp. 27-58.

Grubbs, F.E. (1969): *Procedures for Detecting Outlying Observations in Samples.* Technometrics, Vol. 11, pp. 1-21.

Grubbs, F.E. and Beck, G. (1972): *Extensions of Sample Size and Percentage Points for Significance Tests of Outlying Observations.* Technometrics, Vol. 14, No. 4, pp. 847-854.

Gy, P.M. (1982): *Sampling of Particulate Materials – Theory and Practice.* Elsevier Scientific Publishing Co., Amsterdam, 431 p.

Hald, A. (1949): *Maximum Likelihood Estimation of the Parameters of a Normal Distribution which is Truncated at a Known Point.* Skandinavisk Aktuarietidskrift, Vol. 32, pp. 119-134.

Hald, A. (1952): *Statistical Theory with Engineering Applications.* John Wiley and Sons, New York, 783 p.

Harding, J.P. (1949): *The Use of Probability Paper for the Graphical Analysis of Polymodal Frequency Distributions.* Journal of the Marine Biological Association, Vol. 28, pp. 141-153.

Hasselblad, V. (1966): *Estimation of Parameters for a Mixture of Normal Distributions.* Technometrics, Vol. 8, No. 3, pp. 431-444.

Hoffman, S.J. (1986): *Writing Geochemical Reports.* Association of Exploration Geochemists, Special Volume No. 12, 29 p.

Hosmer, D.W. (1973): *A Comparison of Iterative Maximum Likelihood Estimates of the Parameters of a Mixture of Two Normal Distributions Under Three Different Types of Sample.* Biometrics, Vol. 29, pp. 761-770.

Hathaway, R.J. (1985): *A Constrained Formulation of Maximum Likelihood Estimation for Normal Mixture Distributions.* The Annals of Statistics, Vol. 13, No. 2, pp. 795-800.

Ingamells, C.O. (1974): *Control of Geochemical Error Through Sampling and Subsampling Diagrams.* Geochimica et Cosmochimica Acta, Vol. 38, pp. 1225-1237.

Ingamells, C.O. (1981): *Evaluation of Skewed Exploration Data – The Nugget Effect.* Geochimica et Cosmochimica Acta, Vol. 45, pp. 1209-1216.

Johnson, N.L. and Kotz, S. (1970): *Continuous Univariate Distributions - 1*: John Wiley and Sons, New York, 300 p.

Johnson, R.A. and Wichern, D.W. (1982): *Applied Multivariate Statistical Analysis.* Prentice Hall, Englewood, New Jersey, 594 p.

Jones, T.A. (1968): *Statistical Analysis of Orientation Data.* Journal of Sedimentary Petrology, Vol. 38, No. 1, pp. 61-67.

Jones, T.A. (1972): *Multiple Regression with Correlated Independent Variables. I. Maximum Likelihood and Least-Squares Estimation.* Mathematical Geology, Vol. 4, No. 3, pp. 203-218.

Jones, T.A. (1979): *Fitting Straight Lines When Both Variables Are Subject to Error. 1. Maximum Likelihood and Least-Squares Estimation.* Mathematical Geology, Vol. 11, No. 1, pp. 1-25.

Jones, T.A. and James, W.R. (1969): *Analysis of Bimodal Orientation Data.* Mathematical Geology, Vol. 1, No. 2, pp. 129-135.

Kendall, M.G. and Stuart, A. (1961): *"The Advanced Theory of Statistics".* Hafner Publishing Co., New York, pp. 375-418.

Kermack, K.A. and Haldane, J.B.S. (1950): *Organic Correlation and Allometry.* Biometrika, Vol. 37, pp. 30-41.

Knuth, D.E. (1981): *The Art of Computer Programming : Semi-Numerical Algorithms.* Vol. 2, 2^{nd} Edition, Addison-Wesley, Reading, Mass., 688 p.

Kullback, S. and Leibler, R.A. (1951): *On Information and Sufficiency.* Annals of Mathematical Statistics, Vol. 22, pp. 79-86.

LePeltier, C. (1969): *A Simplified Statistical Treatment of Geochemical Data by Graphical Representation.* Economic Geology, Vol. 64, pp. 538-550.

Lindley, D.V. (1947): *Regression Lines and the Linear Functional Relationship.* Journal of the Royal Statistical Society - Supplement, Vol. 9, pp. 218-244.

Lindquist, L., Lundholm, I., Nisca, D. and Esbensen, K. (1987): *Multivariate Geochemical Modelling and Integration with Petrophysical Data.* Journal of Geochemical Exploration, Vol. 29, pp. 279-294.

Loius, T. (1982): *Finding the Observed Information Matrix When Using the EM Algorithm.* Journal of the Royal Statistical Society – B, Vol. 44, pp. 226-233.

MacDonald, P.D.M. (1975): *Estimation of Finite Mixture Distributions.* in "Applied Statistics", Grupta, R.P., ed., North Holland Publishing Co., Amsterdam, pp. 231-245.

MacDonald, P.D.M. and Pitcher, T.J. (1979): *Age Groups from Size-Frequency Data : A Versitile and Efficient Method of Analyzing Distribution Mixtures.* Journal of the Fishery Research Board of Canada, Vol. 26, pp. 987-1001.

McIntyre, G.A., Brooks, C., Compston, W. and Turek, A. (1966): *The Statistical Assessment of Rb-Sr Isochrons.* Journal of Geophysical Research, Vol. 71, No. 22, pp. 5459-5468.

McLachlan, G.J. and Basford, K.E. (1988): *Mixture Models : Inference and Applications to Clustering.* Marcel Dekker, Inc., New York, 253 p.

Madansky, A. (1959): *The Fitting of Straight Lines When Both Variables Are Subject to Error.* Journal of the American Statistical Society, Vol. 54, pp. 173-205.

Mantoglou, A. (1987): *Digital Simulation of Multivariate Two- and Three-Dimensional Stochastic Processes with a Spectral Turning Bands Method.* Mathematical Geology, Vol. 19, No. 2, pp. 129-149.

Mark, D.M. and Church, M. (1977): *On the Misuse of Regression in Earth Science.* Mathematical Geology, Vol. 9, No. 1, pp. 63-75.

Martin, E.S. (1936): *A Study of an Egyptian Series of Mandibles; With Special Reference to Sexing.* Biometrika, Vol. 28, pp. 948-967.

Matysek, P., Sinclair, A.J. and Fletcher, W.K. (1982): *Rapid Anomaly Recognition and Ranking for Multi-Element Regional Stream Sediment Surveys.* in 'Geological Fieldwork 1981', Paper 1982-1, B.C. Ministry of Energy, Mines and Petroleum Resources, pp. 176-186.

Miesch, A.T. (1981): *Estimation of the Geochemical Threshold and its Statistical Significance.* Journal of Geochemical Exploration, Vol. 16, pp. 49-76.

Moran, P.A.P. (1971): *Estimating Structural and Functional Relationships.* Journal of Multivariate Analysis, Vol. 1, pp. 232-255.

Mundry, E. (1972): *On the Resolution of Mixed Frequency Distributions into Normal Components.* Mathematical Geology, Vol. 4, No. 1, pp. 55-60.

Nash, J.C. (1979): *Compact Numerical Methods for Computers: Linear Algebra and Function Minimization*, Adam Hilger, Ltd., Bristol, pp. 141-152.

Neave, H.R. (1979): *Quick and Simple Tests Based on Extreme Observations.* Journal of Quality Technology, Vol. 11, No. 2, pp. 66-79.

Parr, W.C. (1981): *Minimum Distance Estimation : A Bibliography.* Communications in Statistics A, Vol. 10, pp. 1205-1224.

Parslow, G.R. (1974): *Determination of Background and Threshold in Exploration Geochemistry.* Journal of Geochemical Exploration, Vol. 3, pp. 319-336.

Pearson, K. (1894): *Contribution to the Mathematical Theory of Evolution.* Philosophical Transactions of the Royal Society – A, Vol. 185, pp. 71-110.

Perillo, G.M. and Marone, E. (1986a): *Determination of Optimal Numbers of Class Intervals Using Maximum Entropy.* Mathematical Geology, Vol. 18, No. 4,

pp. 401-408.

Perillo, G.M. and Marone, E. (1986b): *Applications of Maximum Entropy and Optimal Number of Class Interval Concepts: Two Examples.* Mathematical Geology, Vol. 18, No. 5, pp. 465-476.

Popper, K.R. (1968): *The Logic of Scientific Discovery.* Harper and Row, New York, 480 p.

Preston, E.J. (1953): *A Graphical Method for the Analysis of Statistical Distributions into Two Normal Components.* Biometrika, Vol. 40, pp. 460-464.

Quandt, R.E. and Ramsey, J.B. (1978): *Estimating Mixtures of Normal Distributions and Switching Regressions.* Journal of the American Statistical Association, Vol. 73, No. 364, pp. 730-752, (includes several comments and a rejoinder).

Rao, C.R. (1965): *Linear Statistical Inference and Its Applications.* Wiley and Co., New York, 320 p.

Redner, R.A. and Walker, H.F. (1984): *Mixture Densities, Maximum Likelihood and the EM Algorithm.* SIAM Review, Vol. 26, pp. 195-239.

Ripley, B.D. and Thompson, M. (1987): *Regression Techniques for the Detection of Bias.* Analyst, Vol. 112, pp. 377-383.

Rodionov, D.A. (1961): *On the Lognormal Distribution of the Elements in Igneous Rocks.* Geokhimiya, No. 4, pp. 324-327.

Roquin, C. and Zeegers, H. (1987): *Improving Anomaly Selection by Statistical Estimation of Background Variations in Regional Geochemical Prospecting.* Journal of Geochemical Exploration, Vol. 29, pp. 295-316.

Rose, A.W., Hawkes, H.E. and Webb, J.S. (1979): *Geochemistry in Mineral Exploration.* 2^{nd} Edition, Academic Press, New York, 657 p.

Sahu, B.K. (1973): *Comments on 'On the Resolution of Mixed Frequency Distributions into Normal Components'.* Mathematical Geology, Vol. 5, No. 2, pp. 208-209.

Sedgewick, R. (1983): *Algorithms.* Addison-Wesley Publishing Company, Dons Mills, Ontario, 551 p.

Shaw, D.M. (1961): *Element Distribution Laws in Geochemistry.* Geochimica et Cosmochimica Acta, Vol. 23, pp. 116-134.

Sheesley, J.H. (1977): *Tests for Outlying Observations.* Journal of Quality Technology, Vol. 9, No. 1, pp. 38-41.

Silverman, B.W. (1981): *Using Kernel Density Estimates to Investigate Multimodality.* Journal of the Royal Statistical Society – B, Vol. 43, No. 1, pp. 97-99.

Sinclair, A.J. (1974): *Selection of Threshold Values in Geochemical Data Using Probability Graphs.* Journal of Geochemical Exploration, Vol. 3, pp. 129-149.

Sinclair, A.J. (1976): *Application of Probability Graphs in Mineral Exploration.* Association of Exploration Geochemists, Special Volume No. 4, 95 p.

Smith, A.F.M. and Makov, U.E. (1982): *A Quasi-Bayes Sequential Procedure For Mixtures.* Journal of the Royal Statistical Society – B, Vol. 40, pp. 106-111.

Smith, A.F.M., Skene, A.M., Shaw, J.E.H., Naylor, J. and Dransfield, M. (1985): *The Implementation of the Bayesian Paradigm.* Communications in Statistics – A, Vol. 13, pp. 48-67.

Solari, M.E. (1969): *The "Maximum Likelihood Solution" of the Problem of Estimating a Linear Functional Relationship.* Journal of the Royal Statistical Society – B, Vol. 31, No. 2, pp. 372-375.

Stanley, C.R. (1984): *The Geology and Geochemistry of the Daisy Creek Prospect, A Stratabound Copper-Silver Occurrence in Western Montana.* Unpublished M.Sc. Thesis, University of British Columbia, 288 p.

Stanley, C.R. (1986): *Relative Error Analysis of Replicate Geochemical Data : Advantages and Applications.* Programs and Abstracts, GeoExpo - 86 : Exploration

in the North American Cordillera, Association of Exploration Geochemists Regional Symposium, Vancouver, British Columbia, May 1986, pp. 77-78.

Stanley, C.R. (1987): *PROBPLOT - An Interactive Program to Fit Mixtures of Normal (or Log-Normal) Distributions using Maximum Likelihood Optimization Procedures.* Association of Exploration Geochemists, Special Volume No. 14, 1 diskette, 40 p.

Stanley, C.R. and Sinclair, A.J. (1987): *Anomaly Recognition for Multi-Element Geochemical Data – A Background Characterization Approach.* Journal of Geochemical Exploration, Vol. 29, pp. 333-353.

Stanley, C.R. and Sinclair, A.J. (1988): *Univariate Patterns in the Design of Multivariate Analysis Techniques for Geochemical Data Evaluation.* in "Quantitative Analysis of Mineral and Energy Resources", Chung, C.F., Fabbri, A.G. and Sinding-Larsen, R., eds., NATO ASI Series C : Mathematical and Physical Sciences, Reidel Publishing Co., Boston, Vol. 223, pp. 131-143.

Strömgren, B. (1954): *Tables and Diagrams for Dissecting a Frequency Curve into Components by the Half-Invariant Method.* Skandinavisk Aktuarietidskrift, Vol. 17, pp. 7-54.

Tan, W.Y. and Chang, W.C. (1972): *Some Comparisons of the Method of Moments and the Method of Maximum Likelihood in Estimating Parameters of a Mixture of Two Normal Densities.* Journal of the American Statistical Association, Vol. 67, No. 339, pp. 702-708.

Tanaka, S. (1962): *A Method of Analysing a Polymodal Frequency Distribution and its Application to the Length Distribution of the Porgy, Taius tumifrons (T and S).* Journal of the Fishery Research Board of Canada, Vol. 19, pp. 1143-1159.

Tanner, W.F. (1959): *Sample Components Obtained by the Method of Differences.* Journal of Sedimentary Petrology, Vol. 29, No. 3, pp. 408-411.

Teitjen, G.L. and Moore, R.H. (1972): *Some Grubbs-Type Statistics for the Detection of Several Outliers.* Technometrics, Vol. 14, No. 3, pp. 583-597.

Tennant, C.B. and White, M.L. (1959): *Study of Distributions of Some Geochemical Data.* Economic Geology, Vol. 54, pp. 1281-1290.

Thompson, J.B. (1982a): *Composition Space : An Algebraic and Geometric Approach.* in "Characterization of Metamorphism Through Mineral Equilibria", Reviews in Mineralogy, Vol. 10, Ferry, J.M., editor, Mineralogical Society of America, pp. 1-31.

Thompson, J.B. (1982b): *Reaction Space : An Algebraic and Geometric Approach.* in "Characterization of Metamorphism Through Mineral Equilibria", Reviews in Mineralogy, Vol. 10, Ferry, J.M., editor, Mineralogical Society of America, pp. 32-52.

Thompson, M. (1973): *DUPAN 3, A Subroutine for the Interpretation of Duplicated Data in Geochemical Analysis.* Computers and Geosciences, Vol. 4, pp. 333-340.

Thompson, M. (1982): *Regression Methods in the Comparison of Accuracy.* Analyst, Vol. 107, pp. 1169-1180.

Thompson, M. and Howarth, R.J. (1973): *The Rapid Estimation and Control of Precision by Duplicate Determinations.* Analyst, Vol. 98, pp. 153-160.

Thompson, M. and Howarth, R.J. (1976a): *Duplicate Analysis in Geochemical Practice - Part 1. Theoretical Approach and Estimation of Analytical Reproducibility.* Analyst, Vol. 101, pp. 690-698.

Thompson, M. and Howarth, R.J. (1976b): *Duplicate Analysis in Geochemical Practice - Part 2. Examination of Proposed Method and Examples of its Use.* Analyst, Vol. 101, pp. 699-709.

Thompson, M. and Howarth, R.J. (1978): *A New Approach to the Estimation of Analytical Precision.* Journal of Geochemical Exploration, Vol. 9, pp. 23-30.

Till, R. (1973): *The Use of Linear Regression in Geomorphology.* Area, Vol. 5, No. 4, pp. 303-308.

Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985): *Statistical Analysis of Finite Mixture Distributions.* John Wiley and Sons, Toronto, 243 p.

Van Andel, T.H. (1973): *Texture and Dispersal of Sediments in the Panama Basin.* Journal of Geology, Vol. 81, pp. 434-457.

Villegas, C. (1961): *Maximum Likelihood Estimation of a Linear Functional Relationship.* Annals of Mathematical Statistics, Vol. 32, pp. 1048-1062.

Visman, J. (1969): *A General Sampling Theory.* Materials Research and Standards, Vol. 9, No. 11, pp. 8-13.

Visman, J. (1972): *A General Theory of Sampling - Discussion 3.* Journal of Materials, Vol. 7, No. 3, pp. 345-350.

Williams, X.K. (1967): *Statistics in the Interpretation of Geochemical Data.* New Zealand Journal of Geology and Geophysics, Vol. 10, pp. 771-797.

Wolfe, J.H. (1971): *A Monte Carlo Study of the Sampling Distribution of the Likelihood Ratio for Mixtures of Multinomial Distributions.* Technical Bulletin STB 72-2, Naval Personel and Training Research Laboratory, San Diego.

Wolfowitz, J. (1957): *The Minimum Distance Method.* Annals of Mathematical Statistics, Vol. 28, pp. 75-88.

Wurzer, F. (1988): *Application of Robust Statistics in the Analysis of Geochemical Data.* in "Qunatitative Analysis of Mineral and Energy Resources", Chung, C.F., Fabbri, A.G. and Sinding-Larsen, R., eds., NATO ASI Series C : Mathematical and Physical Sciences, Reidel Publishing Co., Boston, Vol. 223, pp. 131-143.

York, D. (1966): *Least-Squares Fitting of a Straight Line.* Canadian Journal of Physics, Vol. 44, pp. 1079-1086.

York, D. (1967): *The Best Isochron.* Earth and Planetary Science Letters, Vol. 2,

pp. 479-482.

York, D. (1969): *Least-Squares Fitting of a Straight Line With Correlated Errors.* Earth and Planetary Science Letters, Vol. 5, pp. 320-324.

Appendix A

Maximum Likelihood Parameter Estimate Comparison

"Statistics is the only profession which demands the right to be wrong 5 % of the time."

Anonymous

A.1 Tabulated Differences of *ML* Parameter Estimates

The following tables present the differences between the *CIDML* parameter estimates, the *RDML* parameter estimates, the stochastic parameter values calculated from the different data sets and the actual population parameter values. Each table refers to a different data set structure, as indicated, with the exception of data set structure # 16, which was not evaluated. These tables have a format similar to Table 3.5 with the exception that the minimum χ^2 parameter estimates were not determined for these data set structures. All differences are expressed relative to the actual population parameter values.

See Table 3.4 for the definitions of all symbols used in the following tables.

Table A.17: Differences Between the *RDML* Parameter Estimates, *CIDML* Parameter Estimates and Stochastically Generated Sample Parameter Values and the Population Parameter Values for Data Set Structure # 2

($\mu_1 = 20$, $\mu_2 = 40$, $\sigma_1 = 5$, $\sigma_2 = 5$, $\varpi(\%) = 50$, **and** $n = 100$)

	$\bar{d}_{\hat{\mu}_1}$	$\bar{s}_{\hat{\mu}_1}$	$\bar{d}_{\hat{\mu}_2}$	$\bar{s}_{\hat{\mu}_2}$	$\bar{d}_{\hat{\sigma}_1}$	$\bar{s}_{\hat{\sigma}_1}$	$\bar{d}_{\hat{\sigma}_2}$	$\bar{s}_{\hat{\sigma}_2}$	$\bar{d}_{\hat{\varpi}(\%)}$	$\bar{s}_{\hat{\varpi}(\%)}$
Stochastic	-0.20	0.51	0.07	0.77	0.04	0.40	0.16	0.40	0.00	0.00
$\ell\ \infty$	-0.61	0.88	-0.33	1.11	-0.36	0.63	0.43	0.55	-1.93	2.61
ℓ 10	0.39	1.55	-0.78	1.32	-0.09	1.33	-1.76	1.23	4.87	7.33
ℓ 15	0.23	1.78	-0.17	1.76	-0.22	1.35	-1.32	1.03	3.42	7.74
ℓ 20	0.24	1.67	-0.26	1.59	-0.17	1.32	-0.84	1.00	2.34	8.02
ℓ 25	-0.17	0.70	-0.49	0.62	-0.31	0.73	-0.37	0.35	0.04	2.92
ℓ 30	-0.29	0.53	-0.67	0.61	-0.37	0.63	-0.15	0.36	-0.77	2.46
ℓ 35	-0.27	0.60	-0.56	0.57	-0.36	0.64	-0.14	0.36	-0.82	2.79
ℓ 40	-0.10	0.80	-0.49	0.67	-0.24	0.72	-0.28	0.42	-0.04	3.57
ℓ 45	-0.33	0.66	-0.50	0.64	-0.40	0.62	-0.22	0.48	-0.75	3.30
ℓ 50	-0.25	0.62	-0.53	0.54	-0.36	0.64	-0.18	0.43	-0.62	2.80

Table A.18: Differences Between the *RDML* Parameter Estimates, *CIDML* Parameter Estimates and Stochastically Generated Sample Parameter Values and the Population Parameter Values for Data Set Structure # 3

$(\mu_1 = 20,\ \mu_2 = 40,\ \sigma_1 = 5,\ \sigma_2 = 5,\ \varpi(\%) = 50,$ **and** $n = 300)$

	$\bar{d}_{\hat{\mu}_1}$	$\bar{s}_{\hat{\mu}_1}$	$\bar{d}_{\hat{\mu}_2}$	$\bar{s}_{\hat{\mu}_2}$	$\bar{d}_{\hat{\sigma}_1}$	$\bar{s}_{\hat{\sigma}_1}$	$\bar{d}_{\hat{\sigma}_2}$	$\bar{s}_{\hat{\sigma}_2}$	$\bar{d}_{\hat{\varpi}(\%)}$	$\bar{s}_{\hat{\varpi}(\%)}$
Stochastic	-0.11	0.25	0.19	0.22	0.02	0.24	-0.18	0.29	0.00	0.00
$\ell\ \infty$	-0.21	0.43	0.12	0.36	-0.11	0.38	-0.15	0.22	-0.39	1.30
ℓ 10	0.34	0.64	-0.13	0.48	-0.30	0.54	-1.32	0.69	1.99	3.08
ℓ 15	0.08	0.33	-0.05	0.39	-0.28	0.27	-0.62	0.34	0.52	1.65
ℓ 20	-0.08	0.30	-0.17	0.26	-0.24	0.27	-0.35	0.25	0.08	1.23
ℓ 25	0.02	0.34	-0.06	0.36	-0.08	0.22	-0.30	0.31	0.38	1.44
ℓ 30	-0.06	0.32	-0.06	0.34	-0.13	0.29	-0.31	0.28	0.21	1.60
ℓ 35	0.01	0.26	-0.04	0.31	-0.09	0.25	-0.28	0.31	0.37	1.42
ℓ 40	-0.04	0.31	-0.08	0.25	-0.09	0.25	-0.23	0.32	0.16	1.52
ℓ 45	0.04	0.36	-0.07	0.36	-0.08	0.26	-0.23	0.34	0.31	1.69
ℓ 50	0.00	0.35	-0.15	0.50	-0.12	0.34	-0.27	0.39	0.05	1.96

Table A.19: Differences Between the *RDML* Parameter Estimates, *CIDML* Parameter Estimates and Stochastically Generated Sample Parameter Values and the Population Parameter Values for Data Set Structure # 4

$(\mu_1 = 20,\ \mu_2 = 40,\ \sigma_1 = 5,\ \sigma_2 = 5,\ \varpi(\%) = 50,$ **and** $n = 400)$

	$\overline{d}_{\hat{\mu}_1}$	$\overline{s}_{\hat{\mu}_1}$	$\overline{d}_{\hat{\mu}_2}$	$\overline{s}_{\hat{\mu}_2}$	$\overline{d}_{\hat{\sigma}_1}$	$\overline{s}_{\hat{\sigma}_1}$	$\overline{d}_{\hat{\sigma}_2}$	$\overline{s}_{\hat{\sigma}_2}$	$\overline{d}_{\hat{\varpi}(\%)}$	$\overline{s}_{\hat{\varpi}(\%)}$
Stochastic	-0.01	0.42	-0.12	0.32	-0.09	0.12	-0.03	0.26	0.00	0.00
$\ell\ \infty$	0.06	0.58	-0.07	0.49	-0.02	0.32	-0.05	0.37	0.31	1.38
ℓ 10	0.54	0.48	-0.40	0.51	-0.04	0.39	-1.60	0.68	3.84	2.82
ℓ 15	0.22	0.37	-0.16	0.38	-0.09	0.31	-0.68	0.49	1.39	1.83
ℓ 20	0.19	0.34	-0.02	0.27	0.08	0.26	-0.42	0.30	1.23	1.58
ℓ 25	0.09	0.30	0.01	0.27	0.02	0.29	-0.27	0.31	0.77	1.53
ℓ 30	0.06	0.32	-0.03	0.31	0.00	0.29	-0.24	0.33	0.60	1.63
ℓ 35	0.11	0.32	0.03	0.33	0.07	0.29	-0.20	0.28	0.79	1.60
ℓ 40	0.10	0.31	0.03	0.31	0.07	0.29	-0.22	0.30	0.80	1.68
ℓ 45	0.07	0.33	-0.03	0.34	0.01	0.26	-0.29	0.37	0.68	1.59
ℓ 50	0.14	0.28	-0.11	0.24	0.02	0.20	-0.18	0.24	0.56	1.25

Table A.20: Differences Between the *RDML* Parameter Estimates, *CIDML* Parameter Estimates and Stochastically Generated Sample Parameter Values and the Population Parameter Values for Data Set Structure # 5

($\mu_1 = 20$, $\mu_2 = 40$, $\sigma_1 = 5$, $\sigma_2 = 5$, $\varpi(\%) = 50$, **and** $n = 500$)

	$\bar{d}_{\hat{\mu}_1}$	$\bar{s}_{\hat{\mu}_1}$	$\bar{d}_{\hat{\mu}_2}$	$\bar{s}_{\hat{\mu}_2}$	$\bar{d}_{\hat{\sigma}_1}$	$\bar{s}_{\hat{\sigma}_1}$	$\bar{d}_{\hat{\sigma}_2}$	$\bar{s}_{\hat{\sigma}_2}$	$\bar{d}_{\hat{\varpi}(\%)}$	$\bar{s}_{\hat{\varpi}(\%)}$
Stochastic	-0.10	0.26	0.03	0.34	-0.16	0.20	-0.02	0.18	0.00	0.00
$\ell\ \infty$	-0.17	0.34	-0.05	0.38	-0.22	0.18	0.04	0.24	-0.38	0.67
ℓ 10	0.28	0.54	-0.19	0.34	-0.35	0.49	-1.34	0.71	2.33	2.89
ℓ 15	-0.09	0.32	-0.28	0.24	-0.33	0.17	-0.34	0.39	-0.27	1.19
ℓ 20	-0.08	0.22	-0.12	0.17	-0.13	0.16	-0.26	0.21	0.01	0.84
ℓ 25	-0.01	0.24	-0.06	0.29	-0.13	0.21	-0.21	0.19	-0.04	0.70
ℓ 30	-0.01	0.18	-0.06	0.13	-0.07	0.14	-0.19	0.18	0.11	0.80
ℓ 35	-0.03	0.13	-0.13	0.15	-0.09	0.11	-0.14	0.14	-0.06	0.63
ℓ 40	0.02	0.21	-0.07	0.16	-0.04	0.14	-0.12	0.19	0.01	0.94
ℓ 45	-0.03	0.27	-0.21	0.27	-0.05	0.20	-0.03	0.27	-0.27	1.20
ℓ 50	-0.04	0.26	-0.15	0.24	0.02	0.17	-0.07	0.27	-0.03	0.98

Table A.21: Differences Between the *RDML* Parameter Estimates, *CIDML* Parameter Estimates and Stochastically Generated Sample Parameter Values and the Population Parameter Values for Data Set Structure # 6

($\mu_1 = 20$, $\mu_2 = 40$, $\sigma_1 = 5$, $\sigma_2 = 5$, $\varpi(\%) = 70$, **and** $n = 200$)

	$\bar{d}_{\hat{\mu}_1}$	$\bar{s}_{\hat{\mu}_1}$	$\bar{d}_{\hat{\mu}_2}$	$\bar{s}_{\hat{\mu}_2}$	$\bar{d}_{\hat{\sigma}_1}$	$\bar{s}_{\hat{\sigma}_1}$	$\bar{d}_{\hat{\sigma}_2}$	$\bar{s}_{\hat{\sigma}_2}$	$\bar{d}_{\hat{\varpi}(\%)}$	$\bar{s}_{\hat{\varpi}(\%)}$
Stochastic	-0.04	0.30	-0.22	0.69	0.11	0.37	0.08	0.39	0.00	0.00
ℓ ∞	-0.01	0.45	-0.06	0.94	0.15	0.35	-0.01	0.48	0.13	1.38
ℓ 10	0.34	0.38	-0.40	0.63	-0.11	0.41	-1.36	0.93	3.42	3.34
ℓ 15	0.06	0.39	-0.19	0.74	-0.18	0.37	-0.88	0.63	1.40	2.32
ℓ 20	0.16	0.30	-0.13	0.53	0.08	0.24	-0.62	0.52	1.53	1.61
ℓ 25	0.13	0.42	0.11	0.64	0.03	0.30	-0.62	0.37	1.24	1.68
ℓ 30	0.04	0.19	-0.09	0.44	-0.01	0.22	-0.58	0.31	0.93	1.13
ℓ 35	0.09	0.22	0.01	0.50	0.07	0.21	-0.53	0.37	1.14	1.34
ℓ 40	0.07	0.20	-0.12	0.48	-0.00	0.21	-0.53	0.34	0.91	1.16
ℓ 45	-0.00	0.21	-0.12	0.43	-0.06	0.20	-0.53	0.28	0.71	1.11
ℓ 50	0.04	0.27	-0.20	0.48	-0.01	0.21	-0.44	0.31	0.66	1.36

Table A.22: Differences Between the *RDML* Parameter Estimates, *CIDML* Parameter Estimates and Stochastically Generated Sample Parameter Values and the Population Parameter Values for Data Set Structure # 7

$(\mu_1 = 20,\ \mu_2 = 40,\ \sigma_1 = 5,\ \sigma_2 = 5,\ \varpi(\%) = 85,$ **and** $n = 200)$

	$\bar{d}_{\hat{\mu}_1}$	$\bar{s}_{\hat{\mu}_1}$	$\bar{d}_{\hat{\mu}_2}$	$\bar{s}_{\hat{\mu}_2}$	$\bar{d}_{\hat{\sigma}_1}$	$\bar{s}_{\hat{\sigma}_1}$	$\bar{d}_{\hat{\sigma}_2}$	$\bar{s}_{\hat{\sigma}_2}$	$\bar{d}_{\hat{\varpi}(\%)}$	$\bar{s}_{\hat{\varpi}(\%)}$
Stochastic	0.11	0.53	0.03	1.00	-0.07	0.28	0.22	0.45	0.00	0.00
$\ell\ \infty$	-0.14	0.48	-1.13	1.57	-1.32	0.39	0.88	0.82	-2.24	2.02
ℓ 10	-0.15	0.51	-1.95	1.87	-0.61	0.33	-2.10	1.12	0.25	3.65
ℓ 15	-0.25	0.46	-2.08	1.89	-0.48	0.31	-0.93	1.03	-0.89	3.13
ℓ 20	-0.20	0.35	-1.56	1.48	-0.33	0.36	-0.89	0.90	-0.14	1.81
ℓ 25	-0.15	0.43	-1.54	1.40	-0.27	0.31	-0.33	1.24	-1.24	2.69
ℓ 30	-0.18	0.33	-1.54	1.40	-0.29	0.31	-0.29	1.24	-1.41	3.16
ℓ 35	-0.24	0.37	-1.59	1.50	-0.24	0.32	-0.22	1.39	-1.39	3.01
ℓ 40	-0.24	0.37	-1.69	1.54	-0.30	0.39	-0.10	1.28	-1.61	3.14
ℓ 45	-0.11	0.38	-1.20	1.11	-0.23	0.37	-0.30	1.05	-0.64	2.42
ℓ 50	-0.17	0.29	-1.31	1.17	-0.24	0.29	-0.26	1.18	-1.03	2.42

Table A.23: Differences Between the *RDML* Parameter Estimates, *CIDML* Parameter Estimates and Stochastically Generated Sample Parameter Values and the Population Parameter Values for Data Set Structure # 8

($\mu_1 = 20$, $\mu_2 = 40$, $\sigma_1 = 5$, $\sigma_2 = 5$, $\varpi(\%) = 95$, **and** $n = 200$)

	$\bar{d}_{\hat{\mu}_1}$	$\bar{s}_{\hat{\mu}_1}$	$\bar{d}_{\hat{\mu}_2}$	$\bar{s}_{\hat{\mu}_2}$	$\bar{d}_{\hat{\sigma}_1}$	$\bar{s}_{\hat{\sigma}_1}$	$\bar{d}_{\hat{\sigma}_2}$	$\bar{s}_{\hat{\sigma}_2}$	$\bar{d}_{\hat{\varpi}(\%)}$	$\bar{s}_{\hat{\varpi}(\%)}$
Stochastic	-0.03	0.22	0.60	1.20	-0.08	0.30	0.03	1.02	0.00	0.00
$\ell\ \infty$	-0.16	0.22	-1.33	2.37	-0.19	0.32	1.07	1.83	-1.27	1.51
ℓ 10	-0.20	0.20	-3.91	3.24	-0.46	0.23	-2.36	0.98	0.31	1.12
ℓ 15	-0.03	0.18	-1.83	1.23	-0.28	0.24	-1.94	0.69	0.72	1.19
ℓ 20	-0.02	0.19	-2.19	1.60	-0.15	0.17	-1.09	0.72	0.21	0.92
ℓ 25	-0.02	0.11	-1.26	1.02	-0.10	0.12	-1.39	0.68	0.35	0.98
ℓ 30	-0.05	0.15	-1.57	1.22	-0.07	0.09	-1.22	0.78	0.30	0.80
ℓ 35	-0.04	0.13	-1.48	1.03	-0.10	0.11	-1.39	0.74	0.27	0.82
ℓ 40	-0.01	0.14	-1.15	1.01	-0.05	0.18	-1.47	0.76	0.53	0.75
ℓ 45	-0.08	0.27	-1.64	1.16	-0.08	0.17	-1.12	0.93	0.18	0.99
ℓ 50	-0.04	0.12	-1.54	1.19	-0.08	0.12	-1.23	0.91	0.30	0.75

Table A.24: Differences Between the *RDML* Parameter Estimates, *CIDML* Parameter Estimates and Stochastically Generated Sample Parameter Values and the Population Parameter Values for Data Set Structure # 9

$(\mu_1 = 20,\ \mu_2 = 25,\ \sigma_1 = 5,\ \sigma_2 = 5,\ \varpi(\%) = 50,$ **and** $n = 200)$

	$\bar{d}_{\hat{\mu}_1}$	$\bar{s}_{\hat{\mu}_1}$	$\bar{d}_{\hat{\mu}_2}$	$\bar{s}_{\hat{\mu}_2}$	$\bar{d}_{\hat{\sigma}_1}$	$\bar{s}_{\hat{\sigma}_1}$	$\bar{d}_{\hat{\sigma}_2}$	$\bar{s}_{\hat{\sigma}_2}$	$\bar{d}_{\hat{\varpi}(\%)}$	$\bar{s}_{\hat{\varpi}(\%)}$
Stochastic	0.10	0.64	-0.06	0.47	0.15	0.29	0.05	0.35	0.00	0.00
$\ell\ \infty$	0.34	1.33	0.05	2.24	0.23	0.90	-0.18	0.58	-0.94	1.48
ℓ 10	-0.56	1.51	0.18	0.97	-0.52	0.85	-1.08	0.49	-0.49	10.04
ℓ 15	0.09	1.01	0.17	1.43	-0.12	0.57	-0.65	0.77	2.55	7.90
ℓ 20	-0.01	1.57	-0.10	1.70	-0.07	0.82	-0.72	0.75	-0.82	8.64
ℓ 25	-0.21	1.22	0.34	1.45	-0.20	0.78	-0.74	0.65	1.40	6.97
ℓ 30	-0.30	0.98	0.17	1.31	-0.12	0.61	-0.68	0.67	-1.06	9.83
ℓ 35	-0.10	1.81	0.17	1.98	-0.15	0.83	-0.96	0.70	4.05	16.49
ℓ 40	-0.49	1.00	0.38	1.15	-0.27	0.70	-0.68	0.58	0.16	5.34
ℓ 45	-0.30	1.08	0.35	1.27	-0.22	0.71	-0.71	0.58	1.66	6.55
ℓ 50	-0.45	1.12	0.40	1.22	-0.28	0.73	-0.71	0.54	0.44	4.90

Table A.25: Differences Between the *RDML* Parameter Estimates, *CIDML* Parameter Estimates and Stochastically Generated Sample Parameter Values and the Population Parameter Values for Data Set Structure # 10

$(\mu_1 = 20,\ \mu_2 = 30,\ \sigma_1 = 5,\ \sigma_2 = 5,\ \varpi(\%) = 50,$ **and** $n = 200)$

	$\bar{d}_{\hat{\mu}_1}$	$\bar{s}_{\hat{\mu}_1}$	$\bar{d}_{\hat{\mu}_2}$	$\bar{s}_{\hat{\mu}_2}$	$\bar{d}_{\hat{\sigma}_1}$	$\bar{s}_{\hat{\sigma}_1}$	$\bar{d}_{\hat{\sigma}_2}$	$\bar{s}_{\hat{\sigma}_2}$	$\bar{d}_{\hat{\varpi}(\%)}$	$\bar{s}_{\hat{\varpi}(\%)}$
Stochastic	0.19	0.44	0.19	0.52	0.21	0.35	-0.03	0.53	0.00	0.00
$\ell\ \infty$	0.08	1.16	0.35	1.15	0.03	0.50	-0.19	0.73	1.37	9.58
ℓ 10	0.24	1.13	0.41	0.69	-0.33	0.68	-1.23	0.56	4.99	8.30
ℓ 15	0.50	1.09	0.74	1.39	-0.04	0.39	-1.05	0.83	7.90	11.95
ℓ 20	0.32	1.23	0.61	1.30	-0.14	0.38	-0.79	0.83	4.66	12.81
ℓ 25	0.66	1.25	1.11	1.13	-0.01	0.52	-0.95	0.82	9.94	11.51
ℓ 30	0.36	1.36	0.75	1.19	-0.04	0.36	-0.86	0.76	7.22	12.89
ℓ 35	0.22	1.65	0.64	1.22	-0.13	0.42	-0.73	0.73	6.19	14.57
ℓ 40	-0.09	1.40	0.74	1.52	-0.24	0.42	-0.79	0.85	4.40	14.95
ℓ 45	0.69	1.19	0.77	1.14	0.07	0.49	-0.77	0.77	9.12	11.99
ℓ 50	0.35	1.25	0.66	1.39	-0.07	0.54	-0.81	0.95	6.18	13.26

Table A.26: Differences Between the *RDML* Parameter Estimates, *CIDML* Parameter Estimates and Stochastically Generated Sample Parameter Values and the Population Parameter Values for Data Set Structure # 11

($\mu_1 = 20$, $\mu_2 = 35$, $\sigma_1 = 5$, $\sigma_2 = 5$, $\varpi(\%) = 50$, **and** $n = 200$)

	$\bar{d}_{\hat{\mu}_1}$	$\bar{s}_{\hat{\mu}_1}$	$\bar{d}_{\hat{\mu}_2}$	$\bar{s}_{\hat{\mu}_2}$	$\bar{d}_{\hat{\sigma}_1}$	$\bar{s}_{\hat{\sigma}_1}$	$\bar{d}_{\hat{\sigma}_2}$	$\bar{s}_{\hat{\sigma}_2}$	$\bar{d}_{\hat{\varpi}(\%)}$	$\bar{s}_{\hat{\varpi}(\%)}$
Stochastic	0.16	0.70	-0.21	0.47	-0.12	0.33	-0.10	0.22	0.00	0.00
$\ell\ \infty$	-1.03	1.21	-1.18	1.69	-0.85	0.60	0.25	0.86	-7.68	9.50
ℓ 10	-0.37	1.48	-0.52	1.22	-0.56	0.73	-0.99	1.04	-1.45	9.99
ℓ 15	-0.38	1.06	-0.73	0.96	-0.55	0.63	-0.46	0.67	-2.42	6.90
ℓ 20	-0.80	1.42	-0.74	1.23	-0.66	0.77	-0.18	0.86	-4.75	9.45
ℓ 25	-0.88	1.28	-0.79	1.27	-0.61	0.73	-0.13	0.86	-5.12	9.35
ℓ 30	-1.14	1.37	-0.85	1.31	-0.74	0.76	-0.03	0.89	-6.10	9.38
ℓ 35	-0.93	1.43	-0.83	1.22	-0.68	0.83	-0.08	0.88	-5.24	9.56
ℓ 40	-1.01	1.44	-0.82	1.26	-0.65	0.81	-0.07	0.85	-5.35	9.51
ℓ 45	-1.10	1.33	-0.87	1.34	-0.72	0.71	-0.02	0.90	-6.06	9.51
ℓ 50	-1.01	1.34	-0.92	1.26	-0.66	0.71	-0.03	0.86	-5.83	9.38

Table A.27: Differences Between the *RDML* Parameter Estimates, *CIDML* Parameter Estimates and Stochastically Generated Sample Parameter Values and the Population Parameter Values for Data Set Structure # 12

($\mu_1 = 20$, $\mu_2 = 45$, $\sigma_1 = 5$, $\sigma_2 = 5$, $\varpi(\%) = 50$, **and** $n = 200$)

	$\bar{d}_{\hat{\mu}_1}$	$\bar{s}_{\hat{\mu}_1}$	$\bar{d}_{\hat{\mu}_2}$	$\bar{s}_{\hat{\mu}_2}$	$\bar{d}_{\hat{\sigma}_1}$	$\bar{s}_{\hat{\sigma}_1}$	$\bar{d}_{\hat{\sigma}_2}$	$\bar{s}_{\hat{\sigma}_2}$	$\bar{d}_{\hat{\varpi}(\%)}$	$\bar{s}_{\hat{\varpi}(\%)}$
Stochastic	0.10	0.48	0.02	0.33	-0.02	0.27	-0.06	0.50	0.00	0.00
$\ell\ \infty$	0.06	0.50	-0.07	0.36	-0.07	0.36	0.07	0.60	0.82	3.24
ℓ 10	0.12	0.43	-0.53	0.35	-0.69	0.52	-1.45	0.58	1.15	1.85
ℓ 15	0.07	0.24	-0.25	0.23	-0.24	0.28	-0.72	0.24	0.94	0.94
ℓ 20	-0.06	0.17	-0.35	0.13	-0.19	0.29	-0.37	0.20	0.52	0.90
ℓ 25	0.06	0.27	-1.33	0.33	-0.08	0.18	-0.32	0.27	0.48	0.71
ℓ 30	-0.01	0.14	-0.23	0.24	-0.02	0.17	-0.22	0.20	0.40	0.75
ℓ 35	-0.03	0.19	-0.27	0.22	-0.05	0.19	-0.16	0.28	0.32	0.72
ℓ 40	-0.00	0.17	-0.25	0.21	-0.09	0.22	-0.15	0.18	0.29	0.77
ℓ 45	-0.03	0.14	-0.25	0.19	-0.04	0.21	-0.21	0.21	0.41	0.76
ℓ 50	0.06	0.17	-0.28	0.31	0.02	0.23	-0.13	0.27	0.34	0.83

Table A.28: Differences Between the *RDML* Parameter Estimates, *CIDML* Parameter Estimates and Stochastically Generated Sample Parameter Values and the Population Parameter Values for Data Set Structure # 13

($\mu_1 = 20$, $\mu_2 = 50$, $\sigma_1 = 5$, $\sigma_2 = 5$, $\varpi(\%) = 50$, **and** $n = 200$)

	$\bar{d}_{\hat{\mu}_1}$	$\bar{s}_{\hat{\mu}_1}$	$\bar{d}_{\hat{\mu}_2}$	$\bar{s}_{\hat{\mu}_2}$	$\bar{d}_{\hat{\sigma}_1}$	$\bar{s}_{\hat{\sigma}_1}$	$\bar{d}_{\hat{\sigma}_2}$	$\bar{s}_{\hat{\sigma}_2}$	$\bar{d}_{\hat{\varpi}(\%)}$	$\bar{s}_{\hat{\varpi}(\%)}$
Stochastic	-0.01	0.59	-0.12	0.46	0.03	0.32	0.00	0.37	0.00	0.00
$\ell\ \infty$	0.02	0.59	-0.07	0.47	0.04	0.33	-0.07	0.36	0.09	0.21
ℓ 10	0.06	0.47	-0.20	0.41	-0.61	0.46	-1.35	0.57	1.04	1.27
ℓ 15	0.09	0.21	-0.16	0.30	-0.23	0.21	-0.74	0.27	0.70	0.47
ℓ 20	0.10	0.16	-0.23	0.13	-0.11	0.16	-0.48	0.22	0.49	0.33
ℓ 25	0.17	0.21	-0.05	0.26	0.03	0.14	-0.47	0.22	0.64	0.45
ℓ 30	0.05	0.22	-0.08	0.22	-0.09	0.07	-0.38	0.20	0.44	0.26
ℓ 35	0.04	0.10	-0.08	0.12	-0.01	0.14	-0.30	0.14	0.44	0.30
ℓ 40	0.05	0.09	-0.13	0.15	-0.05	0.11	-0.25	0.13	0.37	0.29
ℓ 45	0.07	0.11	-0.05	0.12	-0.01	0.11	-0.31	0.11	0.44	0.32
ℓ 50	-0.02	0.07	-0.06	0.15	0.01	0.08	-0.27	0.14	0.39	0.28

Table A.29: Differences Between the *RDML* Parameter Estimates, *CIDML* Parameter Estimates and Stochastically Generated Sample Parameter Values and the Population Parameter Values for Data Set Structure # 14

($\mu_1 = 20$, $\mu_2 = 40$, $\sigma_1 = 5$, $\sigma_2 = 10$, $\varpi(\%) = 50$, **and** $n = 200$)

	$\overline{d}_{\hat{\mu}_1}$	$\overline{s}_{\hat{\mu}_1}$	$\overline{d}_{\hat{\mu}_2}$	$\overline{s}_{\hat{\mu}_2}$	$\overline{d}_{\hat{\sigma}_1}$	$\overline{s}_{\hat{\sigma}_1}$	$\overline{d}_{\hat{\sigma}_2}$	$\overline{s}_{\hat{\sigma}_2}$	$\overline{d}_{\hat{\varpi}(\%)}$	$\overline{s}_{\hat{\varpi}(\%)}$
Stochastic	0.08	0.49	0.28	0.74	-0.12	0.28	0.19	0.47	0.00	0.00
ℓ ∞	0.26	0.83	0.78	1.91	-0.04	0.37	-0.10	0.76	1.42	5.33
ℓ 10	0.32	0.95	1.52	2.06	-0.28	0.69	-2.40	1.44	6.34	6.80
ℓ 15	0.27	0.46	0.58	1.65	-0.13	0.49	-1.09	0.98	2.71	4.94
ℓ 20	0.35	0.84	0.88	1.90	0.07	0.49	-1.26	1.14	3.66	5.99
ℓ 25	0.32	0.67	0.80	1.72	0.10	0.48	-1.12	0.99	3.50	5.41
ℓ 30	0.62	0.85	1.13	2.03	0.03	0.50	-1.21	1.05	3.81	5.93
ℓ 35	0.35	0.63	0.82	1.82	0.12	0.42	-1.04	1.07	2.47	5.00
ℓ 40	0.36	0.69	1.06	1.99	0.17	0.51	-1.12	1.15	3.89	6.00
ℓ 45	0.35	0.70	0.85	1.92	0.13	0.49	-1.02	1.09	3.39	5.95
ℓ 50	0.35	0.65	0.82	1.88	0.13	0.45	-1.04	1.14	3.32	5.72

Table A.30: Differences Between the *RDML* Parameter Estimates, *CIDML* Parameter Estimates and Stochastically Generated Sample Parameter Values and the Population Parameter Values for Data Set Structure # 15

($\mu_1 = 20$, $\mu_2 = 40$, $\sigma_1 = 5$, $\sigma_2 = 15$, $\varpi(\%) = 50$, **and** $n = 200$)

	$\overline{d}_{\hat{\mu}_1}$	$\overline{s}_{\hat{\mu}_1}$	$\overline{d}_{\hat{\mu}_2}$	$\overline{s}_{\hat{\mu}_2}$	$\overline{d}_{\hat{\sigma}_1}$	$\overline{s}_{\hat{\sigma}_1}$	$\overline{d}_{\hat{\sigma}_2}$	$\overline{s}_{\hat{\sigma}_2}$	$\overline{d}_{\hat{\varpi}(\%)}$	$\overline{s}_{\hat{\varpi}(\%)}$
Stochastic	0.07	0.47	-0.08	1.76	0.04	0.21	0.11	1.24	0.00	0.00
ℓ ∞	0.42	0.70	3.07	3.08	0.60	0.32	-2.62	1.76	7.53	5.99
ℓ 10	1.37	1.62	6.90	4.12	0.77	1.45	-5.52	2.74	16.81	7.30
ℓ 15	0.45	0.55	3.00	3.79	0.15	0.63	-3.16	2.12	8.83	8.01
ℓ 20	0.56	0.65	4.25	3.40	0.48	0.72	-3.36	1.93	10.90	7.01
ℓ 25	0.64	0.41	5.10	2.21	0.87	0.41	-3.50	1.12	13.21	3.67
ℓ 30	0.35	0.36	3.46	3.17	0.56	0.51	-2.76	1.49	9.31	6.02
ℓ 35	0.54	0.43	4.07	2.89	0.71	0.53	-3.01	1.44	10.87	5.76
ℓ 40	0.60	0.40	4.62	2.69	0.73	0.49	-3.21	1.47	11.69	5.56
ℓ 45	0.41	0.52	3.57	3.57	0.52	0.60	-2.66	1.90	9.31	7.32
ℓ 50	0.41	0.40	3.80	3.80	0.64	0.47	-2.81	1.69	9.72	6.28

A.2 Graphical Comparison of *ML* Parameter Estimates

The following figures graphically depict the differences between the parameters which are listed in the tables above. At the bottom of each of these plots, "S" refers to the sample statistic differences, "I" refers to the *RDML* parameter estimate differences, and the numbers indicate the number of class intervals used for determination of the *CIDML* parameter estimate differences. All differences are depicted in the parameter units, except the component proportion differences, which are depicted in 100× the parameter units.

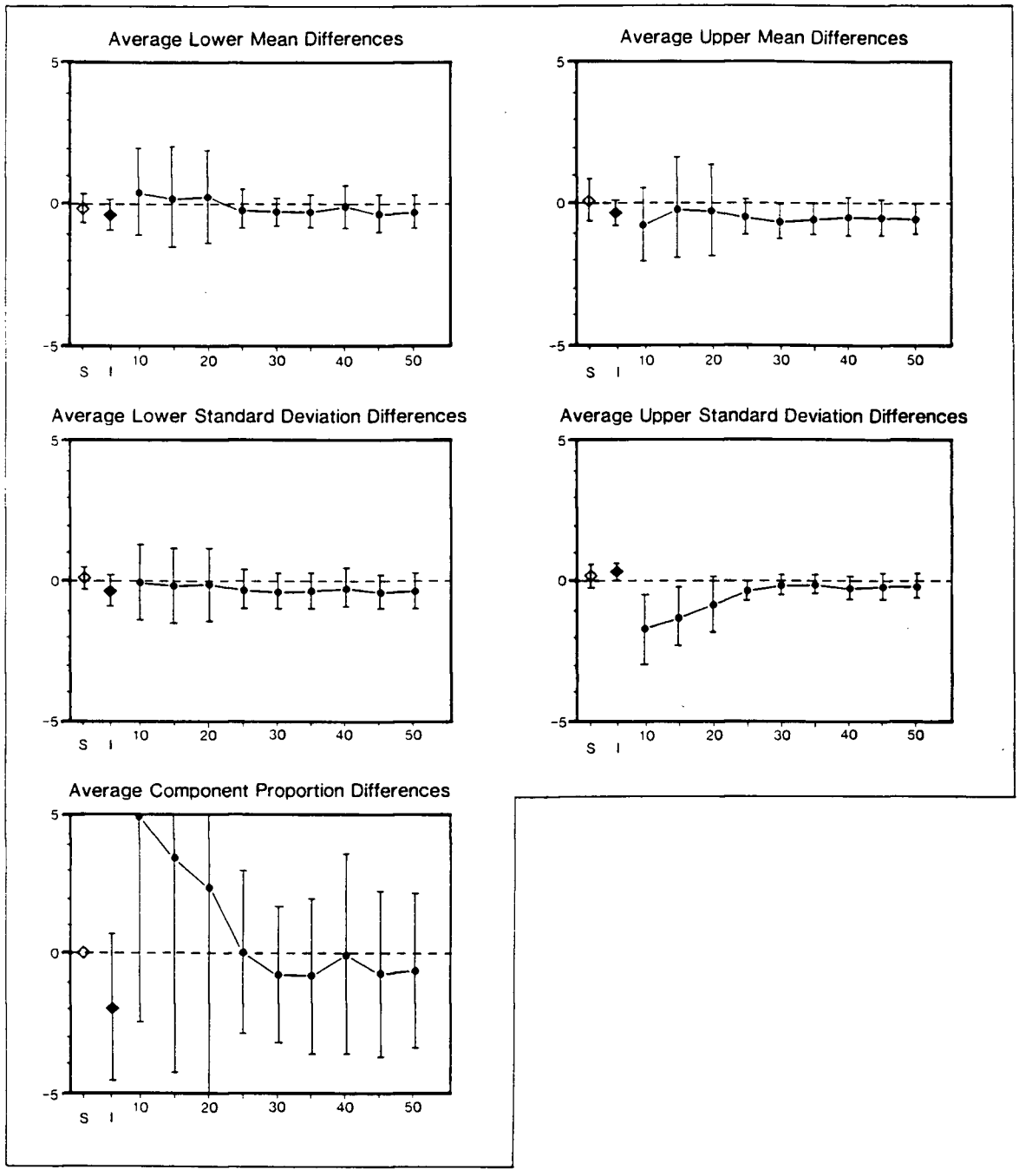


Figure A.26: Parameter Bias of the *RDML* Parameter Estimates, the *CIDML* Parameter Estimates, and the Stochastically Generated Sample Parameter Estimates Relative to the True Population Parameter Estimates for Data Set Structure # 2

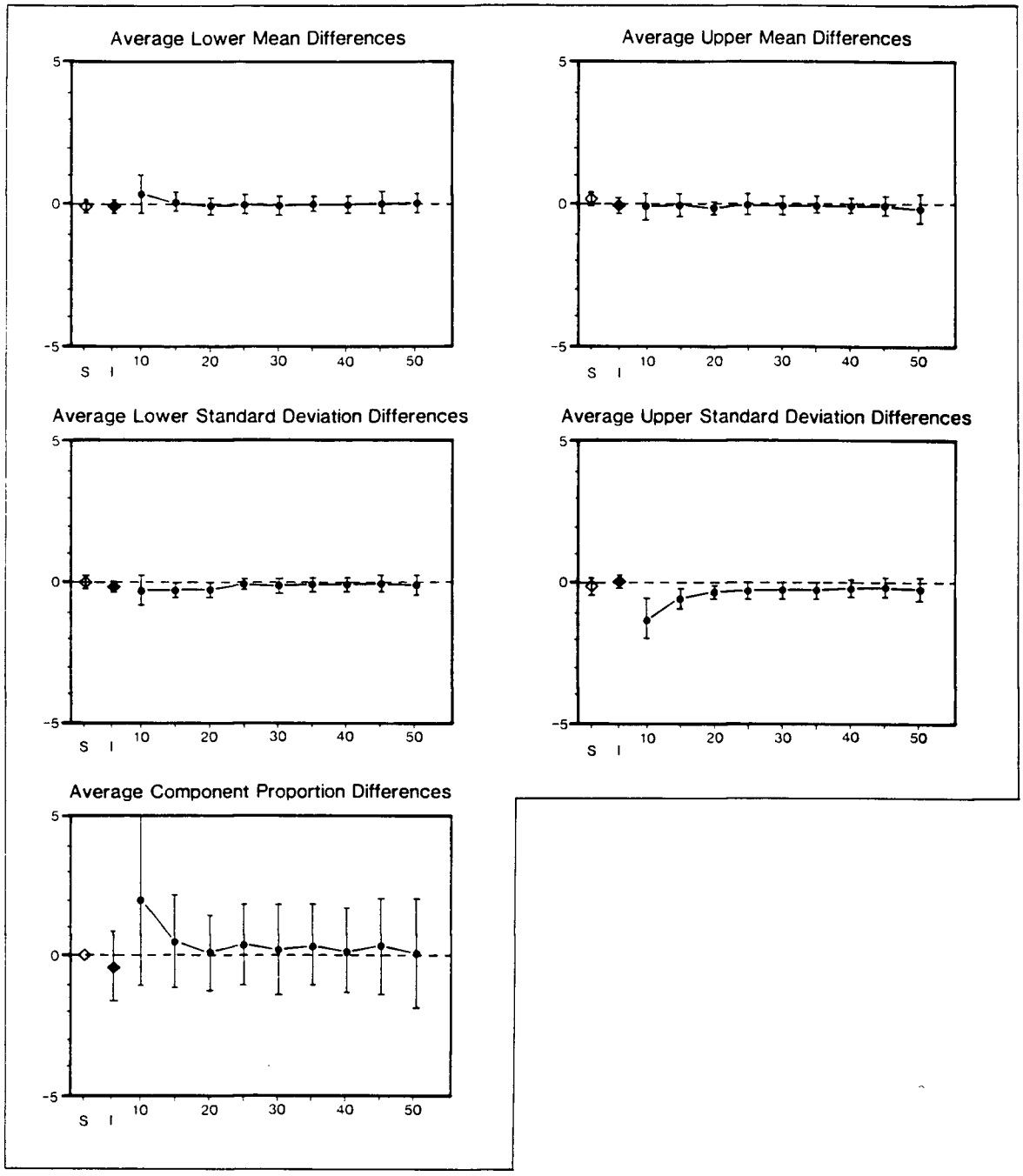


Figure A.27: Parameter Bias of the *RDML* Parameter Estimates, the *CIDML* Parameter Estimates, and the Stochastically Generated Sample Parameter Estimates Relative to the True Population Parameter Estimates for Data Set Structure # 3

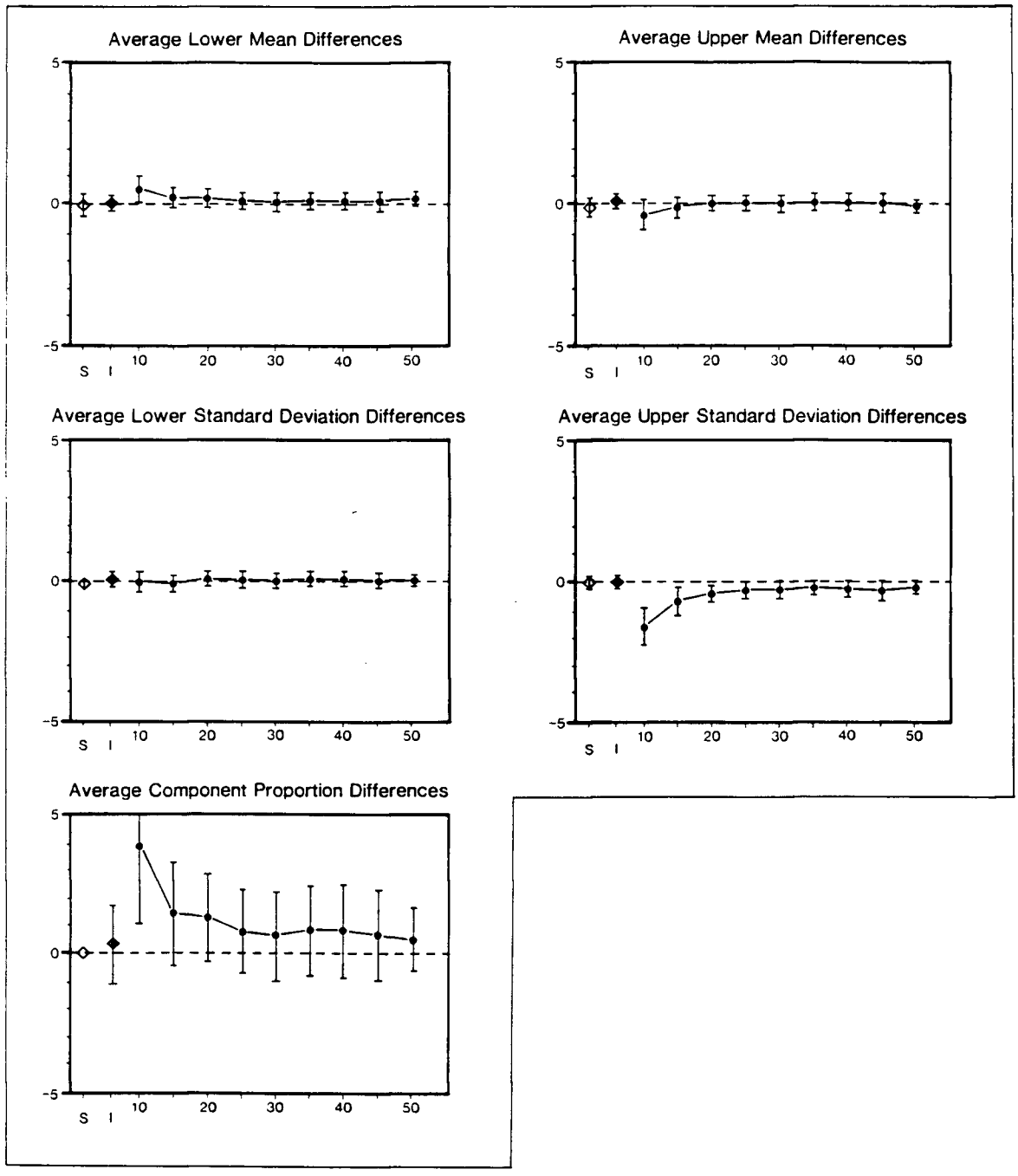


Figure A.28: Parameter Bias of the *RDML* Parameter Estimates, the *CIDML* Parameter Estimates, and the Stochastically Generated Sample Parameter Estimates Relative to the True Population Parameter Estimates for Data Set Structure # 4

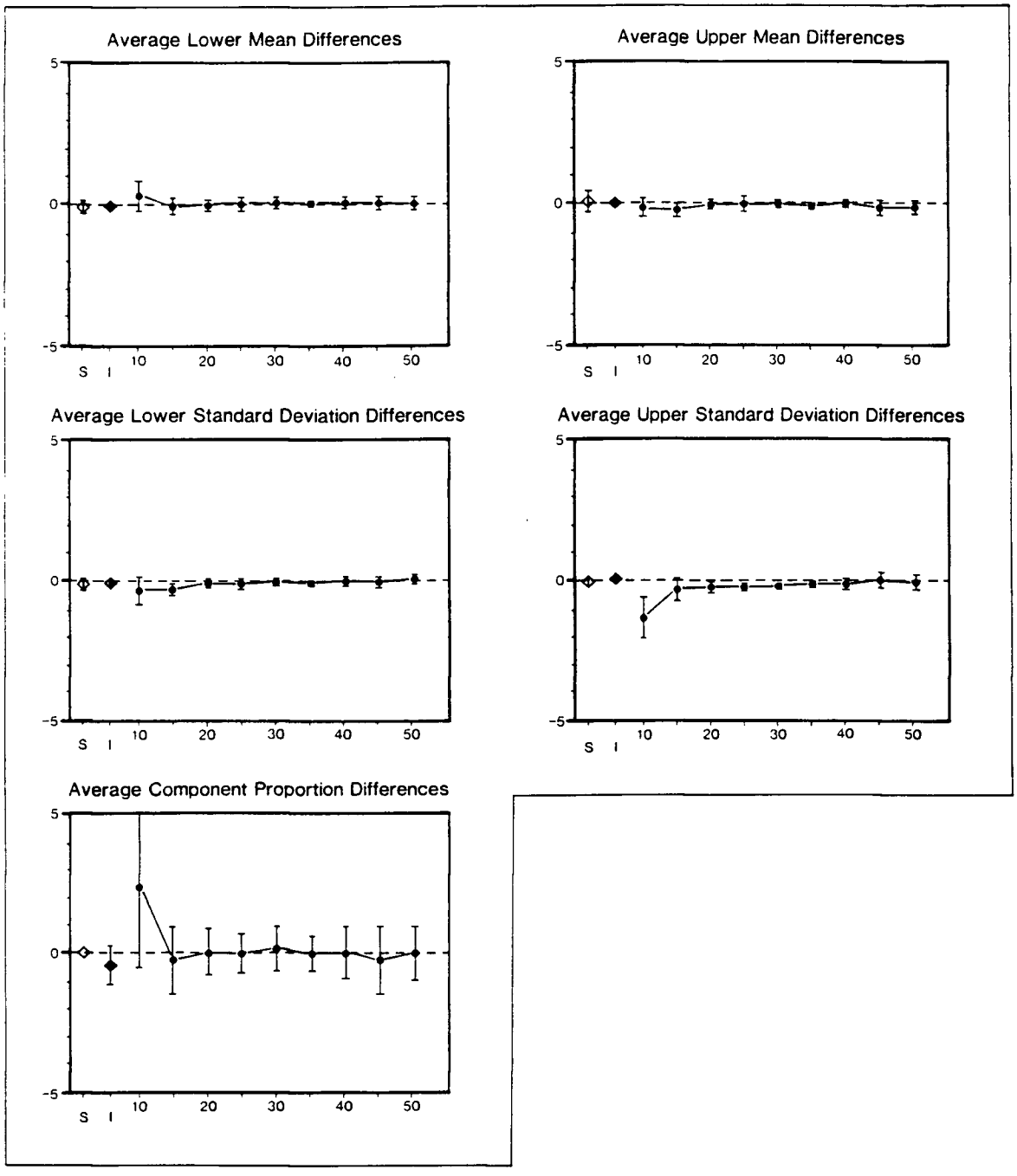


Figure A.29: Parameter Bias of the *RDML* Parameter Estimates, the *CIDML* Parameter Estimates, and the Stochastically Generated Sample Parameter Estimates Relative to the True Population Parameter Estimates for Data Set Structure # 5

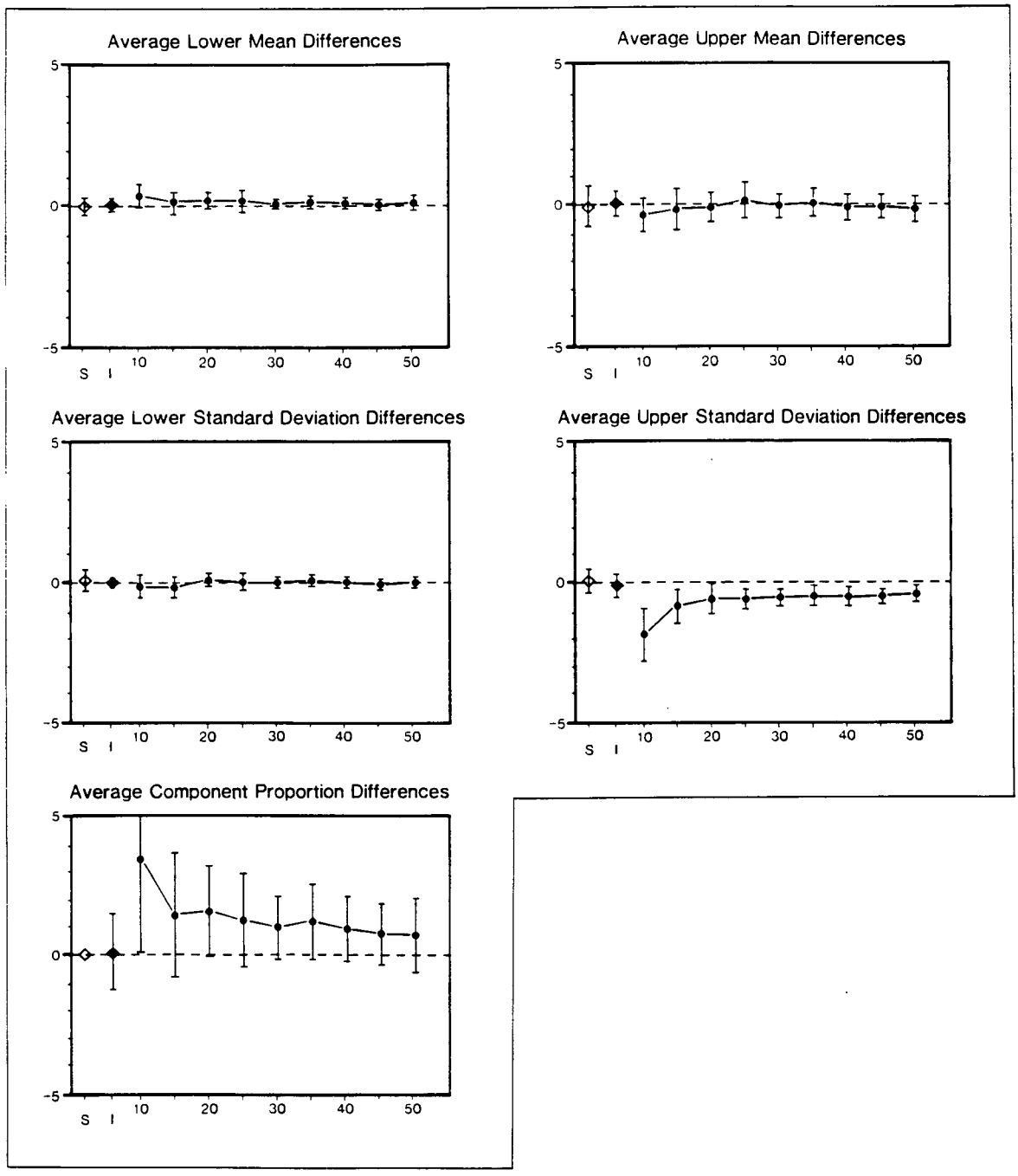


Figure A.30: Parameter Bias of the *RDML* Parameter Estimates, the *CIDML* Parameter Estimates, and the Stochastically Generated Sample Parameter Estimates Relative to the True Population Parameter Estimates for Data Set Structure # 6

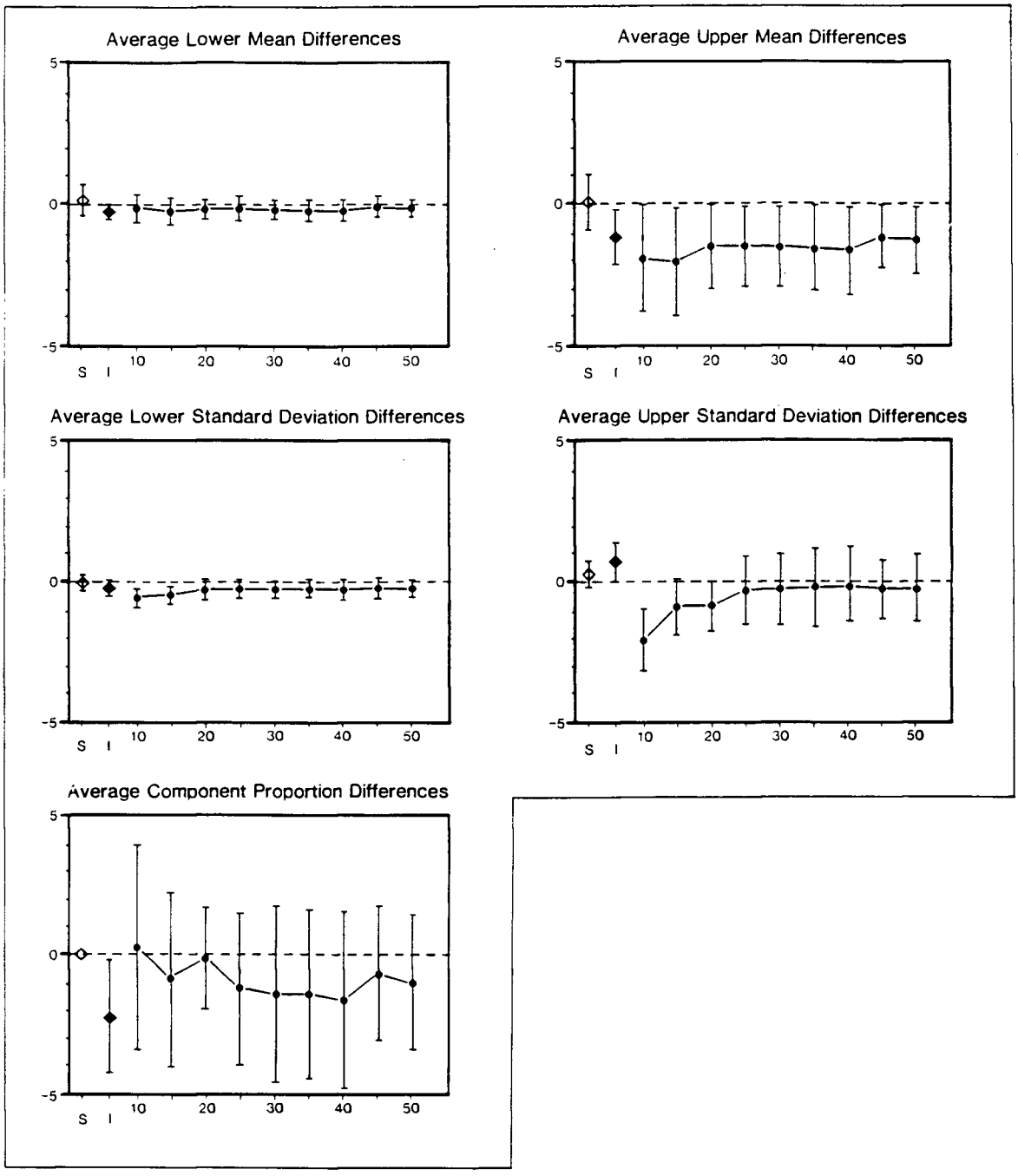


Figure A.31: Parameter Bias of the *RDML* Parameter Estimates, the *CIDML* Parameter Estimates, and the Stochastically Generated Sample Parameter Estimates Relative to the True Population Parameter Estimates for Data Set Structure # 7

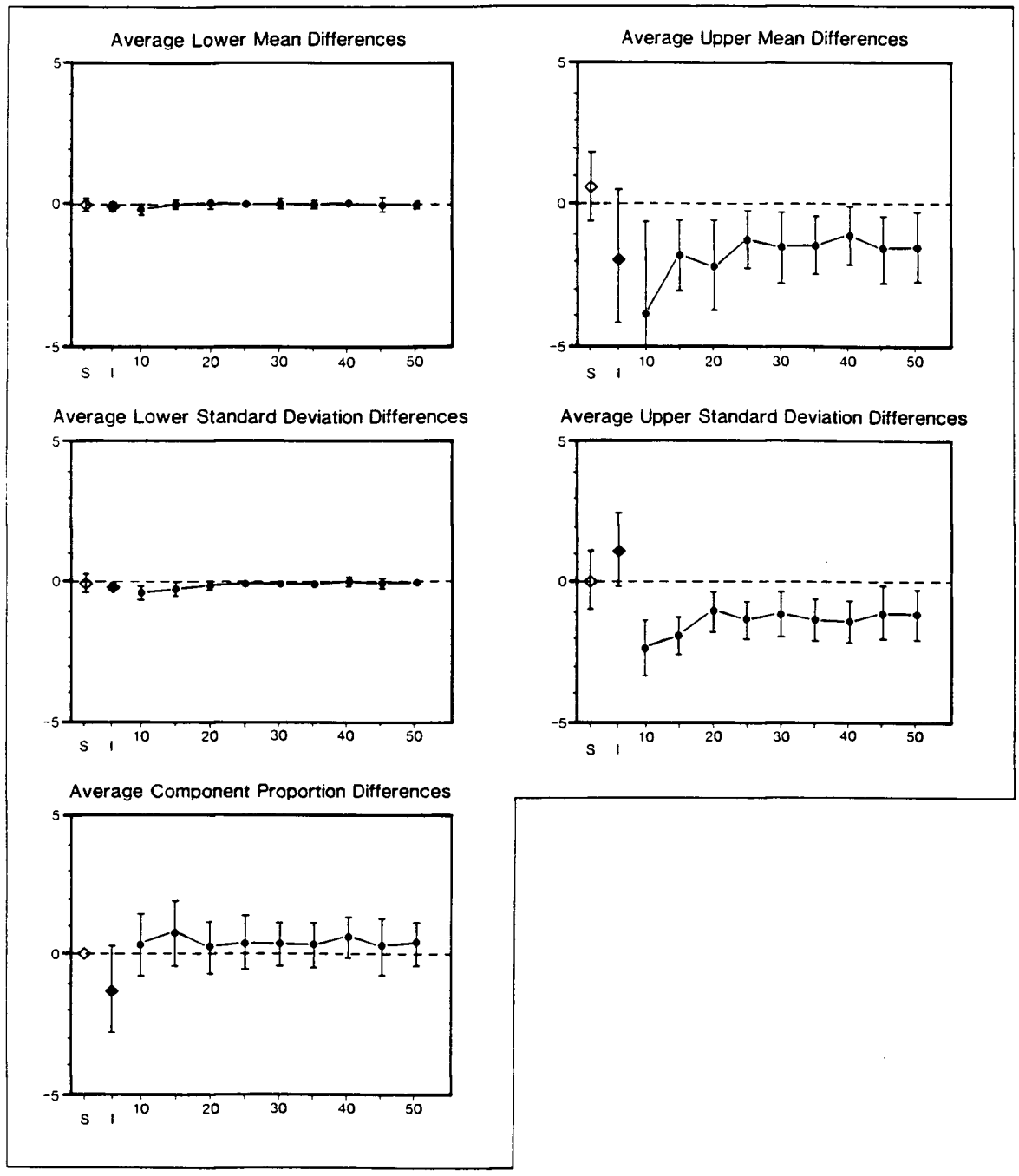


Figure A.32: Parameter Bias of the *RDML* Parameter Estimates, the *CIDML* Parameter Estimates, and the Stochastically Generated Sample Parameter Estimates Relative to the True Population Parameter Estimates for Data Set Structure # 8

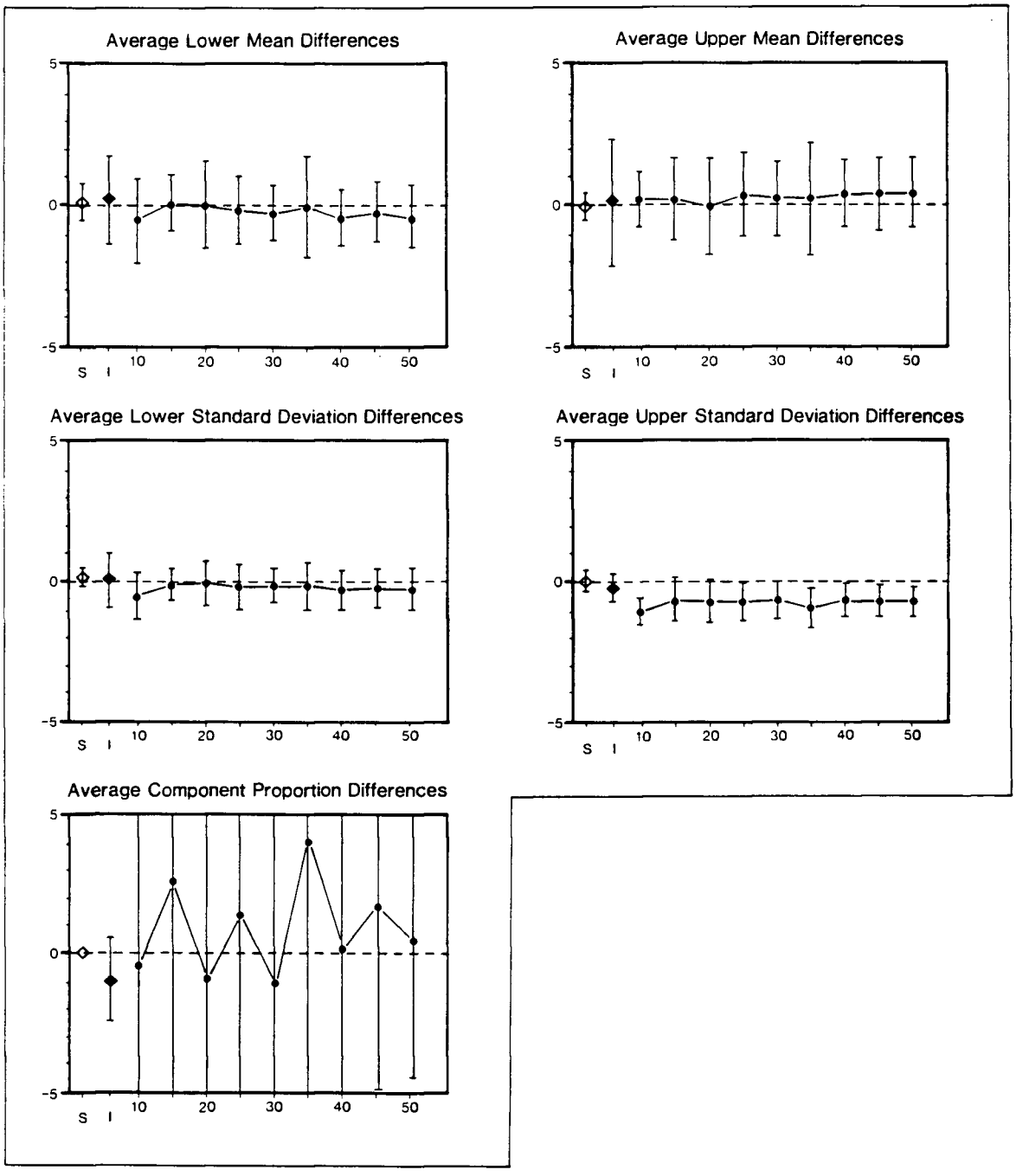


Figure A.33: Parameter Bias of the *RDML* Parameter Estimates, the *CIDML* Parameter Estimates, and the Stochastically Generated Sample Parameter Estimates Relative to the True Population Parameter Estimates for Data Set Structure # 9

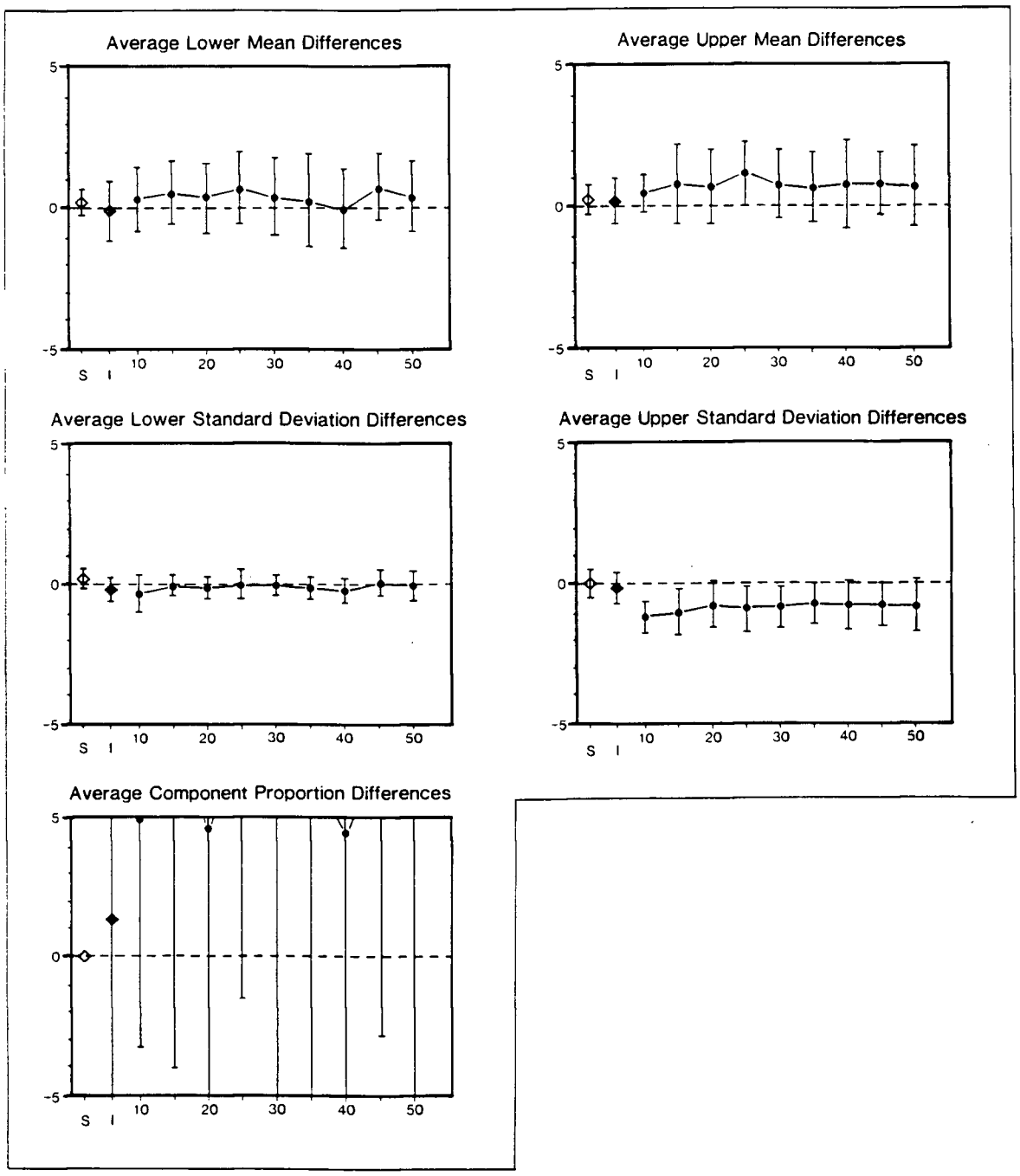


Figure A.34: Parameter Bias of the *RDML* Parameter Estimates, the *CIDML* Parameter Estimates, and the Stochastically Generated Sample Parameter Estimates Relative to the True Population Parameter Estimates for Data Set Structure # 10

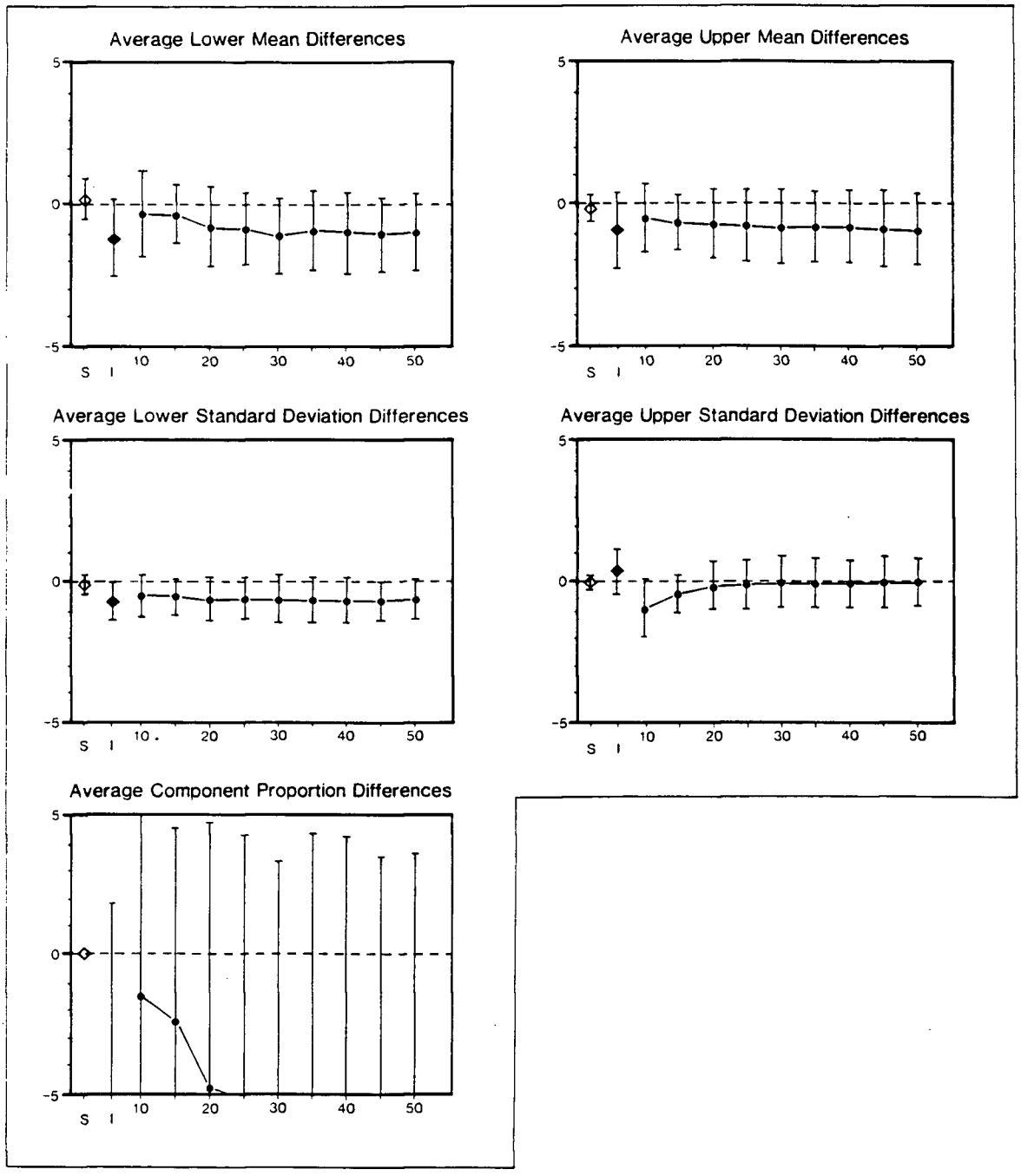


Figure A.35: Parameter Bias of the *RDML* Parameter Estimates, the *CIDML* Parameter Estimates, and the Stochastically Generated Sample Parameter Estimates Relative to the True Population Parameter Estimates for Data Set Structure # 11

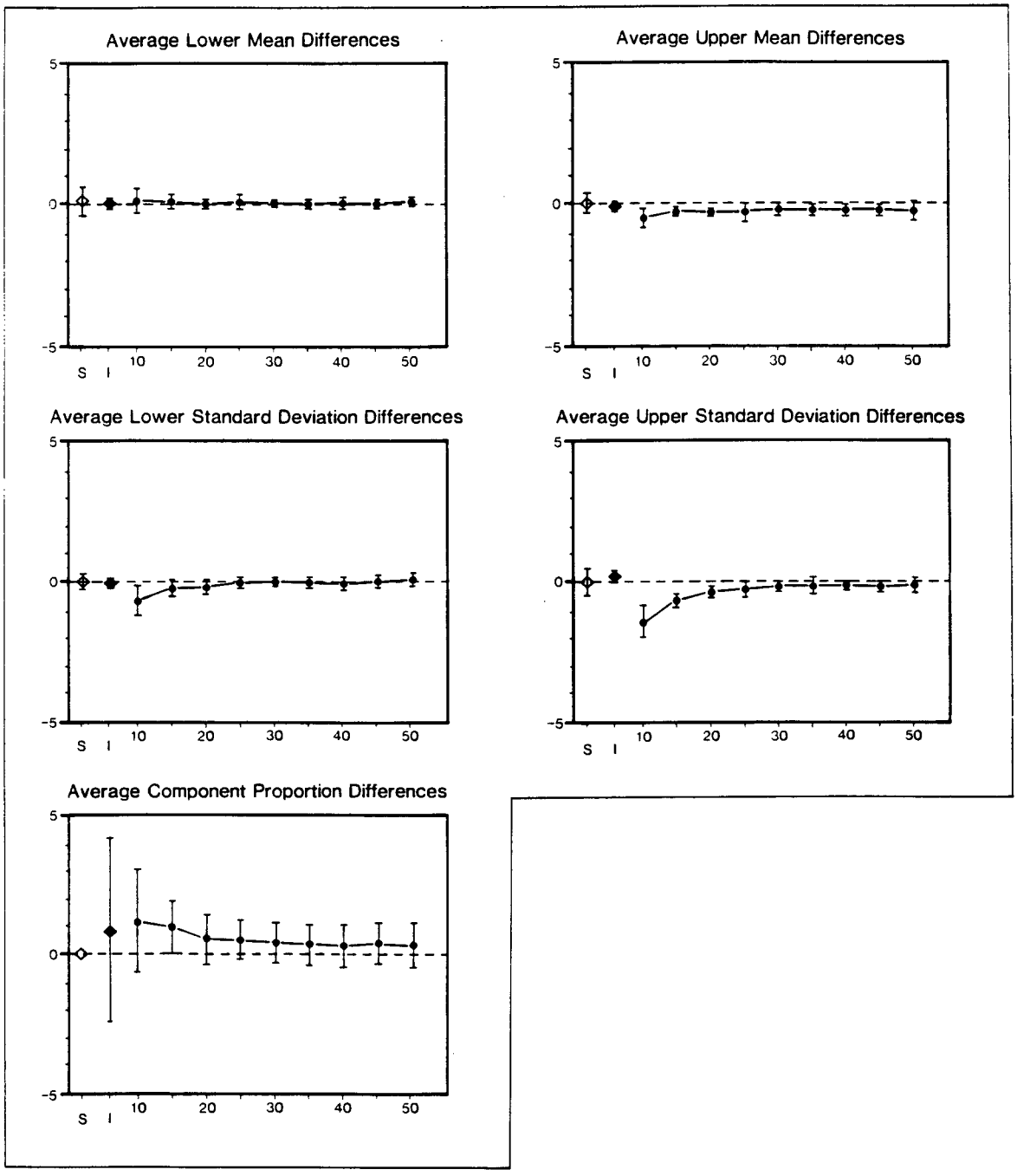


Figure A.36: Parameter Bias of the *RDML* Parameter Estimates, the *CIDML* Parameter Estimates, and the Stochastically Generated Sample Parameter Estimates Relative to the True Population Parameter Estimates for Data Set Structure # 12

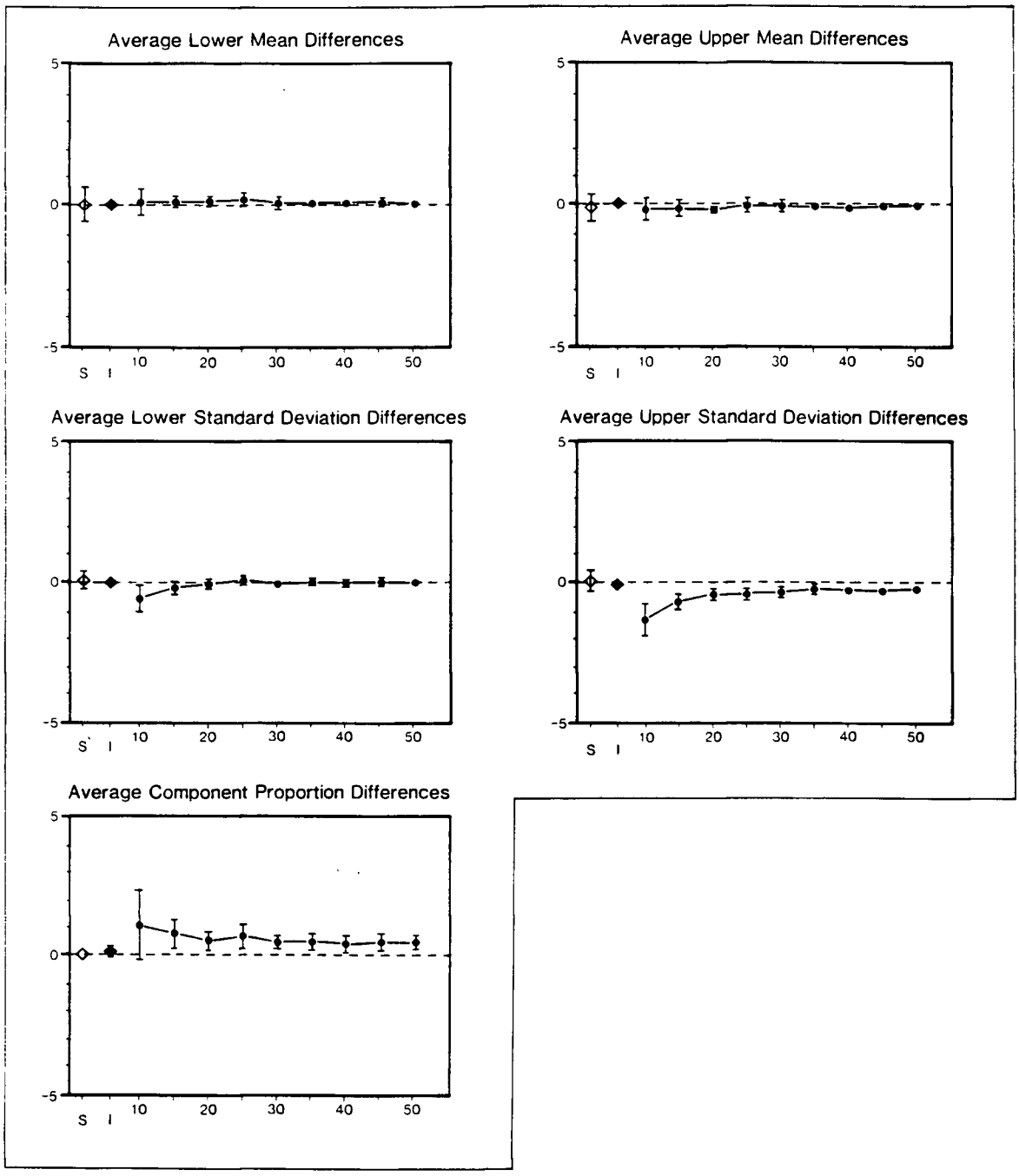


Figure A.37: Parameter Bias of the *RDML* Parameter Estimates, the *CIDML* Parameter Estimates, and the Stochastically Generated Sample Parameter Estimates Relative to the True Population Parameter Estimates for Data Set Structure # 13

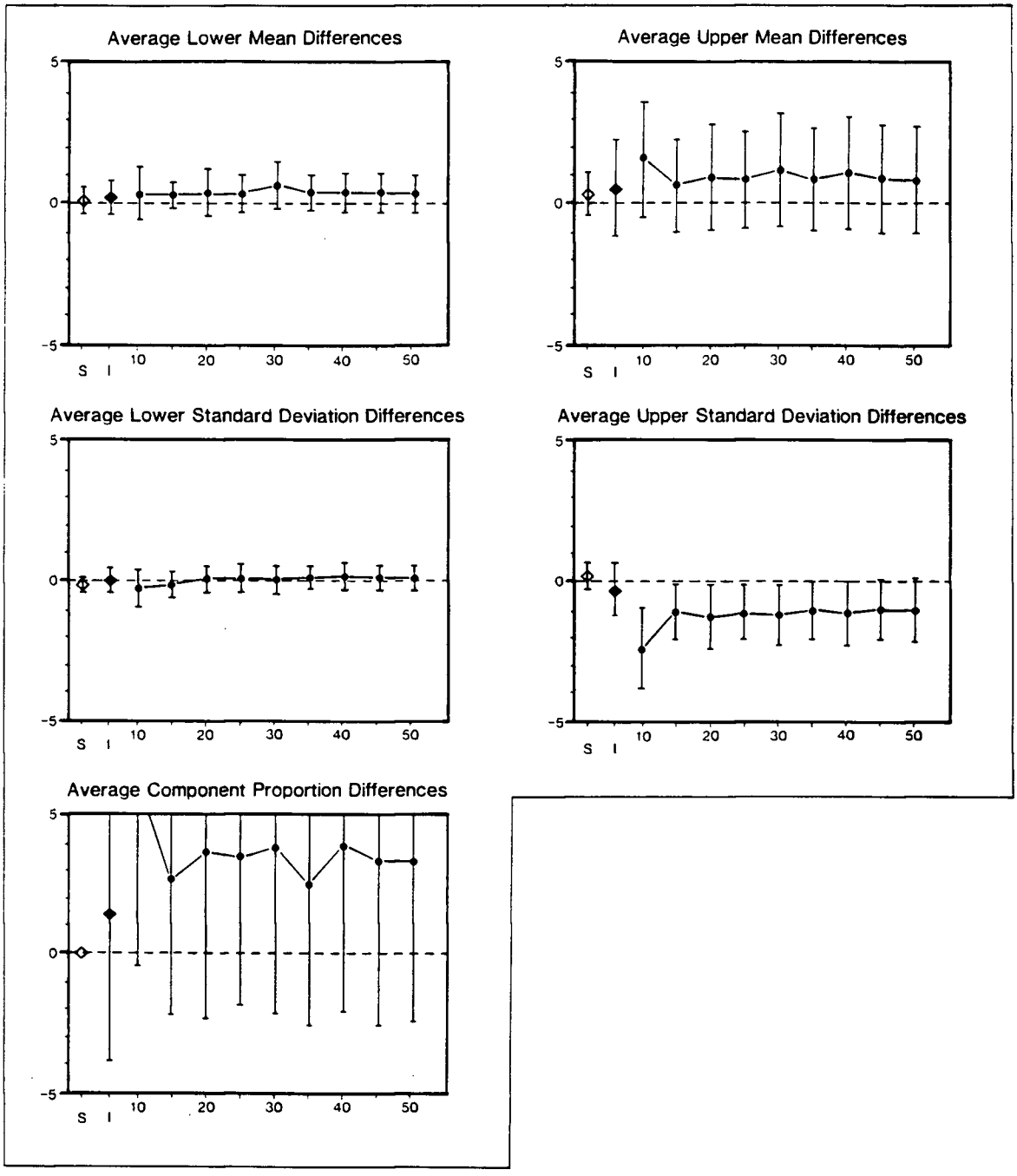


Figure A.38: Parameter Bias of the *RDML* Parameter Estimates, the *CIDML* Parameter Estimates, and the Stochastically Generated Sample Parameter Estimates Relative to the True Population Parameter Estimates for Data Set Structure # 14

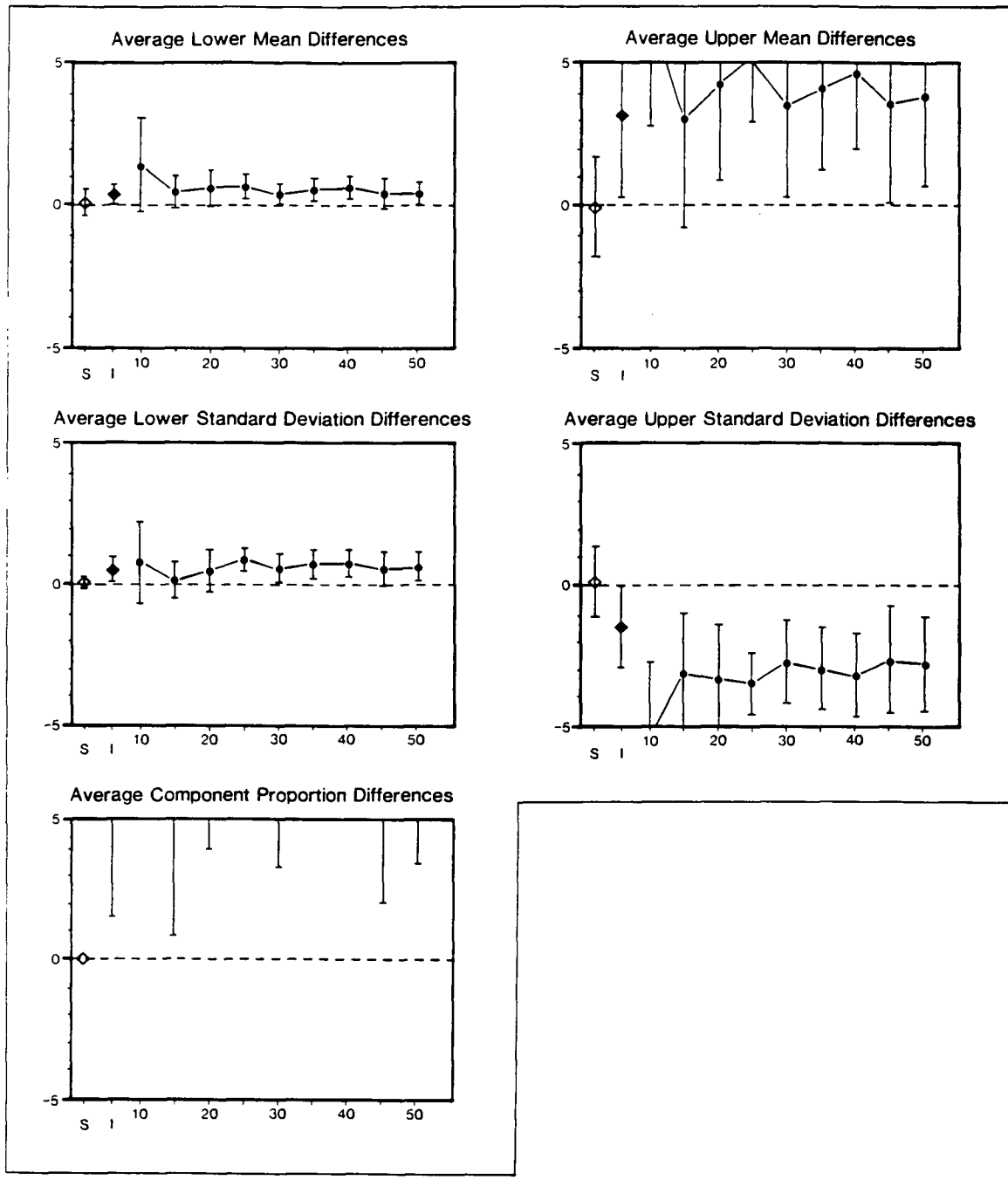


Figure A.39: Parameter Bias of the *RDML* Parameter Estimates, the *CIDML* Parameter Estimates, and the Stochastically Generated Sample Parameter Estimates Relative to the True Population Parameter Estimates for Data Set Structure # 15

A.3 Proportionality of Parameter Estimate Variances and the Number of Observations

The following tables present the average product of the standard deviations of the *CIDML* parameter estimates $\times\sqrt{n}$, where n is the size of the data set.

Table A.31: Comparison of Proportionality of *CIDML* Parameter Estimate Variances With the Number of Observations for 10, 15 and 20 Class Intervals

n	$\hat{s}_{\mu_1}\sqrt{n}$	$\hat{s}_{\mu_2}\sqrt{n}$	$\hat{s}_{\sigma_1}\sqrt{n}$	$\hat{s}_{\sigma_2}\sqrt{n}$	$\hat{s}_{\varpi(\%)}\sqrt{n}$
ℓ 10					
100	15.500	13.200	13.300	12.300	73.300
200	7.212	8.768	6.930	10.324	39.457
300	11.085	8.314	9.353	11.951	53.347
400	9.600	10.200	7.800	13.600	56.400
500	12.075	7.603	10.957	15.876	64.622
ℓ 15					
100	17.800	17.600	13.500	10.300	77.400
200	6.223	7.778	6.081	6.081	27.011
300	5.716	6.755	4.676	5.890	28.579
400	7.400	7.600	6.200	9.800	36.600
500	7.155	5.367	3.801	8.721	26.609
ℓ 20					
100	16.700	15.900	13.200	10.000	80.200
200	5.374	7.071	6.364	5.233	27.436
300	5.196	4.503	4.677	4.330	21.304
400	6.800	5.400	5.200	6.000	31.600
500	4.919	3.801	3.578	4.696	18.783

Table A.32: Comparison of Proportionality of *CIDML* Parameter Estimate Variances With the Number of Observations for 25, 30 and 35 Class Intervals

n	$\hat{s}_{\mu_1}\sqrt{n}$	$\hat{s}_{\mu_2}\sqrt{n}$	$\hat{s}_{\sigma_1}\sqrt{n}$	$\hat{s}_{\sigma_2}\sqrt{n}$	$\hat{s}_{\varpi(\%)}\sqrt{n}$
ℓ 25					
100	7.000	6.000	7.300	3.500	29.200
200	8.202	6.505	6.223	4.950	28.567
300	5.890	6.235	3.811	5.369	24.941
400	6.000	5.400	5.800	6.200	30.600
500	5.367	6.485	4.696	4.249	15.652
ℓ 30					
100	5.300	6.100	6.300	3.600	24.600
200	7.495	6.505	6.647	6.223	32.810
300	5.543	5.890	5.023	4.850	27.713
400	6.400	6.200	5.800	6.600	32.600
500	4.025	2.907	3.130	4.025	17.889
ℓ 35					
100	6.000	5.700	6.400	3.600	27.900
200	6.647	5.515	5.515	5.091	27.294
300	4.503	5.369	4.330	5.369	24.595
400	6.400	6.600	5.800	5.600	32.000
500	2.907	3.354	2.460	3.130	14.087

Table A.33: Comparison of Proportionality of *CIDML* Parameter Estimate Variances With the Number of Observations for 40, 45 and 50 Class Intervals

n	$\hat{s}_{\mu_1}\sqrt{n}$	$\hat{s}_{\mu_2}\sqrt{n}$	$\hat{s}_{\sigma_1}\sqrt{n}$	$\hat{s}_{\sigma_2}\sqrt{n}$	$\hat{s}_{\varpi(\%)}\sqrt{n}$
			ℓ 40		
100	8.000	6.700	7.200	4.200	35.700
200	9.475	6.081	6.930	5.940	32.385
300	5.369	4.330	4.330	5.543	26.327
400	6.200	6.200	5.800	6.000	13.600
500	4.696	3.578	3.130	4.249	21.019
			ℓ 45		
100	6.600	6.400	6.200	4.800	33.000
200	7.495	6.930	6.081	6.505	31.961
300	6.235	6.235	4.503	5.890	29.272
400	6.600	6.800	5.200	7.400	31.800
500	6.037	6.037	4.472	6.037	26.833
			ℓ 50		
100	6.200	5.400	6.400	4.300	28.000
200	4.384	4.101	3.394	6.930	28.284
300	6.062	8.660	5.890	6.755	33.948
400	5.600	4.800	4.000	4.800	25.000
500	5.814	5.367	3.081	6.037	21.913

Appendix B

Likelihood Function Hessian Matrices

"Science bestowed immense new powers on man and at the same time created conditions which were largely beyond his comprehension and still more beyond his control."

Sir Winston Churchill (1949)

The second partial derivative (Hessian) matrix of the logarithm of a likelihood function with respect to the parameters evaluated at the maximum likelihood *ML* solution can be used to estimate the asymptotic variances of the ML estimators of these parameters at the solution :

$$\tilde{H}(\hat{\Psi}_{kl}) = \left[\frac{\partial^2 \ell}{\partial\psi_k \partial\psi_l}\right]_{\Psi=\hat{\Psi}}, \tag{B.129}$$

where $\tilde{H}$ is the Hessian (or second partial derivative) matrix. Specifically :

$$VAR(\hat{\theta}_n) \approx \left[-\ell_n''(\hat{\theta}_n)\right]^{-1}. \tag{B.130}$$

The negative of the Hessian matrix is the observed Fisher information matrix, the inverse of the covariance matrix of the asymptotic parameter estimates. Thus, an estimate of the asymptotic covariance matrix can be calculated from the second partial derivative matrix (Hessian) using the following formula :

$$\tilde{\Sigma} \cong \tilde{I}^{-1} = [-\tilde{H}(\hat{\Psi})]^{-1}. \tag{B.131}$$

where $\tilde{I}$ is the Fisher information matrix.

Two approaches with different likelihood functions are commonly taken. The first utilizes a likelihood function for the raw data and the second involves a likelihood function for the data after it has been cumulated into class intervals. The terms required to determine the entries for both of these likelihood function Hessian matrices are derived below.

B.1 Derivation of the Hessian Matrix for the Raw Data Maximum Likelihood Function for a Mixture of Two Normal Distributions

The probability density function for a mixture of two normal distributions is :

$$p(x|\Psi) = \left(\frac{\varpi}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2}\right) + \left(\frac{(1-\varpi)}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2}\right), \tag{B.132}$$

where $\Psi = (\mu_1, \sigma_1, \mu_2, \sigma_2, \varpi)$. The parameters μ_1, σ_1, μ_2 and σ_2 are the first and second normal distribution means and standard deviations respectively, and ϖ is the percentage of the first component normal distribution. The *RDML* function for this probability density function is :

$$L = \prod_{i=1}^{n} p(x_i|\Psi). \tag{B.133}$$

Taking the natural logarithm gives :

$$\ell = \ln L = \sum_{i=1}^{n} \ln p(x_i|\Psi), \tag{B.134}$$

or :

$$\ell = \ln L = \sum_{i=1}^{n} \ln\left[\left(\frac{\varpi}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\left(\frac{x_i-\mu_1}{\sigma_1}\right)^2}\right) + \left(\frac{(1-\varpi)}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2}\left(\frac{x_i-\mu_2}{\sigma_2}\right)^2}\right)\right]. \tag{B.135}$$

This can be described in the more compact form :

$$\ell = \ln L = \sum_{i=1}^{n} \ln\left[\left(\frac{\varpi}{\sigma_1}\right)\phi(z_{i,1}) + \left(\frac{1-\varpi}{\sigma_2}\right)\phi(z_{i,2})\right], \tag{B.136}$$

where :

$$z_{i,k} = \left(\frac{x_i - \mu_k}{\sigma_k}\right), \tag{B.137}$$

and :

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}. \tag{B.138}$$

Taking the first and second partial derivatives of the log-likelihood function with respect to the 5 parameters (μ_1, σ_1, ϖ, μ_2, σ_2) gives :

$$\frac{\partial \ell}{\partial \psi_k} = \sum_{i=1}^{n} \frac{\partial \ln p(x_i|\Psi)}{\partial \psi_k} = \sum_{i=1}^{n} \frac{1}{p(x_i|\Psi)} \frac{\partial (p(x_i|\Psi))}{\partial \psi_k}, \tag{B.139}$$

and :

$$\frac{\partial^2 \ell}{\partial \psi_k \partial \psi_l} = \sum_{i=1}^{n} \frac{\partial^2 \ln p(x_i|\Psi)}{\partial \psi_k \partial \psi_l} = \sum_{i=1}^{n} \frac{\partial}{\partial \psi_l} \left(\frac{1}{p(x_i|\Psi)} \frac{\partial (p(x_i|\Psi))}{\partial \psi_k}\right), \tag{B.140}$$

which, by the quotient rule, reduces to :

$$\frac{\partial^2 \ell}{\partial \psi_k \partial \psi_l} = \sum_{i=1}^{n} \left(\frac{1}{p(x_i|\Psi)} \frac{\partial^2 (p(x_i|\Psi))}{\partial \psi_k \partial \psi_l} - \frac{1}{p(x_i|\Psi)^2} \left(\frac{\partial (p(x_i|\Psi))}{\partial \psi_k}\right) \left(\frac{\partial (p(x_i|\Psi))}{\partial \psi_l}\right) \right). \tag{B.141}$$

This summation is used to calculate the second partial derivatives of the natural logarithm of the *RDML* function at the *RDML* solution with respect to the parameters. The individual first derivative terms consist of :

$$\frac{\partial p(x_i|\Psi)}{\partial \mu_1} = \left(\frac{\varpi}{\sigma_1^2}\right)(z_{i,1})\phi(z_{i,1}), \tag{B.142}$$

$$\frac{\partial p(x_i|\Psi)}{\partial \sigma_1} = \left(\frac{\varpi}{\sigma_1^2}\right)\phi(z_{i,1})(z_{i,1}^2 - 1), \tag{B.143}$$

$$\frac{\partial p(x_i|\Psi)}{\partial \varpi} = \frac{1}{\sigma_1}\phi(z_{i,1}) - \frac{1}{\sigma_2}\phi(z_{i,2}), \tag{B.144}$$

$$\frac{\partial p(x_i|\Psi)}{\partial \mu_2} = \left(\frac{(1-\varpi)}{\sigma_2^2}\right)(z_{i,2})\phi(z_{i,2}), \tag{B.145}$$

$$\frac{\partial p(x_i|\Psi)}{\partial \sigma_2} = \left(\frac{(1-\varpi)}{\sigma_2^2}\right)\phi(z_{i,2})(z_{i,2}^2 - 1). \tag{B.146}$$

The individual second derivative terms consist of :

$$\frac{\partial^2 p(x_i|\Psi)}{\partial \mu_1^2} = (\frac{\varpi}{\sigma_1^3})\phi(z_{i,1})(z_{i,1}^2 - 1), \tag{B.147}$$

$$\frac{\partial^2 p(x_i|\Psi)}{\partial \mu_2^2} = (\frac{(1-\varpi)}{\sigma_2^3})\phi(z_{i,2})(z_{i,2}^2 - 1), \tag{B.148}$$

$$\frac{\partial^2 p(x_i|\Psi)}{\partial \mu_1 \partial \mu_2} = 0, \tag{B.149}$$

$$\frac{\partial^2 p(x_i|\Psi)}{\partial \sigma_1^2} = (\frac{\varpi}{\sigma_1^3})\phi(z_{i,1})(z_{i,1}^4 - 5z_{i,1}^2 + 2), \tag{B.150}$$

$$\frac{\partial^2 p(x_i|\Psi)}{\partial \sigma_2^2} = (\frac{(1-\varpi)}{\sigma_2^3})\phi(z_{i,2})(z_{i,2}^4 - 5z_{i,2}^2 + 2), \tag{B.151}$$

$$\frac{\partial^2 p(x_i|\Psi)}{\partial \sigma_1 \partial \sigma_2} = 0, \tag{B.152}$$

$$\frac{\partial^2 p(x_i|\Psi)}{\partial \mu_1 \partial \sigma_1} = (\frac{\varpi}{\sigma_1^3})\phi(z_{i,1})(z_{i,1}^3 - 3z_{i,1}), \tag{B.153}$$

$$\frac{\partial^2 p(x_i|\Psi)}{\partial \mu_2 \partial \sigma_2} = (\frac{(1-\varpi)}{\sigma_2^3})\phi(z_{i,2})(z_{i,2}^3 - 3z_{i,2}), \tag{B.154}$$

$$\frac{\partial^2 p(x_i|\Psi)}{\partial \mu_1 \partial \sigma_2} = 0, \tag{B.155}$$

$$\frac{\partial^2 p(x_i|\Psi)}{\partial \mu_2 \partial \sigma_1} = 0, \tag{B.156}$$

$$\frac{\partial^2 p(x_i|\Psi)}{\partial \mu_1 \partial \varpi} = (\frac{1}{\sigma_1^2})(z_{i,1})\phi(z_{i,1}), \tag{B.157}$$

$$\frac{\partial^2 p(x_i|\Psi)}{\partial \sigma_1 \partial \varpi} = (\frac{1}{\sigma_1^2})\phi(z_{i,1})(z_{i,1}^2 - 1), \tag{B.158}$$

$$\frac{\partial^2 p(x_i|\Psi)}{\partial \varpi^2} = 0, \tag{B.159}$$

$$\frac{\partial^2 p(x_i|\Psi)}{\partial \mu_2 \partial \varpi} = (\frac{-1}{\sigma_2^2})(z_{i,2})\phi(z_{i,2}), \tag{B.160}$$

$$\frac{\partial^2 p(x_i|\Psi)}{\partial \sigma_2 \partial \varpi} = (\frac{-1}{\sigma_2^2})\phi(z_{i,2})(z_{i,2}^2 - 1). \tag{B.161}$$

These partial derivatives can then be used to calculate the terms of the Hessian matrix of the natural logarithm of the *RDML* function evaluated at the *RDML* parameter estimates (Behboodian 1972).

B.2 Derivation of the Hessian Matrix for the Class Interval Maximum Likelihood Function for a Mixture of Two Normal Distributions

The likelihood function for the class interval data from a mixture of two normal distributions is :

$$L = \prod_{j=1}^{m} \Big(P(b_{j+1}|\Psi) - P(b_j|\Psi) \Big)^{n_j}, \tag{B.162}$$

where b_j and b_{j+1} are the lower and upper class interval limits, respectively, the n_j are the number of cases which fall in the j'th class interval and :

$$P(x|\Psi) = \int_{-\infty}^{x} p(u|\Psi) du. \tag{B.163}$$

Taking the natural logarithm of the likelihood function gives :

$$\ell = \ln L = \sum_{j=1}^{m} n_j \ln \Big[P(b_{j+1}|\Psi) - P(b_j|\Psi)) \Big]. \tag{B.164}$$

This can be further expanded to a more specific equation expressed in terms of all of the parameters :

$$\begin{aligned} \ell &= \ln L \\ &= \sum_{j=1}^{m} n_j \left(\ln \left[\left(\frac{\varpi}{\sqrt{2\pi}\sigma_1} \int_{-\infty}^{b_{j+1}} e^{-\frac{1}{2}\left(\frac{u-\mu_1}{\sigma_1}\right)^2} du + \frac{(1-\varpi)}{\sqrt{2\pi}\sigma_2} \int_{-\infty}^{b_{j+1}} e^{-\frac{1}{2}\left(\frac{u-\mu_2}{\sigma_2}\right)^2} du \right) \right. \right. \\ &- \left. \left. \left(\frac{\varpi}{\sqrt{2\pi}\sigma_1} \int_{-\infty}^{b_j} e^{-\frac{1}{2}\left(\frac{u-\mu_1}{\sigma_1}\right)^2} du + \frac{(1-\varpi)}{\sqrt{2\pi}\sigma_2} \int_{-\infty}^{b_j} e^{-\frac{1}{2}\left(\frac{u-\mu_2}{\sigma_2}\right)^2} du \right) \right] \right). \end{aligned} \tag{B.165}$$

In more compact form this becomes :

$$\begin{aligned} \ell &= \ln L \\ &= \sum_{j=1}^{m} n_j \ln \Big[\Big(\varpi \Phi(z_{j+1,1}) + (1-\varpi) \Phi(z_{j+1,2}) \Big) - \\ &\quad \Big(\varpi \Phi(z_{j,1}) + (1-\varpi) \Phi(z_{j,2}) \Big) \Big], \end{aligned} \tag{B.166}$$

where :

$$\Phi(z) = \int_{-\infty}^{z} \phi(u) du, \tag{B.167}$$

and :

$$z_{j,k} = \Big(\frac{b_j - \mu_k}{\sigma_k}\Big). \tag{B.168}$$

Taking the first and second partial derivatives of the *CIDML* function with respect to the 5 parameters $(\mu_1, \sigma_1, \varpi, \mu_2, \sigma_2)$ we get :

$$\begin{aligned}\frac{\partial \ell}{\partial \psi_k} &= \sum_{j=1}^{m} n_j \frac{\partial \ln\Big[P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big]}{\partial \psi_k} \\ &= \sum_{j=1}^{m} \frac{n_j}{\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)} \frac{\partial\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial \psi_k},\end{aligned} \tag{B.169}$$

and :

$$\frac{\partial^2 \ell}{\partial \psi_k \partial \psi_l} = \sum_{j=1}^{m} n_j \frac{\partial^2 \ln\Big[P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big]}{\partial \psi_k \partial \psi_l}$$

$$= \sum_{j=1}^{m} n_j \frac{\partial}{\partial \psi_l}\Bigg(\Big(\frac{1}{P(b_{j+1}|\Psi) - P(b_j|\Psi)}\Big)\frac{\partial}{\partial \psi_k}\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)\Bigg), \tag{B.170}$$

which, by the quotient rule, reduces to :

$$\frac{\partial^2 \ell}{\partial \psi_k \partial \psi_l} = \sum_{j=1}^{m} n_j \Bigg(\frac{1}{P(b_{j+1}|\Psi) - P(b_j|\Psi)} \frac{\partial^2\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial \psi_k \partial \psi_l} -$$

$$\frac{1}{\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)^2} \frac{\partial\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial \psi_k} \frac{\partial\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial \psi_l}\Bigg). \tag{B.171}$$

This summation is used to calculate the second partial derivatives of the natural logarithm of the *CIDML* function at the *CIDML* parameter estimates. The individual first derivative terms consist of :

$$\frac{\partial\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial \mu_1} = \frac{-\varpi}{\sigma_1}\Big(\phi(z_{j+1,1}|\theta_1) - \phi(z_{j,1}|\theta_1)\Big), \tag{B.172}$$

$$\frac{\partial\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial \sigma_1} = \frac{-\varpi}{\sigma_1}\Big((z_{j+1,1})\phi(z_{j+1,1}|\theta_1) - (z_{j,1})\phi(z_{j,1}|\theta_1)\Big), \tag{B.173}$$

$$\frac{\partial\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial \varpi} = \Big(\Phi(z_{j+1,1}|\theta_1) - \Phi(z_{j+1,2}|\theta_2) - \Phi(z_{j,1}|\theta_1) + \Phi(z_{j,2}|\theta_2)\Big), \tag{B.174}$$

$$\frac{\partial\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial\mu_2} = \frac{-(1-\varpi)}{\sigma_2}\Big(\phi(z_{j+1,2}|\theta_2) - \phi(z_{j,2}|\theta_2)\Big), \tag{B.175}$$

$$\frac{\partial\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial\sigma_2} = \frac{-(1-\varpi)}{\sigma_2}\Big((z_{j+1,2})\phi(z_{j+1,2}|\theta_2) - (z_{j,2})\phi(z_{j,2}|\theta_2)\Big). \tag{B.176}$$

The individual second derivative terms consist of :

$$\frac{\partial^2\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial\mu_1^2} = \Big(\frac{-\varpi}{\sigma_1^2}\Big)\Big((z_{j+1,1})\phi(z_{j+1,1}|\theta_1) - (z_{j,1})\phi(z_{j,1}|\theta_1)\Big), \tag{B.177}$$

$$\frac{\partial^2\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial\mu_2^2} = \Big(\frac{-(1-\varpi)}{\sigma_2^2}\Big)\Big((z_{j+1,2})\phi(z_{j+1,2}|\theta_2) - (z_{j,2})\phi(z_{j,2}|\theta_2)\Big), \tag{B.178}$$

$$\frac{\partial^2\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial\mu_1\partial\mu_2} = 0, \tag{B.179}$$

$$\frac{\partial^2\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial\sigma_1^2} = \Big(\frac{-\varpi}{\sigma_1^2}\Big)\Big((z_{j+1,1}^3 - 2z_{j+1,1})\phi(z_{j+1,1}|\theta_1) - (z_{j,1}^3 - 2z_{j,1})\phi(z_{j,1}|\theta_1)\Big), \tag{B.180}$$

$$\frac{\partial^2\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial\sigma_2^2} = \Big(\frac{-(1-\varpi)}{\sigma_2^2}\Big)\Big((z_{j+1,2}^3 - 2z_{j+1,2})\phi(z_{j+1,2}|\theta_2) - (z_{j,2}^3 - 2z_{j,2})\phi(z_{j,2}|\theta_2)\Big), \tag{B.181}$$

$$\frac{\partial^2\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial\sigma_1\partial\sigma_2} = 0, \tag{B.182}$$

$$\frac{\partial^2\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial\mu_1\partial\sigma_1} = \Big(\frac{-\varpi}{\sigma_1^2}\Big)\Big((z_{j+1,1}^2 - 1)\phi(z_{j+1,1}|\theta_1) - (z_{j,1}^2 - 1)\phi(z_{j,1}|\theta_1)\Big), \tag{B.183}$$

$$\frac{\partial^2\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial\mu_2\partial\sigma_2} = \Big(\frac{-(1-\varpi)}{\sigma_2^2}\Big)\Big((z_{j+1,2}^2 - 1)\phi(z_{j+1,2}|\theta_2) - (z_{j,2}^2 - 1)\phi(z_{j,2}|\theta_2)\Big), \tag{B.184}$$

$$\frac{\partial^2\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial\mu_1\partial\sigma_2} = 0, \tag{B.185}$$

$$\frac{\partial^2\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial\mu_2\partial\sigma_1} = 0, \tag{B.186}$$

$$\frac{\partial^2\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial\mu_1\partial\varpi} = \Big(\frac{-1}{\sigma_1}\Big)\Big(\phi(z_{j+1,1}|\theta_1) - \phi(z_{j,1}|\theta_1)\Big), \tag{B.187}$$

$$\frac{\partial^2\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial\sigma_1\partial\varpi} = \Big(\frac{-1}{\sigma_1}\Big)\Big((z_{j+1,1})\phi(z_{j+1,1}|\theta_1) - (z_{j,1})\phi(z_{j,1}|\theta_1)\Big), \tag{B.188}$$

$$\frac{\partial^2\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial\varpi^2} = 0, \tag{B.189}$$

$$\frac{\partial^2\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial\mu_2\partial\varpi} = \Big(\frac{1}{\sigma_2}\Big)\Big(\phi(z_{j+1,2}|\theta_2) - \phi(z_{j,2}|\theta_2)\Big), \tag{B.190}$$

$$\frac{\partial^2\Big(P(b_{j+1}|\Psi) - P(b_j|\Psi)\Big)}{\partial\sigma_2\partial\varpi} = \Big(\frac{1}{\sigma_2}\Big)\Big((z_{j+1,2})\phi(z_{j+1,2}|\theta_2) - (z_{j,2})\phi(z_{j,2}|\theta_2)\Big). \tag{B.191}$$

These terms can then be used to calculate the terms of the Hessian matrix of the natural logarithm of the *CIDML* function evaluated at the *CIDML* parameter estimates.

Appendix C

Asymptotic Correlation Estimates

"Without deviation, progress is not possible."

Frank Zappa (1980)

C.1 Raw Data Maximum Likelihood Function

The following tables contain the mean asymptotic correlations of the *RDML* parameter estimates calculated using the inverse of the negative Hessian matrix (observed Fisher information matrix) evaluated at the *RDML* parameter estimates. Correlation matrices for all data set structures except data set structure # 1 (see Table 3.10) are presented.

(Values in parentheses are the standard deviations of the asymptotic correlations for the 10 data sets of the corresponding parameters.)

Table C.34: Asymptotic Correlation Matrix for *RDML* Parameter Estimates for Data Set Structure # 2

$\mu_1 = 20 \quad \sigma_1 = 5 \quad \varpi(\%) = 50 \quad \mu_2 = 40 \quad \sigma_2 = 5 \quad n = 100$

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.41 (0.02)	1.00 (0.00)			
ϖ	0.30 (0.03)	0.31 (0.02)	1.00 (0.00)		
μ_2	0.36 (0.03)	0.37 (0.02)	0.30 (0.03)	1.00 (0.00)	
σ_2	-0.37 (0.02)	-0.36 (0.02)	-0.32 (0.02)	-0.41 (0.02)	1.00 (0.00)

Table C.35: Asymptotic Correlation Matrix for *RDML* Parameter Estimates for Data Set Structure # 3

$\mu_1 = 20 \quad \sigma_1 = 5 \quad \varpi(\%) = 50 \quad \mu_2 = 40 \quad \sigma_2 = 5 \quad n = 300$

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.36 (0.01)	1.00 (0.00)			
ϖ	0.23 (0.00)	0.26 (0.00)	1.00 (0.00)		
μ_2	0.30 (0.01)	0.32 (0.01)	0.23 (0.00)	1.00 (0.00)	
σ_2	-0.32 (0.00)	-0.33 (0.01)	-0.25 (0.00)	-0.35 (0.00)	1.00 (0.00)

Table C.36: Asymptotic Correlation Matrix for *RDML* Parameter Estimates for Data Set Structure # 4

$\mu_1 = 20 \quad \sigma_1 = 5 \quad \varpi(\%) = 50 \quad \mu_2 = 40 \quad \sigma_2 = 5 \quad n = 400$

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.38 (0.01)	1.00 (0.00)			
ϖ	0.26 (0.01)	0.28 (0.01)	1.00 (0.00)		
μ_2	0.32 (0.01)	0.33 (0.01)	0.26 (0.01)	1.00 (0.00)	
σ_2	-0.33 (0.01)	-0.32 (0.01)	-0.28 (0.01)	-0.38 (0.01)	1.00 (0.00)

Table C.37: Asymptotic Correlation Matrix for *RDML* Parameter Estimates for Data Set Structure # 5

$\mu_1 = 20 \quad \sigma_1 = 5 \quad \varpi(\%) = 50 \quad \mu_2 = 40 \quad \sigma_2 = 5 \quad n = 500$

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.33 (0.01)	1.00 (0.00)			
ϖ	0.21 (0.01)	0.23 (0.00)	1.00 (0.00)		
μ_2	0.27 (0.01)	0.29 (0.00)	0.22 (0.00)	1.00 (0.00)	
σ_2	-0.29 (0.00)	-0.29 (0.00)	-0.24 (0.00)	-0.34 (0.00)	1.00 (0.00)

Table C.38: Asymptotic Correlation Matrix for *RDML* Parameter Estimates for Data Set Structure # 6

$\mu_1 = 20 \quad \sigma_1 = 5 \quad \varpi(\%) = 70 \quad \mu_2 = 40 \quad \sigma_2 = 5 \quad n = 200$

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.33 (0.01)	1.00 (0.00)			
ϖ	0.26 (0.01)	0.29 (0.01)	1.00 (0.00)		
μ_2	0.33 (0.01)	0.35 (0.01)	0.31 (0.02)	1.00 (0.00)	
σ_2	-0.31 (0.01)	-0.31 (0.00)	-0.31 (0.01)	-0.44 (0.01)	1.00 (0.00)

Table C.39: Asymptotic Correlation Matrix for *RDML* Parameter Estimates for Data Set Structure # 7

$\mu_1 = 20 \quad \sigma_1 = 5 \quad \varpi(\%) = 85 \quad \mu_2 = 40 \quad \sigma_2 = 5 \quad n = 200$

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.32 (0.03)	1.00 (0.00)			
ϖ	0.32 (0.04)	0.34 (0.03)	1.00 (0.00)		
μ_2	0.37 (0.04)	0.38 (0.02)	0.46 (0.05)	1.00 (0.00)	
σ_2	-0.33 (0.02)	-0.31 (0.01)	-0.43 (0.04)	-0.56 (0.03)	1.00 (0.00)

Table C.40: Asymptotic Correlation Matrix for *RDML* Parameter Estimates for Data Set Structure # 8

$\mu_1 = 20 \quad \sigma_1 = 5 \quad \varpi(\%) = 95 \quad \mu_2 = 40 \quad \sigma_2 = 5 \quad n = 200$

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.22 (0.01)	1.00 (0.00)			
ϖ	0.29 (0.03)	0.32 (0.02)	1.00 (0.00)		
μ_2	0.32 (0.03)	0.35 (0.02)	0.53 (0.06)	1.00 (0.00)	
σ_2	-0.29 (0.02)	-0.30 (0.01)	-0.50 (0.06)	-0.63 (0.05)	1.00 (0.00)

Table C.41: Asymptotic Correlation Matrix for *RDML* Parameter Estimates for Data Set Structure # 9

$\mu_1 = 20 \quad \sigma_1 = 5 \quad \varpi(\%) = 50 \quad \mu_2 = 25 \quad \sigma_2 = 5 \quad n = 200$

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.20 (0.48)	1.00 (0.00)			
ϖ	0.70 (0.15)	-0.12 (0.85)	1.00 (0.00)		
μ_2	0.27 (0.38)	-0.06 (0.60)	0.71 (0.09)	1.00 (0.00)	
σ_2	-0.39 (0.37)	0.27 (0.41)	-0.73 (0.18)	-0.57 (0.08)	1.00 (0.00)

Table C.42: Asymptotic Correlation Matrix for *RDML* Parameter Estimates for Data Set Structure # 10

$\mu_1 = 20 \quad \sigma_1 = 5 \quad \varpi(\%) = 50 \quad \mu_2 = 30 \quad \sigma_2 = 5 \quad n = 200$

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.89 (0.00)	1.00 (0.00)			
ϖ	0.94 (0.00)	0.86 (0.00)	1.00 (0.00)		
μ_2	0.90 (0.00)	0.79 (0.01)	0.94 (0.00)	1.00 (0.00)	
σ_2	-0.77 (0.01)	-0.64 (0.02)	-0.84 (0.01)	-0.88 (0.00)	1.00 (0.00)

Table C.43: Asymptotic Correlation Matrix for *RDML* Parameter Estimates for Data Set Structure # 11

$\mu_1 = 20 \quad \sigma_1 = 5 \quad \varpi(\%) = 50 \quad \mu_2 = 35 \quad \sigma_2 = 5 \quad n = 200$

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.59 (0.01)	1.00 (0.00)			
ϖ	0.58 (0.02)	0.51 (0.01)	1.00 (0.00)		
μ_2	0.57 (0.01)	0.49 (0.01)	0.60 (0.03)	1.00 (0.00)	
σ_2	-0.50 (0.01)	-0.40 (0.01)	-0.56 (0.02)	-0.61 (0.01)	1.00 (0.00)

Table C.44: Asymptotic Correlation Matrix for *RDML* Parameter Estimates for Data Set Structure # 12

$\mu_1 = 20 \quad \sigma_1 = 5 \quad \varpi(\%) = 50 \quad \mu_2 = 45 \quad \sigma_2 = 5 \quad n = 200$

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.17 (0.00)	1.00 (0.00)			
ϖ	0.07 (0.00)	0.10 (0.00)	1.00 (0.00)		
μ_2	0.12 (0.00)	0.16 (0.00)	0.07 (0.00)	1.00 (0.00)	
σ_2	-0.16 (0.00)	-0.21 (0.00)	-0.10 (0.00)	-0.18 (0.00)	1.00 (0.00)

Table C.45: Asymptotic Correlation Matrix for *RDML* Parameter Estimates for Data Set Structure # 13

$\mu_1 = 20 \quad \sigma_1 = 5 \quad \varpi(\%) = 50 \quad \mu_2 = 50 \quad \sigma_2 = 5 \quad n = 200$

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.07 (0.00)	1.00 (0.00)			
ϖ	0.02 (0.00)	0.03 (0.00)	1.00 (0.00)		
μ_2	0.03 (0.00)	0.06 (0.00)	0.02 (0.00)	1.00 (0.00)	
σ_2	-0.06 (0.00)	-0.11 (0.01)	-0.03 (0.00)	-0.06 (0.00)	1.00 (0.00)

Table C.46: Asymptotic Correlation Matrix for *RDML* Parameter Estimates for Data Set Structure # 14

$\mu_1 = 20 \quad \sigma_1 = 5 \quad \varpi(\%) = 50 \quad \mu_2 = 40 \quad \sigma_2 = 10 \quad n = 200$

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.53 (0.01)	1.00 (0.00)			
ϖ	0.57 (0.01)	0.53 (0.01)	1.00 (0.00)		
μ_2	0.55 (0.01)	0.51 (0.01)	0.74 (0.01)	1.00 (0.00)	
σ_2	-0.45 (0.01)	-0.42 (0.01)	-0.67 (0.00)	-0.73 (0.00)	1.00 (0.00)

Table C.47: Asymptotic Correlation Matrix for *RDML* Parameter Estimates for Data Set Structure # 15

$\mu_1 = 20 \quad \sigma_1 = 5 \quad \varpi(\%) = 50 \quad \mu_2 = 40 \quad \sigma_2 = 15 \quad n = 200$

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.39 (0.05)	1.00 (0.00)			
ϖ	0.40 (0.05)	0.61 (0.02)	1.00 (0.00)		
μ_2	0.35 (0.06)	0.59 (0.02)	0.77 (0.01)	1.00 (0.00)	
σ_2	-0.19 (0.07)	-0.45 (0.03)	-0.63 (0.03)	-0.68 (0.03)	1.00 (0.00)

Table C.48: Asymptotic Correlation Matrix for *RDML* Parameter Estimates for Data Set Structure # 16

$\mu_1 = 20 \quad \sigma_1 = 5 \quad \varpi(\%) = 50 \quad \mu_2 = 40 \quad \sigma_2 = 5 \quad n = 50$

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.29 (0.02)	1.00 (0.00)			
ϖ	0.21 (0.02)	0.22 (0.02)	1.00 (0.00)		
μ_2	0.24 (0.02)	0.23 (0.02)	0.22 (0.02)	1.00 (0.00)	
σ_2	-0.24 (0.02)	-0.21 (0.01)	-0.23 (0.02)	-0.32 (0.03)	1.00 (0.00)

C.2 Class Interval Data Maximum Likelihood Function

The following tables contain the mean asymptotic correlations of the *CIDML* parameter estimates calculated using the inverse of the negative Hessian matrix (observed Fisher information matrix) evaluated at the *CIDML* parameter estimates. Correlation matrices for 9 different numbers of class intervals for data set structure # 1 are presented.

(Values in parentheses are the standard deviations of the asymptotic correlations for the 10 data sets of the corresponding parameters.)

Table C.49: Asymptotic Correlation Matrix for *CIDML* Parameter Estimates for Data Set Structure # 1 (Datum) Using 10 Class Intervals

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.31 (0.24)	1.00 (0.00)			
ϖ	0.17 (0.14)	0.47 (0.35)	1.00 (0.00)		
μ_2	0.32 (0.26)	0.47 (0.37)	0.52 (0.39)	1.00 (0.00)	
σ_2	-0.26 (0.23)	-0.21 (0.23)	0.07 (0.16)	-0.27 (0.31)	1.00 (0.00)

Table C.50: Asymptotic Correlation Matrix for *CIDML* Parameter Estimates for Data Set Structure # 1 (Datum) Using 15 Class Intervals

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.37 (0.15)	1.00 (0.00)			
ϖ	0.16 (0.12)	0.38 (0.15)	1.00 (0.00)		
μ_2	0.32 (0.14)	0.43 (0.15)	0.47 (0.11)	1.00 (0.00)	
σ_2	-0.27 (0.11)	-0.24 (0.11)	0.05 (0.09)	-0.30 (0.13)	1.00 (0.00)

Table C.51: Asymptotic Correlation Matrix for *CIDML* Parameter Estimates for Data Set Structure # 1 (Datum) Using 20 Class Intervals

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.34 (0.08)	1.00 (0.00)			
ϖ	0.12 (0.05)	0.33 (0.09)	1.00 (0.00)		
μ_2	0.29 (0.10)	0.39 (0.10)	0.44 (0.08)	1.00 (0.00)	
σ_2	-0.25 (0.10)	-0.22 (0.11)	0.08 (0.11)	-0.27 (0.19)	1.00 (0.00)

Table C.52: Asymptotic Correlation Matrix for *CIDML* Parameter Estimates for Data Set Structure # 1 (Datum) Using 25 Class Intervals

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.38 (0.15)	1.00 (0.00)			
ϖ	0.16 (0.10)	0.36 (0.13)	1.00 (0.00)		
μ_2	0.33 (0.15)	0.42 (0.15)	0.48 (0.10)	1.00 (0.00)	
σ_2	-0.29 (0.13)	-0.25 (0.13)	0.04 (0.14)	-0.30 (0.19)	1.00 (0.00)

Table C.53: Asymptotic Correlation Matrix for *CIDML* Parameter Estimates for Data Set Structure # 1 (Datum) Using 30 Class Intervals

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.34 (0.10)	1.00 (0.00)			
ϖ	0.14 (0.07)	0.37 (0.08)	1.00 (0.00)		
μ_2	0.29 (0.11)	0.40 (0.10)	0.48 (0.07)	1.00 (0.00)	
σ_2	-0.25 (0.10)	-0.21 (0.10)	0.09 (0.12)	-0.23 (0.16)	1.00 (0.00)

Table C.54: Asymptotic Correlation Matrix for *CIDML* Parameter Estimates for Data Set Structure # 1 (Datum) Using 35 Class Intervals

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.35 (0.14)	1.00 (0.00)			
ϖ	0.16 (0.11)	0.36 (0.12)	1.00 (0.00)		
μ_2	0.30 (0.14)	0.40 (0.14)	0.49 (0.09)	1.00 (0.00)	
σ_2	-0.26 (0.11)	-0.22 (0.11)	0.07 (0.11)	-0.23 (0.13)	1.00 (0.00)

Table C.55: Asymptotic Correlation Matrix for *CIDML* Parameter Estimates for Data Set Structure # 1 (Datum) Using 40 Class Intervals

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.38 (0.19)	1.00 (0.00)			
ϖ	0.19 (0.19)	0.40 (0.19)	1.00 (0.00)		
μ_2	0.32 (0.19)	0.43 (0.19)	0.52 (0.15)	1.00 (0.00)	
σ_2	-0.26 (0.12)	-0.22 (0.13)	0.06 (0.16)	-0.23 (0.16)	1.00 (0.00)

Table C.56: Asymptotic Correlation Matrix for *CIDML* Parameter Estimates for Data Set Structure # 1 (Datum) Using 45 Class Intervals

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.40 (0.17)	1.00 (0.00)			
ϖ	0.19 (0.14)	0.38 (0.17)	1.00 (0.00)		
μ_2	0.35 (0.17)	0.44 (0.16)	0.51 (0.13)	1.00 (0.00)	
σ_2	-0.29 (0.12)	-0.25 (0.13)	0.04 (0.13)	-0.27 (0.19)	1.00 (0.00)

Table C.57: Asymptotic Correlation Matrix for *CIDML* Parameter Estimates for Data Set Structure # 1 (Datum) Using 50 Class Intervals

	μ_1	σ_1	ϖ	μ_2	σ_2
μ_1	1.00 (0.00)				
σ_1	0.39 (0.21)	1.00 (0.00)			
ϖ	0.19 (0.13)	0.38 (0.10)	1.00 (0.00)		
μ_2	0.34 (0.17)	0.43 (0.15)	0.51 (0.09)	1.00 (0.00)	
σ_2	-0.29 (0.14)	-0.25 (0.14)	0.06 (0.12)	-0.24 (0.15)	1.00 (0.00)

Appendix D

Derivation of the Multivariate Regression Formula

"There are three type of lies in the world : lies, damned lies, and statistics."

Benjamin Disraeli (1923)

A possible background characterization approach (*BCA*) model which can be used to describe the variation in a 'background' population is a regression function. Unfortuately, an ordinary least squares (*OLS*) regression approach accomodates error in only the dependent variable; thus, it is not rigorously applicable to geochemical data because all geochemical variables are subject to error. Any regression applied to geochemical variables must accommodate these errors.

Geochemical concentration data have error variances which are frequently proportional to concentration (Thompson 1973; Thompson and Howarth 1973, 1976a, 1976b, 1978). Since the expected concentration range of data from the population representing the 'background' geologic material will not, in general, be large (at least not greater than an order of magnitude), the effect of this proportionality is minimized.

Furthermore, in *BCA* applications, the observations which are subjected to a regression analysis are presumed to be derived from a single population representing a single geologic material. As such, every element determination represents an estimate of the true concentration of the geologic material for that element, and thus relationships between elements have a structural rather than functional form (Dolby 1976a). The variance estimate calculated from the observations is an estimate of the unique and

constant error variance for that geological material, and the assumption of a constant error variance for each variable is, therefore, justified.

A multiple linear regression model which can be applied to this type of situation is :

$$\vec{\mathbf{Y}} = b_0 + \sum_{j=1}^{p} b_j \vec{\mathbf{X}}_\mathbf{j}, \tag{D.192}$$

where $\vec{\mathbf{Y}}$ and the $\vec{\mathbf{X}}_\mathbf{j}$'s are random vectors and b_0 and the b_j's are the regression coefficients. In practice, we have only the observations of the X_j and Y variables (which are observed with error). Thus :

$$x_{ij} = X_{ij} + \eta_{ij}, \tag{D.193}$$

and :

$$y_i = Y_i - \epsilon_i, \tag{D.194}$$

and the relationship between the observed element concentrations on the i^{th} case becomes :

$$y_i = b_0 + \sum_{j=1}^{p} b_j(x_{ij} + \eta_{ij}) + \epsilon_i, \tag{D.195}$$

where n is the number of cases, $i = 1, 2, 3, \ldots, n$, p is the number of independent variables, y_i is an observation of the dependent variable (Y_i), $-\epsilon_i$ is the error in an observation of the dependent variable, x_{ij} are the observations of the p independent variables (X_{ij}), and η_{ij} are the errors in the observations of the p independent variables.

Re-arranging this equation and combining the errors into ξ_i gives :

$$y_i = b_0 + \sum_{j=1}^{p} b_j x_{ij} + \xi_i, \tag{D.196}$$

where :

$$\xi_i = \sum_{j=1}^{p} b_j \eta_{ij} + \epsilon_i. \tag{D.197}$$

The usual assumptions in this regression model include :

- all observations for each variable are independent,
- the expectations of all η_{ij} and ϵ_i are 0,
- the variances of all η_{ij} and ϵ_i are σ_j^2 and σ_ϵ^2, respectively (all are constant across the range of the data),
- all observation errors (η_{ij} and ϵ_i) are independent and normally distributed, and
- all covariances between the variables (x_{ij} and y_i) and their errors (η_{ij} and ϵ_i) are 0.

From these assumptions it follows that :

- the expectation of ξ_i is 0, and
- the variance of ξ_i is $\sigma_\xi^2 = \sum_{j=1}^{p} b_j^2 \sigma_j^2 + \sigma_\epsilon^2$.

Thus :

$$y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij} \sim N(0, \sigma_\xi^2). \tag{D.198}$$

Now, the likelihood (L) is :

$$L = \prod_{i=1}^{n} \Big[\frac{1}{\sqrt{2\pi}\sigma_\xi} e^{-\frac{1}{2\sigma_\xi^2}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij})^2} \Big], \tag{D.199}$$

and the logarithm of the likelihood is :

$$\ell = \ln L = n \ln \frac{1}{\sqrt{2\pi}} - \frac{n}{2} \ln \sigma_\xi^2 - \frac{1}{2\sigma_\xi^2} \sum_{i=1}^{n} (y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij})^2. \tag{D.200}$$

If we define λ_j to be the ratio of the error variances :

$$\lambda_j = \frac{\sigma_\epsilon^2}{\sigma_j^2}, \tag{D.201}$$

then :

$$\sigma_\xi^2 = \sigma_\epsilon^2 \Big(1 + \sum_{j=1}^{p} \frac{b_j^2}{\lambda_j} \Big), \tag{D.202}$$

and we can substitute this into the logarithm of the likelihood (Equation D.200) to get :

$$\ell = \ln L = n \ln \frac{1}{\sqrt{2\pi}} - \frac{n}{2} \ln\Big(\sigma_\epsilon^2 (1 + \sum_{j=1}^{p} \frac{b_j^2}{\lambda_j})\Big) - \frac{1}{2\sigma_\epsilon^2 (1 + \sum_{j=1}^{p} \frac{b_j^2}{\lambda_j})} \sum_{i=1}^{n} (y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij})^2. \tag{D.203}$$

D.1 Maximum Likelihood Estimates

Now, to determine the maximum likelihood (ML) estimates of the regression coefficients ($\hat{b}_0$ and the $\hat{b}_j$'s), we must differentiate the logarithm of the likelihood with respect to the b_0 and b_j terms and set each partial derivative to zero.

Differentiating Equation D.203 with respect to b_0 gives :

$$\frac{\partial \ell}{\partial b_0} = \frac{1}{\sigma_\xi^2} \sum_{i=1}^{n} (y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij}), \tag{D.204}$$

which when summed and set to zero, reduces to :

$$0 = \bar{y} - b_0 - \sum_{j=1}^{p} b_j \bar{x}_j, \tag{D.205}$$

or :

$$b_0 = \bar{y} - \sum_{j=1}^{p} b_j \bar{x}_j. \tag{D.206}$$

Differentiating Equation D.203 with respect to b_t, where b_t represents each of the b_j terms, gives the following set of p simultaneous equations :

$$\frac{\partial \ell}{\partial b_t} = \frac{-n b_t}{\lambda_t (1 + \sum_{j=1}^{p} \frac{b_j^2}{\lambda_j})} + \frac{b_t \sum_{i=1}^{n} (y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij})^2}{\sigma_\epsilon^2 (1 + \sum_{j=1}^{p} \frac{b_j^2}{\lambda_j})^2} + \frac{\sum_{i=1}^{n} (y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij})^2 x_{ij}}{\sigma_\epsilon^2 (1 + \sum_{j=1}^{p} \frac{b_j^2}{\lambda_j})}. \tag{D.207}$$

Setting these equations to zero and reducing gives :

$$0 = \sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij})^2 x_{ij} - \frac{b_t}{\lambda_t}\left[n\sigma_\epsilon^2 - \frac{\sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij})^2}{(1 + \sum_{j=1}^{p} \frac{b_j^2}{\lambda_j})}\right]. \quad \text{(D.208)}$$

Unfortunately, this set of simultaneous equations (D.206 and D.208), while when solved produce the *ML* regression coefficients, is neither easily reduced nor capable of being solved numerically in a straightforward manner.

However, by differentiating the logarithm of the likelihood equation (D.203) with respect to σ_ϵ^2 and taking a least squares approach by choosing b_j and b_0 terms which minimize the result, we can more easily determine the *ML* regression coefficients (Moran 1971; Anderson 1984; Fuller 1987). Differentiating Equation D.203 with respect to σ_ϵ^2 gives :

$$\frac{\partial \ell}{\partial \sigma_\epsilon^2} = -\frac{n}{2\sigma_\epsilon^2} + \frac{\sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij})^2}{2\sigma_\epsilon^4(1 + \sum_{j=1}^{p} \frac{b_j^2}{\lambda_j})}. \quad \text{(D.209)}$$

Setting this equation to zero and multiplying both sides by σ_ϵ^4, we get :

$$0 = -\frac{n\sigma_\epsilon^2}{2} + \frac{\sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij})^2}{2(1 + \sum_{j=1}^{p} \frac{b_j^2}{\lambda_j})}, \quad \text{(D.210)}$$

which reduces to :

$$\sigma_\epsilon^2 = \frac{\sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij})^2}{n(1 + \sum_{j=1}^{p} \frac{b_j^2}{\lambda_j})}. \quad \text{(D.211)}$$

Thus, this least squares regression approach produces coefficient estimates of the regression model with error in every variable which are equivalent to the *ML* regression coefficient estimates (Fuller 1987, Ripley and Thompson 1987).

Minimization can be accomplished by using a non-linear numerical optimization procedure such as the SIMPLEX method of Nash (1979) and Caceci and Cacheris (1984). Unfortunately, the sum of squared residual formula of Equation D.211 is not particularly stable, especially when the independent variables are highly correlated

(Jones 1972). Specifically, this equation can be minimized when the b_j terms in the denominator are large. Because of the high correlation between independent variables, when at least one of the b_j terms become highly positive, the others, to compensate, become highly negative.

Using, as an example, a trivariate data set where the dependent variable is regressed against two independent variables which are highly correlated and no b_0 constant term exists, the resulting $\sum_{i=1}^{n} \sigma_\epsilon^2$ surface is a trough in $b_1 : b_2$ space. The axis of this trough has a slope in $b_1 : b_2$ space (where b_1 is the abcissa and b_2 is the ordinate) with a sign opposite to the sign of the correlation between X_1 and X_2 and a magnitude equal to $\frac{\sigma_{X_2}}{\sigma_{X_1}}$ (Figure D.1).

When the trivariate data set has a correlation structure of the form :

$$\begin{pmatrix} 1.0 & r_{12} & r_{13} \\ r_{12} & 1.0 & r_{23} \\ r_{13} & r_{23} & 1.0 \end{pmatrix}, \tag{D.212}$$

then, if $r_{23} = r_{12} = r_{13}$, the b_1 and b_2 *RMA* (see below) values corresponding to the minimum σ_ϵ^2 are indeterminate and the base of the trough is 'equal-valued' at all locations along its length. If $r_{23} > r_{12}$ and $r_{23} > r_{13}$, $r_{23} > r_{12}$ and $r_{23} = r_{13}$ or $r_{23} = r_{12}$ and $r_{23} > r_{13}$, then the *RMA* b_1 and b_2 values are also indeterminate, but the trough has a saddle form (concave-down). As a result, solution of the *ML* regression solution is not possible when substantial correlation exists between the independent variables. Only when $r_{23} < r_{12}$ and $r_{23} < r_{13}$ does the trough have a 'concave-up' form and a true minimum, the *ML* regression solution. Fortunately, correlation structures produced by the regression model of Equation D.192 will generally have structures where independent-independent variable correlations are lower than dependent-independent variable correlations, provided that the variances of the errors are small with respect to the variances of the variables.

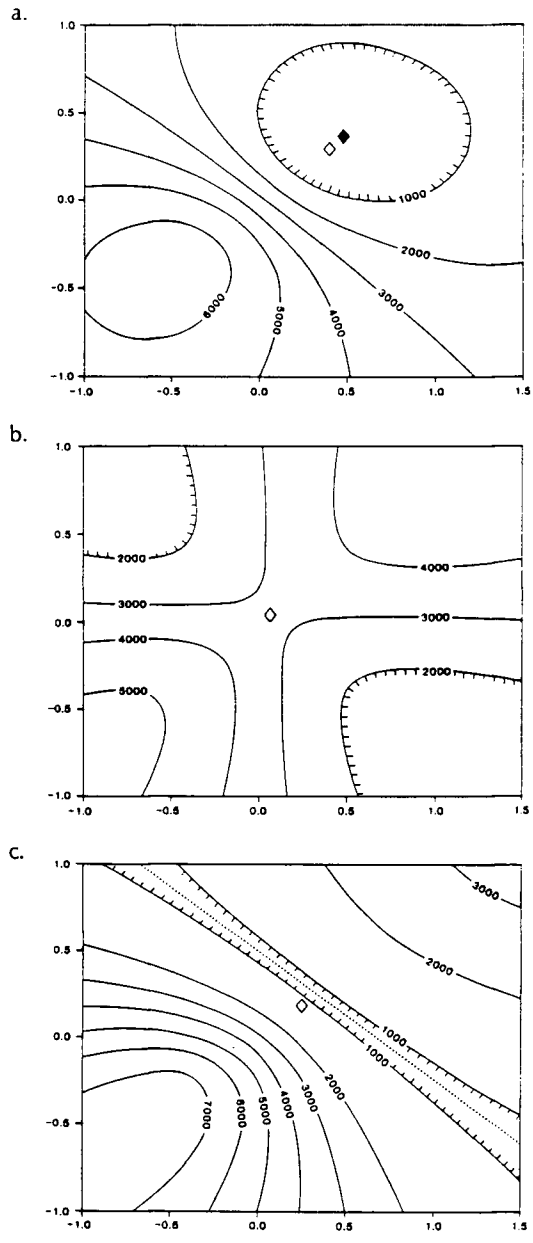


Figure D.40: $E\left(\sum_{i=1}^{n} \sigma_\epsilon^2\right)$ Reduced Major Axis Regression Surface for Various Correlation Structures

Contour plots of *RMA* $E\left(\sum_{i=1}^{n} \sigma_\epsilon^2\right)$ for data set structures # 17 (A), # 31 (B) and # 32 (C). The abcissa on all plots is b_1 and the ordinate is b_2. Filled diamond on plot A represents the *RMA* regression solution (minimum). Open diamonds on each plot represent the corresponding *OLS* regression solution. The dashed diagonal line on C is the minimum axis of the trough (with a slope in $b_1 : b_2$ space of $-\frac{4}{3}$).

D.2 Estimation of λ_j

With the *ML* regression procedure, knowledge (or assumption) of the individual error variances of the independent and dependent variables or the ratios of these error variances are required. This is because the solution, as described in Equation D.211, is under-determined. To illustrate this using a trivariate data set, the parameters of the population can be cast in terms of the regression model (Equation D.192) and estimated. In this case there are 11 population parameters which need to be approximated (μ_y, μ_{x_1}, μ_{x_2}, $\sigma^2_{x_1}$, $\sigma^2_{x_2}$, $\sigma^2_{\eta_1}$, $\sigma^2_{\eta_2}$, $\sigma_{x_1x_2}$, b_0, b_1 and b_2), but only 9 equations :

$$\bar{y} = \hat{b}_0 + \hat{b}_1\hat{\mu}_{x_1} + \hat{b}_2\hat{\mu}_{x_2}, \tag{D.213}$$

$$\bar{x}_1 = \hat{\mu}_{x_1}, \tag{D.214}$$

$$\bar{x}_2 = \hat{\mu}_{x_2}, \tag{D.215}$$

$$s^2_y = \hat{b}^2_1\hat{\sigma}^2_{x_1} + \hat{b}^2_2\hat{\sigma}^2_{x_2} + \hat{\sigma}^2_\epsilon, \tag{D.216}$$

$$s^2_{x_1} = \hat{\sigma}^2_{x_1} + \hat{\sigma}^2_{\eta_1}, \tag{D.217}$$

$$s^2_{x_2} = \hat{\sigma}^2_{x_2} + \hat{\sigma}^2_{\eta_2}, \tag{D.218}$$

$$s_{yx_1} = \hat{b}_1\hat{\sigma}^2_{x_1} + \hat{b}_2\hat{\sigma}_{x_1x_2}, \tag{D.219}$$

$$s_{yx_2} = \hat{b}_2\hat{\sigma}^2_{x_2} + \hat{b}_1\hat{\sigma}_{x_1x_2}, \tag{D.220}$$

and :

$$s_{x_1x_2} = \hat{\sigma}_{x_1x_2}. \tag{D.221}$$

We can calculate the estimates of the parameters for Equations D.214, D.215 and D.221, and substitute these into Equation D.213 to solve for μ_{x_1}, μ_{x_2}, $\sigma_{x_1x_2}$ and b_0, but this still leaves 5 equations and 7 unknowns.

Solution requires that certain assumptions be made in order to determine the *ML* regression estimates. The most common assumption made is that the ratios of the error

variances are known (Madansky 1959; Kendall and Stuart 1961; Mark and Church 1977; Jones 1979). This eliminates the p degrees of freedom (in general, equal to the number of independent variables), allowing determination of a unique solution.

Estimation of the ratios of the error variances is commonly achieved through the use of replicate analysis to estimate the variances (Thompson 1973; Thompson and Howarth 1973, 1976a, 1976b, 1978; Thompson 1982) and allow calculation of the error variance ratios (λ_j's). In the absence of these estimates, or when only a portion of the information required is known, several different approaches have been employed in bivariate applications to solve the *ML* regression (Madansky 1959; Kendall and Stuart 1961; Birch 1964; Till 1973; Dolby 1976b; Mark and Church 1977; Jones 1979). Some of these approaches can also readily be employed in multivariate applications.

The basic difference between the approaches lies in amount of information available about the error variances and severity of the assumptions employed. The various amounts of information regarding the error variances include cases where :

- all error variances are known (thus all error variance ratios can be calculated),
- the error variances are not known, but all of the ratios of the error variances between the dependent variable and each independent variable are known,
- some of the error variances or some of the error variance ratios are known, and
- error variances and error variance ratios are not known.

In the first case, a direct determination of each error variance ratio (λ_j) can be calculated and these can be used to minimize the *ML* regression equation and determine the b_j and b_0 terms. In the bivariate case, minimization of Equation D.211 produces

estimates identical to those produced by the formula :

$$\hat{b} = \sqrt{\frac{s_y^2 - s_\epsilon^2}{s_x^2 - s_\eta^2}}, \tag{D.222}$$

of Madansky (1959), Dolby (1976b), and Kendall and Stuart (1961). Thus, application of this approach is merely a generalization of the above bivariate formula. This approach has been used extensively in geochonological applications (York 1966; McIntyre et al. 1966; York 1967; Brooks et al. 1968; York 1969; Brooks et al. 1972)

If the ratios of the error variances are known, but the individual error variances are not known, simple substitution of these λ_j terms into the *ML* equation will lead to a solution (Madansky 1959; Kendall and Stuart 1961; Birch 1964; Till 1973; Dolby 1976b; Mark and Church 1977; Jones 1979). This also produces a result identical to the result obtained, in the bivariate case, using the formula :

$$\bar{b} = \frac{s_y^2 - \lambda s_x^2 + \sqrt{(s_y^2 - \lambda s_x^2)^2 + 4(s_{xy} - \rho\sqrt{\lambda}s_x^2)(\lambda s_{xy} - \rho\sqrt{\lambda}s_y^2)}}{2(s_{xy} - \rho\sqrt{\lambda}s_x^2)}, \tag{D.223}$$

of Jones (1979), which when, in this case $\rho = 0$ (the errors are uncorrelated), reduces to the formula :

$$\bar{b} = \frac{s_y^2 - \lambda s_x^2 + \sqrt{(s_y^2 - \lambda s_x^2)^2 + 4\lambda s_{xy}^2}}{2s_{xy}}, \tag{D.224}$$

of Kermack and Haldane (1950), Madansky (1959), Kendall and Stuart (1961), Davies and Goldsmith (1972) and Jones (1979).

If only some of the error variances are known, or similarly, if only some of the error variance ratios are known, no multivariate approach analogous to the bivariate formulae :

$$\hat{b} = \frac{s_y^2 - s_\epsilon^2}{s_{xy}}, \tag{D.225}$$

or :

$$\hat{b} = \frac{s_{xy}}{s_x^2 - s_\eta^2}, \tag{D.226}$$

of Madansky (1959) and Kendall and Stuart (1961) exists. In cases such as these, the geochemist is advised to estimate the unknown variances by whatever means possible (e.g. - replicate analysis) or to use the approach described below.

In the last case, where neither the error variances nor their ratios are known, two assumptions may be employed to approximate the ratios of the error variances. The assumption that all error variances are equal ($\lambda_j = 1$) produces what is referred to as a "major axis" (*MA*) regression solution, while the assumption that the error variances are proportional to the ratio of the variances of the actual observations :

$$\lambda_j = \frac{s_y^2}{s_{x_j}^2}, \tag{D.227}$$

produces what is referred to as a "reduced major axis" (*RMA*) regression solution. In the absence of specific over-riding experimental or theoretical considerations, the *RMA* regression solution is preferred over the *MA* solution because it produces estimates of the b_j's which are scale invariant.

Justification of the use of the *RMA* approach is found in the nature of the structural relation which this regression exemplifies. Specifically, since all observations of each variable are estimates of the 'true' value of that variable, the variance of these observations is, by definition, an estimate of the observation error variance. Using a λ_j which is proportional to the ratios of the observed sample variances is, thus, an appropriate estimate of the error variance ratio (for references regarding *ML* regression on a linear functional relation see : Lindley 1947; Villegas 1961; Solari 1969).

Solution of the *ML* regression in this way is analogous to solution of a bivariate case using Equation D.224 of Madansky (1959), Kendall and Stuart (1961), Dolby (1976b) and Jones (1979) where the error variance ratios are known.

D.3 Relationship to Principal Components

The *ML* regression coefficients estimated using Equation D.211 can be shown to be related to the *PC* of the multivariate distribution defined by the regression model (Equation D.192). Substituting :

$$\check{y}_i = y_i - \bar{y}, \tag{D.228}$$

and :

$$\check{x}_{ij} = x_{ij} - \bar{x}_j, \tag{D.229}$$

into Equation D.211 produces 'corrected' values of the variables (centered about 0), thus eliminating the b_0 term (Equation D.206), and giving :

$$\sigma_\epsilon^2 = \frac{\sum_{i=1}^{n}(\check{y}_i - \sum_{j=1}^{p} b_j \check{x}_{ij})^2}{n(1 + \sum_{j=1}^{p} \frac{b_j^2}{\lambda_j})}. \tag{D.230}$$

Knowledge of the λ_j terms allows scaling of the X_j variables such that $\lambda_j = 1$. Thus, multiplying each X_j variable by $\sqrt{\lambda_j}$ creates new independent variables with all new scaled λ_j (λ_{sj}) equalling one. The regression now is 'orthogonal' and corresponds, in a bivariate case, to the major axis of the data (Fuller 1988). Including the denominator in the summation of squares after the above multiplication gives :

$$\sigma_\epsilon^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\check{y}_{si} - \sum_{j=1}^{p} b_{sj}\check{x}_{sij}}{\sqrt{1 + \sum_{j=1}^{p} b_{sj}^2}}\right)^2, \tag{D.231}$$

where $\check{y}_{si}$ and $\check{x}_{sij}$ are the corrected and scaled values and the b_{sj} are the coefficients of the scaled data.

If we define :

$$\vec{z}_i = \begin{pmatrix} \check{y}_{si} \\ \check{x}_{si1} \\ \vdots \\ \check{x}_{sip} \end{pmatrix}, \tag{D.232}$$

and :

$$\vec{a} = \frac{1}{1 + \sum_{j=1}^{p} b_{sj}^2} \begin{pmatrix} 1 \\ -b_{s1} \\ \vdots \\ -b_{sp} \end{pmatrix} = \begin{pmatrix} a_y \\ a_1 \\ \vdots \\ a_p \end{pmatrix}, \tag{D.233}$$

then Equation D.231 can be cast in matrix notation and minimized, provided that the length of $\vec{a}$ is unity ($||\vec{a}|| = 1$). Thus :

$$\min_{||\vec{a}||=1} \sigma_\epsilon^2 = \min_{||\vec{a}||=1} \frac{1}{n} \sum_{i=1}^{n} (\vec{a}^T \vec{z}_i)^2 = \min_{||\vec{a}||=1} \frac{1}{n} \sum_{i=1}^{n} \vec{a}^T \vec{z}_i \vec{z}_i^T \vec{a}. \tag{D.234}$$

Since $\vec{a}$ is not involved in the summation :

$$\min_{||\vec{a}||=1} \sigma_\epsilon^2 = \min_{||\vec{a}||=1} \vec{a}^T \left(\frac{1}{n} \sum_{i=1}^{n} \vec{z}_i \vec{z}_i^T \right) \vec{a} = \min_{||\vec{a}||=1} \vec{a}^T \tilde{S}_s \vec{a}, \tag{D.235}$$

where $\tilde{S}_s$ is the observed covariance matrix of the scaled variables.

Thus, $\vec{a}$ corresponds to the eigenvector (PC) associated with the **smallest** eigenvalue of the scaled covariance matrix. This can be demonstrated conceptually (geometrically) by the following argument (Figure D.3) :

- the regression model defined above corresponds to fitting a hyperplane through the data;
- scaling the data converts all $\lambda_j = 1$, so that minimization of the sum of squared residuals is made on the perpendiculars to the regression hyperplane (orthogonal regression);
- the eigenvector associated with the smallest eigenvalue is the vector of the residual variance after removal of the variances associated with all other eigenvectors;
- this eigenvector would, thus, define the direction of the perpendiculars to the regression hyperplane.

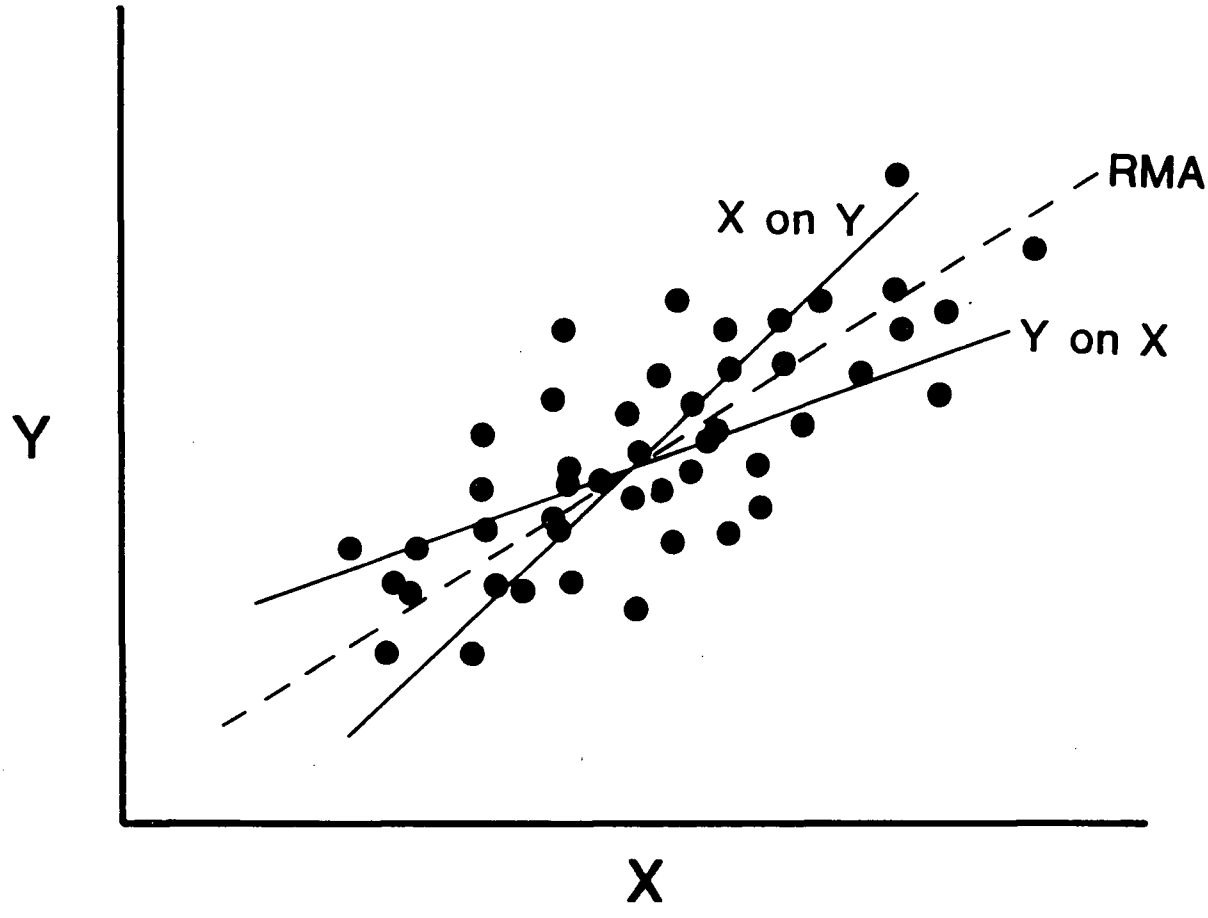


Figure D.41: Bivariate Example of the Relationships Between *OLS* Regression, *ML* Regression (*RMA* and *MA*) and *PC*

OLS regression lines consist of those labelled 'Y on X' (where $\sigma_\eta = 0$ and the residuals are vertical) and 'X on Y' (where $\sigma_\epsilon = 0$ and the residuals are horizontal). The *ML* regressions (*RMA* and *MA*) and first *PC* are all represented by 'RMA'. In this case, the *RMA* and *MA* regression lines are identical because $\sigma_\eta = \sigma_\epsilon$ was assumed (thus the residuals are perpendicular to the 'RMA' line). If $\sigma_\eta < \sigma_\epsilon$ then the *RMA* would have a slope between that of 'Y on X' and 'RMA', and if $\sigma_\eta > \sigma_\epsilon$ the slope would be between 'RMA' and 'X on Y' (the residuals to be minimized would not be perpendicular to the resulting line for both of these cases). In this bivariate case, the first *PC* is equivalent to the *MA* regression slope; however, in multivariate space, the *MA* corresponds to the hyper-plane defined by the eigenvectors corresponding to all non-smallest eigenvalues.

Clearly, if the smallest eigenvalue is not unique, the solution is indeterminate.

Thus, calculation of the eigenvectors and values of the scaled covariance matrix represents another method for determining the coefficients of a *ML* regression. Estimates of the scaled b_j terms (b_{sj}) can be calculated from :

$$b_{sj} = -\frac{a_j}{a_y}, \tag{D.236}$$

and then un-scaled to represent the true b_j values associated with the observed data by dividing each by its corresponding $\sqrt{\lambda_j}$. The b_0 constant can then be determined using Equation D.206.

Appendix E

Truncated Distribution Parameter Estimation

"Statistics means never having to say you're certain."

Anonymous

E.1 Truncated Multivariate Normal Parameter Estimates

While truncation of a normal distribution at some known point (x_0) yields different estimates of the mean and standard deviation of the population, this bias may be corrected and estimates of the *ML* population parameters calculated from these truncated sample estimates (Hald 1949; Cohen 1950, 1957, 1959, 1961).

To determine the expectations of the means, variances and covariances of a multivariate normal distribution which has been truncated at x_0, some known value of one of the variables, the nature and relationships of the different types of variables must be considered. Specifically, there is one variable used to truncate the distribution, and others, correlated with it, which are not used. The expected values and variances of the variable used for truncation, as well as the other correlated variables, can all be calculated; however, different approaches are required for each.

The quantities which must be evaluated include :

$$E(X|X \leq x_0), \tag{E.237}$$

$$VAR(X|X \leq x_0), \tag{E.238}$$

where X is the variable used to truncate the distribution,

$$E(Y|X \leq x_0), \tag{E.239}$$

$$VAR(Y|X \leq x_0), \tag{E.240}$$

where Y is not the variable used to truncate the distribution,

$$COV(X,Y|X \leq x_0), \tag{E.241}$$

and :

$$COV(Y,W|X \leq x_0), \tag{E.242}$$

where W and Y are both variables which were not used to truncate the distribution.

E.1.1 Truncated Variable Parameters

To determine the expected value of X, the variable used for truncation, the relevant probabilities must be evaluated :

$$Pr(X \leq x|X \leq x_0) = \frac{Pr(X \leq x, X \leq x_0)}{Pr(X \leq x_0)}. \tag{E.243}$$

If $x \geq x_0$, then this equals 1; however, if $x \leq x_0$, then :

$$\frac{Pr(X \leq x, X \leq x_0)}{Pr(X \leq x_0)} = \frac{Pr(X \leq x)}{Pr(X \leq x_0)}. \tag{E.244}$$

Substituting the appropriate normal distribution functions into this equation and differentiating with respect to x gives :

$$f(x|X \leq x_0) = \frac{1}{\Phi(z_0)} \frac{1}{\sigma_x} \phi\left(\frac{x - \mu_x}{\sigma_x}\right), \tag{E.245}$$

when $x \leq x_0$, and 0 when $x > x_0$, where :

$$z_0 = \frac{x_0 - \mu_x}{\sigma_x}. \tag{E.246}$$

Thus, the expectation of X given $X \leq x_0$ can be determined by multiplying x by $f(x|X \leq x_0)$ and integrating the result from $-\infty$ to x_0, giving :

$$E(X|X \leq x_0) = \int_{-\infty}^{x_0} x f(x|X \leq x_0) dx = \int_{-\infty}^{x_0} x \frac{1}{\Phi(z_0)} \frac{1}{\sigma_x} \phi\left(\frac{x - \mu_x}{\sigma_x}\right) dx, \tag{E.247}$$

Performing a change of variables, where :

$$z = \frac{x - \mu_x}{\sigma_x}, \tag{E.248}$$

and :

$$\sigma_x dz = dx, \tag{E.249}$$

gives :

$$E(X|X \leq x_0) = \frac{1}{\Phi(z_0)} \int_{-\infty}^{z_0} (\mu_x + z\sigma_x)\phi(z) dz, \tag{E.250}$$

which reduces to :

$$E(X|X \leq x_0) = \frac{1}{\Phi(z_0)} \left(\mu_x \int_{-\infty}^{z_0} \phi(z) dz + \sigma_x \int_{-\infty}^{z_0} z\phi(z) dz\right), \tag{E.251}$$

and :

$$E(X|X \leq x_0) = \mu_x + \left(\frac{1}{\Phi(z_0)} \sigma_x \int_{-\infty}^{z_0} z\phi(z) dz\right). \tag{E.252}$$

Taking this integral :

$$E(X|X \leq x_0) = \mu_x + \left(\frac{1}{\Phi(z_0)} \sigma_x \Big[-\phi(z)\Big]_{-\infty}^{z_0}\right), \tag{E.253}$$

and evaluating it from $-\infty$ to x_0, yields :

$$E(X|X \leq x_0) = \mu_x + \sigma_x E_{z_0}, \tag{E.254}$$

where

$$E_{z_0} = \frac{-\phi(z_0)}{\Phi(z_0)}. \tag{E.255}$$

This result is identical to that of Hald (1949) and Cohen (1950, 1957, 1959, 1961).

Similarly, the variance of X given $X \leq x_0$ is determined through multiplication of $f(x|X \leq x_0)$ by $(x - E(X|X \leq x_0)^2$ and integration of the result from $-\infty$ to x_0, giving :

$$\begin{aligned} VAR(X|X \leq x_0) &= \int_{-\infty}^{x_0} (x - E(X|X \leq x_0))^2 f(x|X \leq x_0) dx \\ &= \int_{-\infty}^{x_0} (x - E(X|X \leq x_0))^2 \frac{1}{\Phi(z_0)} \frac{1}{\sigma_x} \phi\left(\frac{x - \mu_x}{\sigma_x}\right) dx. \end{aligned} \quad \text{(E.256)}$$

Substituting the appropriate terms for $E(X|X \leq x_0)$ gives :

$$VAR(X|X \leq x_0) = \frac{1}{\Phi(z_0)} \int_{-\infty}^{x_0} (x - (\mu_x + \sigma_x E_{z_0}))^2 \frac{1}{\sigma_x} \phi\left(\frac{x - \mu_x}{\sigma_x}\right) dx. \quad \text{(E.257)}$$

Performing the same change of variables (as above) gives :

$$VAR(X|X \leq x_0) = \frac{1}{\Phi(z_0)} \int_{-\infty}^{z_0} (z\sigma_x - \sigma_x E_{z_0})^2 \phi(z) dz, \quad \text{(E.258)}$$

which reduces to :

$$VAR(X|X \leq x_0) = \frac{\sigma_x^2}{\Phi(z_0)} \int_{-\infty}^{z_0} (z^2 - 2zE_{z_0} + E_{z_0}^2) \phi(z) dz. \quad \text{(E.259)}$$

Now, adding and subtracting 1 to the quadratic term inside the integral, rearranging and then integrating by parts gives :

$$\begin{aligned} VAR(X|X \leq x_0) &= \frac{\sigma_x^2}{\Phi(z_0)} \int_{-\infty}^{z_0} (z^2 - 1)\phi(z) dz + \frac{\sigma_x^2}{\Phi(z_0)} \int_{-\infty}^{z_0} \phi(z) dz - \\ & 2E_{z_0} \frac{\sigma_x^2}{\Phi(z_0)} \int_{-\infty}^{z_0} z\phi(z) dz + E_{z_0}^2 \frac{\sigma_x^2}{\Phi(z_0)} \int_{-\infty}^{z_0} \phi(z) dz. \end{aligned} \quad \text{(E.260)}$$

Taking the integral gives :

$$\begin{aligned} VAR(X|X \leq x_0) &= \frac{\sigma_x^2}{\Phi(z_0)} \Big[-z\phi(z)\Big]_{-\infty}^{z_0} + \frac{\sigma_x^2}{\Phi(z_0)} \Big[\Phi(z)\Big]_{-\infty}^{z_0} - \\ & \frac{\sigma_x^2}{\Phi(z_0)} 2E_{z_0} \Big[-\phi(z)\Big]_{-\infty}^{z_0} + \frac{\sigma_x^2}{\Phi(z_0)} E_{z_0}^2 \Big[\Phi(z)\Big]_{-\infty}^{z_0}, \end{aligned} \quad \text{(E.261)}$$

and evaluating the result from $-\infty$ to z_0 gives (see Equation E.255) :

$$VAR(X|X \leq x_0) = \sigma_x^2 \left(z_0 E_{z_0} + 1 - 2E_{z_0}^2 + E_{z_0}^2\right), \quad \text{(E.262)}$$

which reduces to :

$$VAR(X|X \leq x_0) = \sigma_x^2 V_{z_0}, \tag{E.263}$$

where :

$$V_{z_0} = 1 + z_0 E_{z_0} - E_{z_0}^2. \tag{E.264}$$

This result is also identical to that of Hald (1949) and Cohen (1950, 1957, 1959, 1961).

Thus, for any variable Z distributed $N(0,1)$ which is truncated at z_0, by substituting 0 and 1 for μ_x and σ_x into equations E.254 and E.263, the expected value and variance can be shown to be E_{z_0} and V_{z_0}, respectively.

E.1.2 Un-Truncated Variable Parameters

To detemine the equations for the other four distribution parameters, we may cast all of the variables in terms of the independent variables Z_1, Z_2 and Z_3, which are each distributed $N(0,1)$. Thus :

$$X = \mu_x + \sigma_x Z_1, \tag{E.265}$$

$$Y = \mu_y + C_1 Z_1 + C_2 Z_2, \tag{E.266}$$

and :

$$W = \mu_w + D_1 Z_1 + D_2 Z_2 + D_3 Z_3. \tag{E.267}$$

The variances and covariances of each of these variables can be written in terms of C_1 and C_2 (for Y), and D_1, D_2 and D_3 (for W). This produces the following equations :

$$C_1^2 + C_2^2 = \sigma_y^2, \tag{E.268}$$

$$C_1 \sigma_x = \rho_{xy} \sigma_x \sigma_y, \tag{E.269}$$

and

$$D_1^2 + D_2^2 + D_3^2 = \sigma_w, \tag{E.270}$$

$$D_1\sigma_x = \rho_{xw}\sigma_x\sigma_w, \tag{E.271}$$

$$C_1D_1 + C_2D_2 = \rho_{yw}\sigma_y\sigma_w. \tag{E.272}$$

These can be solved simultaneously for C_1, C_2, D_1, D_2 and D_3, producing :

$$C_1 = \rho_{xy}\sigma_y, \tag{E.273}$$

$$C_2 = \sigma_y\sqrt{1-\rho_{xy}^2}, \tag{E.274}$$

$$D_1 = \rho_{xw}\sigma_w, \tag{E.275}$$

$$D_2 = \frac{\sigma_w(\rho_{yw} - \rho_{xy}\rho_{xw})}{\sqrt{1-\rho_{xy}^2}}, \tag{E.276}$$

and :

$$D_3 = \sigma_w\sqrt{1-\rho_{xw}^2 - \frac{(\rho_{yw} - \rho_{xy}\rho_{xw})^2}{1-\rho_{xy}^2}}. \tag{E.277}$$

Now, substituting the formulae of the variables X, Y and W in terms of Z_1, Z_2 and Z_3 into the formula for the expectations, variances and covariances of the variables truncated at x_0, we can derive the appropriate equations.

Since $\{X \leq x_0\}$ if and only if $\{Z \leq z_0\}$, the conditioning event $\{X \leq x_0\}$ can be replaced by $\{Z \leq z_0\}$. Thus :

$$E(Y|X \leq x_0) = E(\mu_y + C_1Z_1 + C_2Z_2|Z_1 \leq z_0), \tag{E.278}$$

and the individual expectations are :

$$E(Y|X \leq x_0) = \mu_y + C_1E(Z_1|Z_1 \leq z_0) + C_2E(Z_2|Z_1 \leq z_0). \tag{E.279}$$

Thus, since $E(Z_2|Z_1 \leq z_0) = 0$ because Z_1 and Z_2 are independent random variables :

$$\begin{aligned} E(Y|X \leq x_0) &= \mu_y + C_1E_{z_0} + 0 \\ &= \mu_y + \rho_{xy}\sigma_yE_{z_0}. \end{aligned} \tag{E.280}$$

By a similar derivation :

$$E(W|X \leq x_0) = \mu_w + \rho_{xw}\sigma_w E_{z_0}. \tag{E.281}$$

Derivation of the variance of the non-truncating variables is achieved in a similar manner, where :

$$VAR(Y|X \leq x_0) = VAR(\mu_y + C_1 Z_1 + C_2 Z_2 | Z_1 \leq z_0), \tag{E.282}$$

which reduces to :

$$\begin{aligned} VAR(Y|X \leq x_0) &= C_1^2 VAR(Z_1|Z_1 \leq z_0) + C_2^2 VAR(Z_2|Z_1 \leq z_0) + \\ &\quad 2C_1 C_2 COV(Z_1 Z_2|Z_1 \leq z_0). \end{aligned} \tag{E.283}$$

Evaluating the individual terms gives :

$$VAR(Y|X \leq x_0) = C_1^2 V_{z_0} + C_2^2 + 0 \tag{E.284}$$

After reduction and substitution of the constants :

$$VAR(Y|X \leq x_0) = \sigma_y^2(1 + \rho_{xy}^2(V_{z_0} - 1)). \tag{E.285}$$

By a similar derivation :

$$VAR(W|X \leq x_0) = \sigma_w^2(1 + \rho_{xw}^2(V_{z_0} - 1)). \tag{E.286}$$

The covariance between the truncating variable and any non-truncating variable can be determined by the following. First :

$$COV(X, Y|X \leq x_0) = COV(\mu_x + \sigma_x Z_1, \mu_y + C_1 Z_1 + C_2 Z_2 | Z_1 \leq z_0), \tag{E.287}$$

which reduces to :

$$COV(X, Y|X \leq x_0) = \sigma_x C_1 VAR(Z_1|Z_1 \leq z_0) + \sigma_x C_2 COV(Z_1, Z_2|Z_1 \leq z_0). \tag{E.288}$$

Evaluating the individual terms gives :

$$COV(X,Y|X \leq x_0) = \sigma_x C_1 V_{z_0} + 0, \tag{E.289}$$

which, after reduction and substitution of the constants, becomes :

$$COV(X,Y|X \leq x_0) = \rho_{xy}\sigma_x\sigma_y V_{z_0}. \tag{E.290}$$

By a similar derivation :

$$COV(X,W|X \leq x_0) = \rho_{xw}\sigma_x\sigma_w V_{z_0}. \tag{E.291}$$

Finally, the covariance between two non-truncating variables is :

$$\begin{aligned} COV(Y,W|X \leq x_0) &= COV(\mu_y + C_1Z_1 + C_2Z_2, \\ &\quad \mu_w + D_1Z_1 + D_2Z_2 + D_3Z_3|Z_1 \leq z_0). \end{aligned} \tag{E.292}$$

Evaluating the individual terms gives :

$$\begin{aligned} COV(Y,W|X \leq x_0) &= C_1D_2COV(Z_1,Z_2|Z_1 \leq z_0) + C_1D_3COV(Z_1,Z_3|Z_1 \leq z_0) + \\ &\quad C_2D_1COV(Z_1,Z_2|Z_1 \leq z_0) + C_2D_3COV(Z_2,Z_3|Z_1 \leq z_0) + \\ &\quad C_1D_1VAR(Z_1|Z_1 \leq z_0) + C_2D_2VAR(Z_2|Z_1 \leq z_0), \end{aligned} \tag{E.293}$$

and reducing these individual terms, we get :

$$COV(Y,W|X \leq x_0) = C_1D_1V_{z_0} + C_2D_2, \tag{E.294}$$

which, after substitution, becomes :

$$COV(Y,W|X \leq x_0) = \rho_{xy}\rho_{xw}\sigma_y\sigma_w(V_{z_0} - 1) + \sigma_y\sigma_w\rho_{yw}. \tag{E.295}$$

E.2 Solution of Simultaneous Equations

The expected values, variances and covariances for truncated multivariate normal distributions (derived above) are :

$$E(X|X \leq x_0) = \mu_{tx} = \mu_x + \sigma_x E_{z_0}, \tag{E.296}$$

$$E(Y|X \leq x_0) = \mu_{ty} = \mu_y + \rho_{xy}\sigma_y E_{z_0}, \tag{E.297}$$

$$E(W|X \leq x_0) = \mu_{tw} = \mu_w + \rho_{xw}\sigma_w E_{z_0}, \tag{E.298}$$

$$VAR(X|X \leq x_0) = \sigma_{tx}^2 = \sigma_x^2 V_{z_0}, \tag{E.299}$$

$$VAR(Y|X \leq x_0) = \sigma_{ty}^2 = \sigma_y^2(1 + \rho_{xy}^2(V_{z_0} - 1)), \tag{E.300}$$

$$VAR(W|X \leq x_0) = \sigma_{tw}^2 = \sigma_w^2(1 + \rho_{xw}^2(V_{z_0} - 1)), \tag{E.301}$$

$$COV(X,Y|X \leq x_0) = \sigma_{txy} = \rho_{xy}\sigma_x\sigma_y V_{z_0}, \tag{E.302}$$

$$COV(X,W|X \leq x_0) = \sigma_{txw} = \rho_{xw}\sigma_x\sigma_w V_{z_0}, \tag{E.303}$$

$$COV(Y,W|X \leq x_0) = \sigma_{tyw} = \rho_{xy}\rho_{xw}\sigma_y\sigma_w(V_{z_0} - 1) + \sigma_y\sigma_w\rho_{yw}, \tag{E.304}$$

where :

$$E_{z_0} = \frac{-\phi(z_0)}{\Phi(z_0)}, \tag{E.305}$$

and :

$$V_{z_0} = 1 + z_0 E_{z_0} - E_{z_0}^2. \tag{E.306}$$

However, in this application, the parameters of the truncated distribution may be estimated from the data. Estimation of the parameters of the underlying distribution requires recasting these equations and solving them simultaneously.

The first requirement for solution is knowledge of the truncation value x_0 which was used to truncate the normally distributed data. In the general case, equations E.296 and E.299 are solved in an iterative non-linear manner using the truncated data. This

yields estimates of μ_x and σ_x, and thus of z_0. Estimation of E_{z_0} and V_{z_0} follows directly. With knowledge of z_0, μ_x and σ_x, we are left with 7 equations and 7 unknowns, which can be solved analytically through substitution.

In this application, μ_x, σ_x and z_0 may be estimated directly because they have already been determined from the ML solution parameters of a mixture of normal distributions. As a result, solution of the 7 unknowns is the primary task, and can be accomplished by algebraically recasting these equations in terms of the truncated parameters.

First, rearranging the equations for σ_{txy}, σ_{tx}, σ_{ty}, and σ_{tw}, in terms of the untruncated parameters gives :

$$\rho_{xy} = \frac{\sigma_{txy}}{V_{z_0}\sigma_x\sigma_y}, \tag{E.307}$$

$$\boxed{\sigma_x = \frac{\sigma_{tx}}{\sqrt{V_{z_0}}},} \tag{E.308}$$

$$\sigma_y = \frac{\sigma_{ty}}{\sqrt{1+\rho_{xy}^2(V_{z_0}-1)}}, \tag{E.309}$$

and :

$$\sigma_w = \frac{\sigma_{tw}}{\sqrt{1+\rho_{xw}^2(V_{z_0}-1)}}. \tag{E.310}$$

Substituting the recast standard deviation equations into the correlation equation, above, produces :

$$\rho_{xy} = \frac{\sigma_{txy}}{V_{z_0}}\frac{\sqrt{V_{z_0}}}{\sigma_{tx}}\frac{\sqrt{1+\rho_{xy}^2(V_{z_0}-1)}}{\sigma_{ty}}, \tag{E.311}$$

which, because $\rho_{txy} = \sigma_{txy}/\sigma_{tx}\sigma_{ty}$, reduces to :

$$\rho_{xy} = \rho_{txy}\frac{\sqrt{1+\rho_{xy}^2(V_{z_0}-1)}}{\sqrt{V_{z_0}}}. \tag{E.312}$$

Combining terms and solving for ρ_{xy} gives :

$$\boxed{\rho_{xy} = \frac{\rho_{txy}}{\sqrt{V_{z_0}-\rho_{txy}^2(V_{z_0}-1)}},} \tag{E.313}$$

or :

$$\boxed{\sigma_{xy} = \frac{\sigma_{txy}}{\sqrt{V_{z_0}}\sqrt{V_{z_0} - \left(\frac{\sigma_{txy}}{\sigma_{tx}\sigma_{ty}}\right)^2 (V_{z_0} - 1)}}.} \tag{E.314}$$

By similar derivation :

$$\boxed{\rho_{xw} = \frac{\rho_{txw}}{\sqrt{V_{z_0} - \rho_{txw}^2 (V_{z_0} - 1)}},} \tag{E.315}$$

or :

$$\boxed{\sigma_{xw} = \frac{\sigma_{txw}}{\sqrt{V_{z_0}}\sqrt{V_{z_0} - \left(\frac{\sigma_{txw}}{\sigma_{tx}\sigma_{tw}}\right)^2 (V_{z_0} - 1)}}.} \tag{E.316}$$

Now, substituting the untruncated correlation equations (E.313 and E.315) into the recast standard deviation equations (E.309 and E.310) we get :

$$\sigma_y = \frac{\sigma_{ty}}{\sqrt{\frac{1+\rho_{txy}^2 (V_{z_0} - 1)}{V_{z_0} - \rho_{txy}^2 (V_{z_0} - 1)}}}, \tag{E.317}$$

and :

$$\sigma_w = \frac{\sigma_{tw}}{\sqrt{\frac{1+\rho_{txw}^2 (V_{z_0} - 1)}{V_{z_0} - \rho_{txw}^2 (V_{z_0} - 1)}}}, \tag{E.318}$$

which reduce to :

$$\boxed{\sigma_y = \frac{\sigma_{ty}}{\sqrt{V_{z_0}}}\sqrt{V_{z_0} - \rho_{txy}^2 (V_{z_0} - 1)},} \tag{E.319}$$

and :

$$\boxed{\sigma_w = \frac{\sigma_{tw}}{\sqrt{V_{z_0}}}\sqrt{V_{z_0} - \rho_{txw}^2 (V_{z_0} - 1)},} \tag{E.320}$$

or :

$$\boxed{\sigma_y = \frac{\sigma_{ty}}{\sqrt{V_{z_0}}}\sqrt{V_{z_0} - \left(\frac{\sigma_{txy}}{\sigma_{tx}\sigma_{ty}}\right)^2 (V_{z_0} - 1)},} \tag{E.321}$$

and :

$$\boxed{\sigma_w = \frac{\sigma_{tw}}{\sqrt{V_{z_0}}}\sqrt{V_{z_0} - \left(\frac{\sigma_{txw}}{\sigma_{tx}\sigma_{tw}}\right)^2 (V_{z_0} - 1).}} \tag{E.322}$$

Substituting all of these relations into the covariance relation between non-truncating variables, we get :

$$\sigma_{tyw} = \frac{\sigma_{ty}\sqrt{V_{z_0} - \rho_{xy}^2(V_{z_0}-1)}}{\sqrt{V_{z_0}}} \frac{\sigma_{tw}\sqrt{V_{z_0} - \rho_{xw}^2(V_{z_0}-1)}}{\sqrt{V_{z_0}}} \rho_{yw} \times \left(1 + \frac{\rho_{xy}}{\sqrt{V_{z_0} - \rho_{txy}^2(V_{z_0}-1)}} \frac{\rho_{xw}}{\sqrt{V_{z_0} - \rho_{txw}^2(V_{z_0}-1)}} \frac{(V_{z_0}-1)}{\rho_{yw}}\right), \quad \text{(E.323)}$$

which reduces to :

$$\boxed{\rho_{yw} = \frac{V_{z_0}\rho_{tyw} + (1-V_{z_0})\rho_{txy}\rho_{txw}}{\sqrt{V_{z_0} - \rho_{txy}^2(V_{z_0}-1)}\sqrt{V_{z_0} - \rho_{txw}^2(V_{z_0}-1)}},} \quad \text{(E.324)}$$

or :

$$\boxed{\sigma_{yw} = \sigma_{tyw} + \left(\frac{(1-V_{z_0})}{V_{z_0}} \frac{\sigma_{txy}\sigma_{txw}}{\sigma_{tx}^2}\right).} \quad \text{(E.325)}$$

Finally, the untruncated distribution means can be calculated through substitution of the standard deviations derived above, to give :

$$\boxed{\mu_x = \mu_{tx} - \frac{\sigma_{tx}}{\sqrt{V_{z_0}}} E_{z_0},} \quad \text{(E.326)}$$

$$\boxed{\mu_y = \mu_{ty} - \frac{\sigma_{ty}\rho_{txy}}{\sqrt{V_{z_0}}} E_{z_0},} \quad \text{(E.327)}$$

and :

$$\boxed{\mu_w = \mu_{tw} - \frac{\sigma_{tw}\rho_{txw}}{\sqrt{V_{z_0}}} E_{z_0}.} \quad \text{(E.328)}$$

Appendix F

BCA Parameter Comparison

"Statistics are like a bikini. What they reveal is tantalizing; what they conceal is vital."

Anonymous

The following tables present the average of 10 means, standard deviations, covariances, correlations, determinants, *OLS* regression coefficients, *RMA* regression coefficients, multiple correlation coefficients, least squared values, eigenvectors and eigenvalues for the population, statistical sample, truncated statistical sample (at 5, 15, 30 and 50 % truncation) and truncation corrected statistical sample for data set structures # 17 and # 23.

(Values in parentheses are the standard deviations of the calculated parameters for the 10 data sets used in this simulation.)

Table F.58: Population, Sample, Truncated Sample and Truncation Corrected Sample Parameter Estimates for the Number of Observations, Means and Standard Deviations of Multivariate Data Set Structure # 17

Truncation (%)	n	$\bar{y}$	$\bar{x}_1$	$\bar{x}_2$	s_y	s_{x_1}	s_{x_2}
Population	200.0	20.00	40.00	60.00	4.00	6.00	8.00
	(0.0)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Sample	200.0	19.98	39.93	60.11	3.97	5.88	8.03
	(0.0)	(0.30)	(0.56)	(0.40)	(0.19)	(0.22)	(0.42)
5	190.5	19.56	39.51	59.50	3.58	5.64	7.65
	(2.8)	(0.27)	(0.48)	(0.47)	(0.13)	(0.21)	(0.33)
15	171.5	18.94	38.83	58.65	3.21	5.35	7.36
	(4.7)	(0.26)	(0.51)	(0.45)	(0.15)	(0.22)	(0.37)
30	141.6	18.08	37.95	57.46	2.85	5.14	7.15
	(5.6)	(0.26)	(0.52)	(0.53)	(0.15)	(0.31)	(0.35)
50	98.7	16.79	36.67	55.57	2.44	4.89	6.85
	(7.4)	(0.14)	(0.40)	(0.56)	(0.21)	(0.34)	(0.51)
5	190.5	19.99	39.96	60.12	3.98	5.92	8.04
Corrected	(2.8)	(0.26)	(0.48)	(0.45)	(0.14)	(0.22)	(0.37)
15	171.5	20.04	39.95	60.23	4.01	5.89	8.14
Corrected	(4.7)	(0.23)	(0.51)	(0.42)	(0.19)	(0.24)	(0.44)
30	141.6	20.09	40.00	60.45	4.05	5.92	8.34
Corrected	(5.6)	(0.21)	(0.59)	(0.48)	(0.21)	(0.35)	(0.47)
50	98.7	20.03	39.99	60.37	4.05	5.92	8.41
Corrected	(7.4)	(0.21)	(0.62)	(1.41)	(0.36)	(0.44)	(0.96)

Table F.59: Population, Sample, Truncated Sample and Truncation Corrected Sample Parameter Estimates for the Number of Observations, Means and Standard Deviations of Multivariate Data Set Structure # 23

Truncation (%)	n	$\bar{y}$	$\bar{x}_1$	$\bar{x}_2$	s_y	s_{x_1}	s_{x_2}
Population	200.0	20.00	40.00	60.00	4.00	6.00	8.00
	(0.0)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Sample	200.0	20.06	39.99	59.97	4.04	6.08	7.68
	(0.0)	(0.31)	(0.52)	(0.31)	(0.17)	(0.26)	(0.37)
5	188.9	19.58	39.75	59.83	3.61	6.03	7.66
	(2.2)	(0.28)	(0.51)	(0.35)	(0.19)	(0.23)	(0.30)
15	167.9	18.89	39.54	59.66	3.20	6.02	7.70
	(3.7)	(0.23)	(0.60)	(0.20)	(0.25)	(0.21)	(0.32)
30	138.6	18.00	39.14	59.49	2.80	6.05	7.75
	(5.0)	(0.28)	(0.52)	(0.43)	(0.23)	(0.28)	(0.33)
50	100.7	16.84	38.60	59.39	2.40	5.99	7.65
	(7.5)	(0.30)	(0.61)	(0.57)	(0.22)	(0.23)	(0.53)
5	188.9	20.02	39.93	59.93	4.02	6.07	7.67
Corrected	(2.2)	(0.27)	(0.53)	(0.35)	(0.21)	(0.25)	(0.31)
15	167.9	19.98	40.06	59.92	4.00	6.14	7.73
Corrected	(3.7)	(0.22)	(0.69)	(0.19)	(0.31)	(0.26)	(0.34)
30	138.6	19.99	40.08	59.99	3.99	6.23	7.79
Corrected	(5.0)	(0.24)	(0.76)	(0.55)	(0.33)	(0.37)	(0.37)
50	100.7	20.01	40.19	60.51	3.97	6.23	7.79
Corrected	(7.5)	(0.20)	(0.77)	(1.27)	(0.36)	(0.30)	(0.58)

Table F.60: Population, Sample, Truncated Sample and Truncation Corrected Sample Parameter Estimates for the Covariances, Correlations and Determinants of Multivariate Data Set Structure # 17

Truncation (%)	r_{yx_1}	r_{yx_2}	$r_{x_1x_2}$	$\|R\|$	s_{yx_1}	s_{yx_2}	$s_{x_1x_2}$	$\|S\|$
Population	0.700	0.700	0.100	0.108	16.80	22.40	4.80	3981.31
	(0.000)	(0.000)	(0.000)	(0.000)	(0.00)	(0.00)	(0.00)	(0.00)
Sample	0.698	0.707	0.106	0.103	16.33	22.64	4.98	3573.10
	(0.036)	(0.046)	(0.086)	(0.017)	(1.66)	(3.28)	(4.18)	(550.56)
5	0.662	0.666	0.009	0.122	13.36	18.27	0.26	2901.22
	(0.038)	(0.051)	(0.090)	(0.016)	(1.13)	(2.38)	(4.08)	(411.05)
15	0.613	0.626	-0.092	0.148	10.52	14.85	-3.76	2357.20
	(0.043)	(0.055)	(0.098)	(0.018)	(1.05)	(2.17)	(4.23)	(380.20)
30	0.563	0.591	-0.185	0.170	8.24	12.06	-6.98	1853.45
	(0.039)	(0.066)	(0.093)	(0.023)	(0.86)	(1.92)	(3.99)	(275.32)
50	0.514	0.528	-0.293	0.195	6.14	9.01	-10.08	1309.67
	(0.057)	(0.124)	(0.132)	(0.028)	(0.92)	(2.85)	(5.28)	(317.91)
5	0.700	0.704	0.096	0.097	16.51	22.57	4.51	3467.22
Corrected	(0.037)	(0.049)	(0.086)	(0.015)	(1.39)	(2.94)	(4.20)	(558.91)
15	0.695	0.707	0.091	0.094	16.43	23.19	4.29	3448.89
Corrected	(0.041)	(0.050)	(0.091)	(0.017)	(1.63)	(3.39)	(4.45)	(707.23)
30	0.696	0.720	0.099	0.084	16.72	24.47	4.81	3313.34
Corrected	(0.037)	(0.058)	(0.085)	(0.021)	(1.74)	(3.90)	(4.19)	(705.79)
50	0.702	0.706	0.095	0.082	16.89	24.79	4.82	3075.80
Corrected	(0.051)	(0.117)	(0.129)	(0.048)	(2.54)	(7.83)	(6.51)	(1328.40)

Table F.61: Population, Sample, Truncated Sample and Truncation Corrected Sample Parameter Estimates for the Covariances, Correlations and Determinants of Multivariate Data Set Structure # 23

Truncation (%)	r_{yx_1}	r_{yx_2}	$r_{x_1x_2}$	$\|R\|$	s_{yx_1}	s_{yx_2}	$s_{x_1x_2}$	$\|S\|$
Population	0.300	0.100	0.100	0.896	7.20	3.20	4.80	33030.14
	(0.000)	(0.000)	(0.000)	(0.000)	(0.00)	(0.00)	(0.00)	(0.00)
Sample	0.288	0.132	0.109	0.883	7.12	4.11	5.19	31253.22
	(0.086)	(0.050)	(0.083)	(0.060)	(2.36)	(1.58)	(4.12)	(2677.27)
5	0.249	0.112	0.100	0.907	5.44	3.10	4.68	25062.22
	(0.090)	(0.038)	(0.086)	(0.053)	(2.03)	(1.04)	(4.09)	(1482.16)
15	0.250	0.097	0.106	0.904	4.82	2.36	4.96	19785.18
	(0.095)	(0.068)	(0.094)	(0.058)	(1.89)	(1.60)	(4.50)	(2192.13)
30	0.220	0.089	0.099	0.916	3.78	1.91	4.72	15711.69
	(0.109)	(0.064)	(0.102)	(0.056)	(1.98)	(1.34)	(4.91)	(1557.26)
50	0.200	0.107	0.090	0.917	2.90	1.95	4.21	11031.12
	(0.093)	(0.103)	(0.101)	(0.042)	(1.41)	(1.85)	(4.78)	(1827.77)
5	0.274	0.125	0.095	0.892	6.72	3.83	4.52	31041.58
Corrected	(0.097)	(0.042)	(0.078)	(0.061)	(2.51)	(1.28)	(3.78)	(1794.82)
15	0.305	0.120	0.096	0.871	7.53	3.69	4.59	31086.13
Corrected	(0.112)	(0.084)	(0.079)	(0.078)	(2.95)	(2.50)	(4.00)	(3357.57)
30	0.302	0.126	0.083	0.865	7.68	3.87	4.16	32165.50
Corrected	(0.144)	(0.089)	(0.079)	(0.090)	(4.02)	(2.71)	(4.22)	(3207.92)
50	0.315	0.174	0.076	0.829	7.97	5.38	3.83	30647.05
Corrected	(0.140)	(0.167)	(0.064)	(0.089)	(3.88)	(5.08)	(3.58)	(5156.75)

Table F.62: Population, Sample, Truncated Sample and Truncation Corrected Sample Parameter Estimates for the Ordinary Least Squares Regression Coefficients, R^2 and Average Sum of Squared Residuals for Multivariate Data Set Structure # 17

Truncation (%)	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_0$	R^2	$\overline{\sum(\hat{y}_i - y_i)^2}$
Population	0.424	0.318	-16.061	0.891	1.75
	(0.000)	(0.000)	(0.000)	(0.000)	(0.00)
Sample	0.426	0.317	-16.097	0.896	1.63
	(0.013)	(0.011)	(0.604)	(0.017)	(0.19)
5	0.416	0.310	-15.337	0.870	1.64
	(0.012)	(0.011)	(0.372)	(0.017)	(0.19)
15	0.407	0.301	-14.508	0.831	1.72
	(0.014)	(0.013)	(0.661)	(0.021)	(0.21)
30	0.386	0.288	-13.146	0.785	1.72
	(0.022)	(0.015)	(0.980)	(0.033)	(0.27)
50	0.366	0.269	-11.587	0.707	1.68
	(0.032)	(0.021)	(1.621)	(0.077)	(0.22)
5	0.428	0.319	-16.273	0.901	1.55
Corrected	(0.013)	(0.011)	(0.313)	(0.014)	(0.18)
15	0.433	0.321	-16.587	0.905	1.52
Corrected	(0.016)	(0.013)	(0.659)	(0.016)	(0.19)
30	0.431	0.321	-16.525	0.915	1.38
Corrected	(0.026)	(0.016)	(1.259)	(0.021)	(0.25)
50	0.433	0.315	-16.347	0.916	1.27
Corrected	(0.038)	(0.022)	(2.008)	(0.049)	(0.46)

Table F.63: Population, Sample, Truncated Sample and Truncation Corrected Sample Parameter Estimates for the Ordinary Least Squares Regression Coefficients, R^2 and Average Sum of Squared Residuals for Multivariate Data Set Structure # 23

Truncation (%)	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_0$	R^2	$\sum(\hat{y}_i - y_i)^2$
Population	0.195	0.035	10.067	0.095	14.48
	(0.000)	(0.000)	(0.000)	(0.000)	(0.00)
Sample	0.183	0.053	9.577	0.101	14.60
	(0.055)	(0.018)	(2.717)	(0.054)	(1.48)
5	0.144	0.042	11.356	0.067	12.15
	(0.053)	(0.014)	(2.511)	(0.044)	(1.45)
15	0.128	0.029	12.097	0.058	9.65
	(0.049)	(0.022)	(2.550)	(0.040)	(1.72)
30	0.100	0.023	12.709	0.041	7.52
	(0.051)	(0.017)	(2.105)	(0.034)	(1.17)
50	0.079	0.026	12.204	0.032	5.52
	(0.044)	(0.037)	(2.427)	(0.021)	(0.98)
5	0.175	0.052	9.932	0.095	14.65
Corrected	(0.061)	(0.017)	(2.917)	(0.058)	(1.84)
15	0.191	0.046	9.556	0.117	14.21
Corrected	(0.068)	(0.032)	(3.614)	(0.072)	(2.58)
30	0.188	0.048	9.505	0.124	14.04
Corrected	(0.088)	(0.033)	(3.705)	(0.085)	(2.53)
50	0.194	0.071	7.748	0.163	13.27
Corrected	(0.095)	(0.088)	(5.603)	(0.087)	(2.45)

Table F.64: Population, Sample, Truncated Sample and Truncation Corrected Sample Parameter Estimates for the Reduced Major Axis Regression Coefficients, R^2 and Average Sum of Squared Residuals for Multivariate Data Set Structure # 17

Truncation (%)	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_0$	R^2	$\overline{\sum(\hat{y}_i - y_i)^2}$
Population	0.448	0.336	-18.097	0.941	0.94
	(0.000)	(0.000)	(0.000)	(0.000)	(0.00)
Sample	0.449	0.334	-18.043	0.944	0.88
	(0.016)	(0.010)	(0.620)	(0.009)	(0.11)
5	0.445	0.331	-17.708	0.936	0.81
	(0.016)	(0.010)	(0.389)	(0.008)	(0.11)
15	0.443	0.328	-17.521	0.926	0.76
	(0.018)	(0.011)	(0.568)	(0.009)	(0.11)
30	0.431	0.322	-16.756	0.917	0.67
	(0.025)	(0.014)	(0.919)	(0.012)	(0.13)
50	0.426	0.313	-16.242	0.905	0.56
	(0.038)	(0.016)	(1.487)	(0.014)	(0.10)
5	0.450	0.335	-18.122	0.947	0.83
Corrected	(0.016)	(0.010)	(0.355)	(0.007)	(0.10)
15	0.455	0.337	-18.396	0.949	0.81
Corrected	(0.019)	(0.011)	(0.557)	(0.008)	(0.09)
30	0.449	0.335	-18.124	0.955	0.73
Corrected	(0.027)	(0.015)	(1.164)	(0.010)	(0.12)
50	0.452	0.330	-17.951	0.957	0.67
Corrected	(0.039)	(0.017)	(1.582)	(0.023)	(0.21)

Table F.65: Population, Sample, Truncated Sample and Truncation Corrected Sample Parameter Estimates for the Reduced Major Axis Regression Coefficients, R^2 and Average Sum of Squared Residuals for Multivariate Data Set Structure # 23

Truncation (%)	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_0$	R^2	$\overline{\sum(\hat{y}_i - y_i)^2}$
Population	0.667	0.000	-6.667	0.300	11.20
	(0.000)	(0.000)	(0.000)	(0.000)	(0.00)
Sample	0.660	0.028	-8.048	0.299	11.36
	(0.155)	(0.202)	(7.534)	(0.082)	(1.49)
5	0.703	-0.073	-4.011	0.264	9.59
	(0.285)	(0.330)	(9.847)	(0.087)	(1.49)
15	0.603	-0.076	-0.409	0.265	7.53
	(0.134)	(0.202)	(7.993)	(0.096)	(1.61)
30	0.411	0.012	1.119	0.251	5.85
	(0.329)	(0.257)	(7.707)	(0.084)	(0.95)
50	0.498	-0.078	2.271	0.251	4.24
	(0.629)	(0.536)	(10.684)	(0.083)	(0.70)
5	0.691	0.009	-8.137	0.286	11.55
Corrected	(0.186)	(0.234)	(8.304)	(0.096)	(1.91)
15	0.633	0.031	-7.230	0.315	11.00
Corrected	(0.112)	(0.156)	(6.804)	(0.115)	(2.44)
30	0.521	0.105	-7.412	0.328	10.73
Corrected	(0.309)	(0.215)	(6.542)	(0.114)	(2.20)
50	0.465	0.179	-9.843	0.376	9.86
Corrected	(0.203)	(0.232)	(10.889)	(0.118)	(2.17)

Table F.66: Population, Sample, Truncated Sample and Truncation Corrected Sample Parameter Estimates for the Eigenvector Coefficients for Multivariate Data Set Structure # 17

Truncation (%)	e_{11}	e_{12}	e_{21}	e_{23}	e_{32}	e_{33}
Population	0.463	0.315	0.294	-0.451	-0.440	-0.324
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Sample	0.454	0.299	0.302	-0.434	-0.441	-0.323
	(0.050)	(0.117)	(0.040)	(0.116)	(0.014)	(0.011)
5	0.385	0.159	0.342	-0.288	-0.433	-0.317
	(0.061)	(0.143)	(0.051)	(0.147)	(0.013)	(0.010)
15	0.306	-0.007	0.392	-0.110	-0.426	-0.310
	(0.068)	(0.151)	(0.056)	(0.163)	(0.016)	(0.012)
30	0.226	-0.177	0.430	0.083	-0.407	-0.298
	(0.066)	(0.145)	(0.050)	(0.163)	(0.025)	(0.014)
50	0.210	-0.167	0.402	0.304	-0.388	-0.281
	(0.158)	(0.451)	(0.143)	(0.325)	(0.035)	(0.020)
5	0.453	0.291	0.305	-0.427	-0.443	-0.324
Corrected	(0.053)	(0.123)	(0.041)	(0.124)	(0.014)	(0.010)
15	0.450	0.278	0.313	-0.415	-0.448	-0.326
Corrected	(0.071)	(0.154)	(0.053)	(0.154)	(0.018)	(0.012)
30	0.445	0.268	0.312	-0.404	-0.444	-0.325
Corrected	(0.071)	(0.137)	(0.044)	(0.141)	(0.027)	(0.016)
50	0.437	0.254	0.296	-0.170	-0.445	-0.320
Corrected	(0.165)	(0.336)	(0.160)	(0.470)	(0.039)	(0.020)

Table F.67: Population, Sample, Truncated Sample and Truncation Corrected Sample Parameter Estimates for the Eigenvector Coefficients for Multivariate Data Set Structure # 23

Truncation (%)	e_{11}	e_{12}	e_{21}	e_{23}	e_{32}	e_{33}
Population	0.092	0.187	0.307	-0.216	-0.315	-0.034
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Sample	0.127	0.243	0.280	-0.279	-0.293	-0.057
	(0.051)	(0.201)	(0.078)	(0.213)	(0.077)	(0.027)
5	0.088	0.213	0.204	-0.231	-0.213	-0.046
	(0.036)	(0.197)	(0.077)	(0.201)	(0.074)	(0.021)
15	0.061	0.193	0.168	-0.203	-0.173	-0.030
	(0.047)	(0.221)	(0.067)	(0.228)	(0.065)	(0.027)
30	0.045	0.185	0.123	-0.190	-0.126	-0.024
	(0.037)	(0.219)	(0.066)	(0.224)	(0.064)	(0.020)
50	0.042	0.171	0.093	-0.174	-0.094	-0.028
	(0.032)	(0.191)	(0.063)	(0.190)	(0.054)	(0.043)
5	0.117	0.223	0.268	-0.254	-0.281	-0.059
Corrected	(0.046)	(0.192)	(0.093)	(0.200)	(0.089)	(0.026)
15	0.108	0.201	0.287	-0.232	-0.298	-0.049
Corrected	(0.083)	(0.230)	(0.102)	(0.252)	(0.100)	(0.042)
30	0.107	0.196	0.277	-0.224	-0.288	-0.052
Corrected	(0.080)	(0.222)	(0.135)	(0.242)	(0.130)	(0.044)
50	0.136	0.189	0.259	-0.220	-0.286	-0.081
Corrected	(0.134)	(0.125)	(0.138)	(0.156)	(0.138)	(0.117)

Table F.68: Population, Sample, Truncated Sample and Truncation Corrected Sample Parameter Estimates for the Eigenvalues for Multivariate Data Set Structure # 17

Truncation (%)	λ_{e1}	λ_{e2}	λ_{e3}
Population	75.875	38.772	1.353
	(0.000)	(0.000)	(0.000)
Sample	76.726	36.999	1.265
	(8.270)	(3.922)	(0.145)
5	66.188	35.816	1.228
	(5.774)	(3.251)	(0.133)
15	59.541	32.525	1.218
	(6.164)	(2.408)	(0.124)
30	55.744	28.927	1.155
	(5.742)	(2.524)	(0.144)
50	50.881	25.325	1.057
	(11.783)	(7.503)	(0.109)
5	76.649	37.844	1.199
Corrected	(6.965)	(4.144)	(0.134)
15	78.509	37.532	1.168
Corrected	(7.757)	(3.698)	(0.146)
30	82.248	38.121	1.064
Corrected	(8.883)	(5.274)	(0.188)
50	84.681	37.786	0.985
Corrected	(18.885)	(6.444)	(0.372)

Table F.69: Population, Sample, Truncated Sample and Truncation Corrected Sample Parameter Estimates for the Eigenvalues for Multivariate Data Set Structure # 23

Truncation (%)	λ_{e1}	λ_{e2}	λ_{e3}
Population	65.194	37.178	13.627
	(0.000)	(0.000)	(0.000)
Sample	61.573	37.006	13.850
	(6.459)	(2.828)	(1.531)
5	60.696	35.779	11.660
	(4.902)	(2.089)	(1.443)
15	61.389	35.317	9.243
	(5.040)	(2.357)	(1.563)
30	62.035	35.387	7.260
	(5.416)	(2.703)	(1.174)
50	60.417	34.740	5.323
	(8.541)	(2.631)	(0.890)
5	61.019	37.120	13.877
Corrected	(5.162)	(2.621)	(1.854)
15	62.102	38.222	13.336
Corrected	(5.966)	(3.265)	(2.457)
30	63.075	39.651	13.150
Corrected	(7.286)	(4.191)	(2.348)
50	63.380	40.198	12.272
Corrected	(10.275)	(4.205)	(2.234)

Appendix G

Residual and Score Summary

"**Significance Level** : a natural constant, like π and e, whose value is 0.05."

S.J. Senn (1988)

The following tables present the average of 10 means, standard deviations and skewnesses of the residuals or scores of the truncated data for the *OLS* regression model, *RMA* regression model, *PC* # 2, *PC* # 3, and the radial (elliptical) distance between *PC* # 2 and # 3 scores and *PC* # 1 (the major axis) calculated for both the truncated statistical sample and truncation corrected statistical sample for data set structures # 17 and # 23.

(Values in parentheses are the standard deviations of the calculated parameters for the 10 data sets used in this simulation.)

Table G.70: Means, Standard Deviations, Skewnesses, Minima and Maxima of Ordinary Least Squares Regression Residuals of the Truncated Data for Data Set Structure # 17

Truncation (%)	$\bar{n}$	Mean	Standard Deviation	Skewness	Minimum	Maximum
5	9.5	-1.08	1.25	0.11	-2.84	0.86
		(0.52)	(0.36)	(0.66)	(1.10)	(1.07)
15	28.5	-1.03	1.21	-0.07	-3.48	1.43
		(0.32)	(0.18)	(0.48)	(0.77)	(0.76)
30	58.4	-1.26	1.18	-0.09	-3.96	1.31
		(0.32)	(0.16)	(0.25)	(0.77)	(0.71)
50	101.3	-1.52	1.21	-0.11	-4.76	1.49
		(0.30)	(0.12)	(0.15)	(0.88)	(0.69)
5	9.5	-0.83	1.28	0.11	-2.63	1.17
Corrected		(0.54)	(0.37)	(0.66)	(1.13)	(1.11)
15	28.5	-0.52	1.27	-0.05	-3.13	2.11
Corrected		(0.34)	(0.19)	(0.47)	(0.72)	(0.82)
30	58.4	-0.50	1.27	-0.04	-3.35	2.30
Corrected		(0.40)	(0.19)	(0.23)	(0.71)	(0.89)
50	101.3	-0.42	1.29	-0.00	-3.89	2.85
Corrected		(0.34)	(0.15)	(0.25)	(0.87)	(0.88)

Table G.71: Means, Standard Deviations, Skewnesses, Minima and Maxima of Reduced Major Axis Regression Residuals of the Truncated Data for Data Set Structure # 17

Truncation (%)	$\bar{n}$	Mean	Standard Deviation	Skewness	Minimum	Maximum
5	9.5	-0.57	1.33	0.10	-2.44	1.50
		(0.56)	(0.39)	(0.65)	(1.16)	(1.16)
15	28.5	-0.47	1.30	-0.04	-3.14	2.23
		(0.33)	(0.19)	(0.47)	(0.68)	(0.81)
30	58.4	-0.66	1.27	-0.04	-3.52	2.15
		(0.36)	(0.18)	(0.23)	(0.69)	(0.83)
50	101.3	-0.74	1.28	0.00	-4.17	2.47
		(0.34)	(0.14)	(0.25)	(0.87)	(0.76)
5	9.5	-0.45	1.35	0.09	-2.35	1.65
Corrected		(0.56)	(0.39)	(0.64)	(1.17)	(1.18)
15	28.5	-0.25	1.33	-0.03	-2.99	2.54
Corrected		(0.34)	(0.20)	(0.47)	(0.67)	(0.84)
30	58.4	-0.32	1.32	-0.03	-3.29	2.59
Corrected		(0.40)	(0.20)	(0.23)	(0.67)	(0.95)
50	101.3	-0.30	1.33	0.02	-3.86	3.07
Corrected		(0.36)	(0.16)	(0.28)	(0.90)	(0.92)

Table G.72: Means, Standard Deviations, Skewnesses, Minima and Maxima of PC # 2 Scores of the Truncated Data for Data Set Structure # 17

Truncation (%)	$\bar{n}$	Mean	Standard Deviation	Skewness	Minimum	Maximum
5	9.5	7.32	4.22	0.39	1.48	14.93
		(2.70)	(1.49)	(0.54)	(3.51)	(4.06)
15	28.5	8.42	4.67	0.17	-1.12	18.96
		(2.20)	(0.87)	(0.52)	(4.35)	(3.16)
30	58.4	9.29	4.63	0.29	-0.52	21.25
		(1.30)	(0.51)	(0.41)	(3.25)	(3.76)
50	101.3	9.38	5.10	0.44	-1.11	24.84
		(2.18)	(0.63)	(0.35)	(3.83)	(4.30)
5	9.5	5.31	4.48	0.35	-1.06	13.36
Corrected		(2.46)	(1.50)	(0.51)	(3.21)	(3.80)
15	28.5	4.94	5.31	0.06	-6.13	16.68
Corrected		(2.34)	(0.77)	(0.46)	(4.99)	(1.85)
30	58.4	4.62	5.46	-0.05	-8.73	17.60
Corrected		(1.47)	(0.22)	(0.32)	(2.41)	(1.61)
50	101.3	5.02	5.61	0.03	-9.18	19.43
Corrected		(2.42)	(0.39)	(0.30)	(4.06)	(3.17)

Table G.73: Means, Standard Deviations, Skewnesses, Minima and Maxima of PC # 3 Scores of the Truncated Data for Data Set Structure # 17

Truncation (%)	$\bar{n}$	Mean	Standard Deviation	Skewness	Minimum	Maximum
5	9.5	0.75	1.13	-0.12	-1.01	2.33
		(0.48)	(0.33)	(0.66)	(0.98)	(1.00)
15	28.5	0.71	1.10	0.05	-1.56	2.95
		(0.29)	(0.17)	(0.47)	(0.71)	(0.64)
30	58.4	0.93	1.08	0.07	-1.42	3.38
		(0.32)	(0.15)	(0.23)	(0.66)	(0.67)
50	101.3	1.15	1.10	0.06	-1.59	4.09
		(0.30)	(0.10)	(0.17)	(0.63)	(0.75)
5	9.5	0.59	1.15	-0.11	-1.19	2.20
Corrected		(0.49)	(0.34)	(0.65)	(1.00)	(1.01)
15	28.5	0.41	1.14	0.04	-1.96	2.75
Corrected		(0.30)	(0.17)	(0.46)	(0.73)	(0.61)
30	58.4	0.48	1.13	0.04	-2.02	3.02
Corrected		(0.36)	(0.17)	(0.22)	(0.81)	(0.62)
50	101.3	0.50	1.15	-0.01	-2.40	3.58
Corrected		(0.32)	(0.13)	(0.27)	(0.79)	(0.74)

Table G.74: Means, Standard Deviations, Skewnesses, Minima and Maxima of PC # 2 and # 3 Radial Distance of the Truncated Data for Data Set Structure # 17

Truncation (%)	$\bar{n}$	Mean	Standard Deviation	Skewness	Minimum	Maximum
5	9.5	1.05	0.63	0.42	0.32	2.18
		(0.30)	(0.25)	(0.47)	(0.10)	(0.80)
15	28.5	1.01	0.64	0.92	0.22	2.63
		(0.20)	(0.10)	(0.40)	(0.10)	(0.54)
30	58.4	1.20	0.71	0.86	0.21	3.15
		(0.23)	(0.12)	(0.32)	(0.12)	(0.73)
50	101.3	1.48	0.84	0.89	0.25	4.20
		(0.25)	(0.13)	(0.26)	(0.15)	(0.94)
5	9.5	1.00	0.66	0.34	0.21	2.18
Corrected		(0.30)	(0.28)	(0.56)	(0.08)	(0.84)
15	28.5	0.94	0.67	0.96	0.13	2.67
Corrected		(0.19)	(0.13)	(0.38)	(0.06)	(0.50)
30	58.4	1.09	0.74	0.88	0.09	3.23
Corrected		(0.22)	(0.18)	(0.23)	(0.07)	(0.84)
50	101.3	1.32	0.92	1.00	0.06	4.63
Corrected		(0.50)	(0.38)	(0.30)	(0.04)	(2.08)

Table G.75: Means, Standard Deviations, Skewnesses, Minima and Maxima of Ordinary Least Squares Regression Residuals of the Background Data for Data Set Structure # 17

Background (%)	$\bar{n}$	Mean	Standard Deviation	Skewness	Minimum	Maximum
Sample	200.0	0.00	1.28	-0.01	-3.68	3.52
		(0.00)	(0.07)	(0.18)	(0.70)	(0.66)
95	190.5	0.00	1.25	0.00	-3.50	3.45
		(0.00)	(0.07)	(0.17)	(0.64)	(0.66)
85	171.5	-0.00	1.24	0.01	-3.47	3.36
		(0.00)	(0.06)	(0.18)	(0.66)	(0.64)
70	141.6	0.00	1.20	0.06	-3.16	3.24
		(0.00)	(0.07)	(0.28)	(0.91)	(0.59)
50	98.7	-0.00	1.13	0.10	-2.75	3.03
		(0.00)	(0.06)	(0.25)	(0.32)	(0.55)
95	190.5	0.04	1.26	0.00	-3.47	3.55
Corrected		(0.01)	(0.07)	(0.19)	(0.69)	(0.66)
85	171.5	0.10	1.26	0.01	-3.38	3.60
Corrected		(0.02)	(0.07)	(0.21)	(0.76)	(0.66)
70	141.6	0.17	1.23	0.03	-3.09	3.57
Corrected		(0.03)	(0.08)	(0.29)	(0.98)	(0.67)
50	98.7	0.26	1.20	0.04	-2.61	3.41
Corrected		(0.12)	(0.07)	(0.27)	(0.40)	(0.66)

Table G.76: Means, Standard Deviations, Skewnesses, Minima and Maxima of Reduced Major Axis Regression Residuals of the Background Data for Data Set Structure # 17

Background (%)	$\bar{n}$	Mean	Standard Deviation	Skewness	Minimum	Maximum
Sample	200.0	0.00	1.29	0.00	-3.71	3.67
		(0.00)	(0.08)	(0.20)	(0.72)	(0.57)
95	190.5	0.00	1.27	-0.00	-3.55	3.60
		(0.00)	(0.07)	(0.20)	(0.74)	(0.62)
85	171.5	0.00	1.27	0.01	-3.51	3.56
		(0.00)	(0.07)	(0.21)	(0.78)	(0.61)
70	141.6	0.00	1.23	0.02	-3.30	3.38
		(0.00)	(0.08)	(0.28)	(0.98)	(0.68)
50	98.7	-0.00	1.19	0.02	-2.91	3.07
		(0.00)	(0.07)	(0.27)	(0.36)	(0.65)
95	190.5	0.02	1.28	0.00	-3.54	3.66
Corrected		(0.00)	(0.07)	(0.21)	(0.76)	(0.61)
85	171.5	0.06	1.29	0.01	-3.48	3.70
Corrected		(0.01)	(0.07)	(0.22)	(0.82)	(0.61)
70	141.6	0.09	1.27	0.03	-3.28	3.63
Corrected		(0.02)	(0.09)	(0.27)	(0.99)	(0.67)
50	98.7	0.14	1.23	0.03	-2.84	3.39
Corrected		(0.06)	(0.08)	(0.26)	(0.39)	(0.68)

Table G.77: Means, Standard Deviations, Skewnesses, Minima and Maxima of *PC* # 2 Scores of the Background Data for Data Set Structure # 17

Background (%)	$\bar{n}$	Mean	Standard Deviation	Skewness	Minimum	Maximum
Sample	200.0	0.00	6.08	0.00	-16.16	16.22
		(0.00)	(0.32)	(0.21)	(2.49)	(1.64)
95	190.5	0.00	5.98	-0.01	-15.28	15.39
		(0.00)	(0.27)	(0.18)	(1.93)	(1.51)
85	171.5	0.00	5.70	-0.10	-15.21	14.12
		(0.00)	(0.21)	(0.23)	(1.99)	(1.74)
70	141.6	0.00	5.37	-0.22	-14.95	12.12
		(0.00)	(0.24)	(0.26)	(2.11)	(1.62)
50	98.7	0.00	4.99	-0.28	-13.77	11.83
		(0.00)	(0.67)	(0.26)	(1.99)	(2.86)
95	190.5	0.00	6.02	0.01	-15.93	15.56
Corrected		(0.00)	(0.29)	(0.22)	(2.59)	(1.80)
85	171.5	0.00	5.87	0.04	-15.21	14.97
Corrected		(0.00)	(0.25)	(0.26)	(2.72)	(2.18)
70	141.6	0.00	5.80	0.09	-14.70	14.52
Corrected		(0.00)	(0.35)	(0.30)	(3.02)	(2.40)
50	98.7	0.00	5.54	0.08	-13.09	13.99
Corrected		(0.00)	(0.55)	(0.32)	(2.62)	(3.06)

Table G.78: Means, Standard Deviations, Skewnesses, Minima and Maxima of *PC* # 3 Scores of the Background Data for Data Set Structure # 17

Background (%)	$\bar{n}$	Mean	Standard Deviation	Skewness	Minimum	Maximum
Sample	200.0	0.00	1.12	0.01	-3.13	3.23
		(0.00)	(0.06)	(0.19)	(0.54)	(0.62)
95	190.5	0.00	1.11	0.00	-3.09	3.09
		(0.00)	(0.06)	(0.19)	(0.56)	(0.62)
85	171.5	0.00	1.10	-0.00	-3.01	3.07
		(0.00)	(0.06)	(0.20)	(0.57)	(0.63)
70	141.6	0.00	1.07	-0.04	-2.88	2.84
		(0.00)	(0.06)	(0.29)	(0.57)	(0.84)
50	98.7	0.00	1.03	-0.06	-2.68	2.50
		(0.00)	(0.05)	(0.26)	(0.53)	(0.28)
95	190.5	0.00	1.11	0.00	-3.12	3.09
Corrected		(0.00)	(0.06)	(0.20)	(0.55)	(0.64)
85	171.5	0.00	1.11	-0.01	-3.11	3.06
Corrected		(0.00)	(0.06)	(0.22)	(0.54)	(0.68)
70	141.6	0.00	1.09	-0.03	-3.02	2.89
Corrected		(0.00)	(0.07)	(0.28)	(0.59)	(0.85)
50	98.7	0.00	1.06	-0.03	-2.81	2.55
Corrected		(0.00)	(0.06)	(0.26)	(0.57)	(0.31)

Table G.79: Means, Standard Deviations, Skewnesses, Minima and Maxima of *PC* # 2 and # 3 Radial Distance of the Background Data for Data Set Structure # 17

Background (%)	$\bar{n}$	Mean	Standard Deviation	Skewness	Minimum	Maximum
Sample	200.0	0.80	0.55	1.06	0.04	2.87
		(0.05)	(0.03)	(0.25)	(0.02)	(0.33)
95	190.5	0.81	0.55	1.06	0.04	2.95
		(0.05)	(0.02)	(0.24)	(0.02)	(0.36)
85	171.5	0.81	0.55	1.05	0.04	2.91
		(0.05)	(0.03)	(0.28)	(0.02)	(0.41)
70	141.6	0.83	0.57	1.09	0.04	2.98
		(0.05)	(0.03)	(0.32)	(0.02)	(0.52)
50	98.7	0.86	0.59	1.01	0.05	2.85
		(0.06)	(0.03)	(0.20)	(0.03)	(0.39)
95	190.5	0.83	0.57	1.06	0.04	3.05
Corrected		(0.06)	(0.04)	(0.25)	(0.02)	(0.41)
85	171.5	0.86	0.59	1.05	0.04	3.15
Corrected		(0.09)	(0.06)	(0.30)	(0.03)	(0.58)
70	141.6	0.92	0.65	1.07	0.04	3.43
Corrected		(0.13)	(0.10)	(0.33)	(0.02)	(0.75)
50	98.7	1.06	0.76	1.01	0.06	3.60
Corrected		(0.38)	(0.27)	(0.21)	(0.03)	(1.31)

Table G.80: Means, Standard Deviations, Skewnesses, Minima and Maxima of Ordinary Least Squares Regression Residuals of the Truncated Data for Data Set Structure # 23

Truncation (%)	$\bar{n}$	Mean	Standard Deviation	Skewness	Minimum	Maximum
5	11.1	-7.81	1.57	-0.23	-10.70	-5.36
		(0.47)	(0.36)	(0.38)	(1.17)	(0.93)
15	32.1	-6.91	1.76	-0.66	-11.35	-4.05
		(0.25)	(0.22)	(0.37)	(1.17)	(0.61)
30	61.4	-6.39	2.04	-0.80	-12.27	-3.23
		(0.26)	(0.22)	(0.24)	(1.12)	(0.69)
50	99.3	-6.24	2.38	-0.74	-13.43	-2.37
		(0.34)	(0.18)	(0.22)	(1.07)	(0.64)
5	11.1	-7.26	1.69	-0.16	-10.30	-4.52
Corrected		(0.51)	(0.40)	(0.35)	(1.17)	(1.09)
15	32.1	-5.74	1.93	-0.49	-10.35	-2.32
Corrected		(0.34)	(0.24)	(0.32)	(1.28)	(0.87)
30	61.4	-4.37	2.22	-0.57	-10.42	-0.35
Corrected		(0.24)	(0.25)	(0.22)	(1.17)	(1.03)
50	99.3	-3.22	2.61	-0.45	-10.75	2.14
Corrected		(0.43)	(0.24)	(0.28)	(1.27)	(1.33)

Table G.81: Means, Standard Deviations, Skewnesses, Minima and Maxima of Reduced Major Axis Regression Residuals of the Truncated Data for Data Set Structure # 23

Truncation (%)	$\bar{n}$	Mean	Standard Deviation	Skewness	Minimum	Maximum
5	11.1	-5.54	4.52	-0.01	-12.89	2.15
		(1.22)	(1.95)	(0.18)	(2.74)	(4.34)
15	32.1	-5.78	3.80	-0.16	-14.28	1.88
		(1.13)	(0.86)	(0.31)	(2.46)	(1.81)
30	61.4	-5.65	3.70	-0.26	-15.44	2.10
		(0.86)	(0.83)	(0.23)	(2.80)	(1.19)
50	99.3	-5.15	4.60	-0.22	-18.44	5.52
		(1.37)	(4.15)	(0.21)	(11.04)	(11.10)
5	11.1	-5.19	4.32	0.04	-12.13	2.44
Corrected		(0.91)	(1.18)	(0.22)	(2.36)	(3.24)
15	32.1	-4.81	3.88	-0.14	-13.32	3.03
Corrected		(1.18)	(0.56)	(0.30)	(2.83)	(1.29)
30	61.4	-3.88	4.06	-0.18	-14.41	4.78
Corrected		(0.74)	(0.61)	(0.34)	(2.30)	(1.70)
50	99.3	-3.07	4.02	-0.04	-14.08	6.84
Corrected		(0.88)	(0.49)	(0.27)	(2.35)	(1.53)

Table G.82: Means, Standard Deviations, Skewnesses, Minima and Maxima of PC # 2 Scores of the Truncated Data for Data Set Structure # 23

Truncation (%)	$\bar{n}$	Mean	Standard Deviation	Skewness	Minimum	Maximum
5	11.1	4.97	5.35	0.02	-3.96	13.80
		(2.12)	(1.08)	(0.44)	(2.34)	(2.84)
15	32.1	3.57	5.62	-0.04	-8.28	15.94
		(1.94)	(0.75)	(0.39)	(3.76)	(2.88)
30	61.4	3.22	5.66	-0.06	-10.72	16.33
		(1.11)	(0.35)	(0.47)	(2.68)	(2.96)
50	99.3	3.04	5.77	0.02	-11.39	18.10
		(1.16)	(0.28)	(0.31)	(3.25)	(2.32)
5	11.1	5.32	5.28	0.01	-3.53	13.96
Corrected		(2.12)	(1.07)	(0.42)	(2.34)	(2.87)
15	32.1	4.20	5.51	-0.03	-7.53	16.32
Corrected		(1.82)	(0.78)	(0.37)	(3.53)	(2.66)
30	61.4	3.98	5.53	-0.05	-9.39	16.83
Corrected		(1.13)	(0.37)	(0.44)	(2.41)	(2.48)
50	99.3	3.93	5.71	0.06	-9.94	18.76
Corrected		(1.16)	(0.27)	(0.28)	(3.11)	(2.00)

Table G.83: Means, Standard Deviations, Skewnesses, Minima and Maxima of *PC* # 3 Scores of the Truncated Data for Data Set Structure # 23

Truncation (%)	$\bar{n}$	Mean	Standard Deviation	Skewness	Minimum	Maximum
5	11.1	7.34	1.78	0.06	4.35	10.42
		(0.66)	(0.38)	(0.32)	(1.20)	(1.17)
15	32.1	6.67	1.83	0.55	3.51	11.08
		(0.38)	(0.20)	(0.35)	(0.83)	(1.18)
30	61.4	6.25	2.06	0.74	2.90	12.06
		(0.30)	(0.22)	(0.24)	(0.79)	(1.14)
50	99.3	6.16	2.38	0.71	2.16	13.26
		(0.34)	(0.18)	(0.24)	(0.80)	(1.06)
5	11.1	6.92	2.04	-0.01	3.38	10.35
Corrected		(0.77)	(0.44)	(0.28)	(1.44)	(1.26)
15	32.1	6.13	2.18	0.33	2.10	11.07
Corrected		(0.65)	(0.25)	(0.33)	(1.25)	(1.33)
30	61.4	5.57	2.42	0.39	0.91	11.85
Corrected		(0.55)	(0.30)	(0.26)	(1.44)	(1.26)
50	99.3	5.33	2.75	0.30	-0.74	13.10
Corrected		(0.64)	(0.30)	(0.30)	(1.85)	(1.21)

Table G.84: Means, Standard Deviations, Skewnesses, Minima and Maxima of *PC* # 2 and # 3 Radial Distance of the Truncated Data for Data Set Structure # 23

Truncation (%)	$\bar{n}$	Mean	Standard Deviation	Skewness	Minimum	Maximum
5	11.1	0.69	0.13	0.50	0.52	0.94
		(0.08)	(0.05)	(0.39)	(0.06)	(0.20)
15	32.1	0.78	0.20	0.65	0.47	1.27
		(0.12)	(0.05)	(0.36)	(0.07)	(0.33)
30	61.4	0.91	0.29	0.82	0.49	1.74
		(0.14)	(0.07)	(0.24)	(0.08)	(0.40)
50	99.3	1.21	0.45	0.77	0.51	2.60
		(0.19)	(0.10)	(0.18)	(0.10)	(0.59)
5	11.1	0.57	0.12	0.53	0.42	0.80
Corrected		(0.07)	(0.05)	(0.43)	(0.06)	(0.19)
15	32.1	0.53	0.15	0.58	0.29	0.91
Corrected		(0.09)	(0.04)	(0.37)	(0.06)	(0.27)
30	61.4	0.49	0.17	0.76	0.23	0.98
Corrected		(0.07)	(0.05)	(0.17)	(0.05)	(0.26)
50	99.3	0.51	0.21	0.64	0.13	1.16
Corrected		(0.08)	(0.05)	(0.24)	(0.04)	(0.32)

Table G.85: Means, Standard Deviations, Skewnesses, Minima and Maxima of Ordinary Least Squares Regression Residuals of the Background Data for Data Set Structure # 23

Background (%)	$\bar{n}$	Mean	Standard Deviation	Skewness	Minimum	Maximum
Sample	200.0	-0.00	3.83	-0.03	-10.21	10.27
		(0.00)	(0.20)	(0.16)	(1.13)	(1.24)
95	188.9	0.00	3.47	0.22	-7.22	9.81
		(0.00)	(0.21)	(0.14)	(0.62)	(1.38)
85	167.9	-0.00	3.07	0.43	-5.67	9.16
		(0.00)	(0.26)	(0.14)	(0.65)	(1.47)
70	138.6	-0.00	2.71	0.58	-4.63	8.40
		(0.00)	(0.23)	(0.15)	(0.42)	(1.54)
50	100.7	0.00	2.31	0.75	-3.80	7.32
		(0.00)	(0.21)	(0.18)	(0.57)	(1.42)
95	188.9	0.39	3.48	0.20	-7.05	10.13
Corrected		(0.03)	(0.21)	(0.14)	(0.71)	(1.31)
85	167.9	0.97	3.10	0.36	-5.33	10.05
Corrected		(0.12)	(0.26)	(0.12)	(0.93)	(1.39)
70	138.6	1.74	2.78	0.41	-3.97	9.99
Corrected		(0.21)	(0.23)	(0.13)	(0.82)	(1.45)
50	100.7	2.65	2.48	0.40	-2.75	9.92
Corrected		(0.33)	(0.22)	(0.22)	(1.45)	(1.26)

Table G.86: Means, Standard Deviations, Skewnesses, Minima and Maxima of Reduced Major Axis Regression Residuals of the Background Data for Data Set Structure # 23

Background (%)	$\bar{n}$	Mean	Standard Deviation	Skewness	Minimum	Maximum
Sample	200.0	-0.00	4.93	-0.03	-14.09	12.92
		(0.00)	(0.77)	(0.19)	(2.67)	(1.92)
95	188.9	0.00	5.15	0.00	-14.37	13.55
		(0.00)	(1.89)	(0.21)	(6.15)	(4.31)
85	167.9	0.00	4.36	0.07	-11.54	11.82
		(0.00)	(0.96)	(0.19)	(3.64)	(2.79)
70	138.6	0.00	3.91	0.05	-9.64	10.35
		(0.00)	(0.82)	(0.19)	(2.33)	(2.27)
50	100.7	-0.00	4.29	0.08	-11.28	10.98
		(0.00)	(3.86)	(0.23)	(12.50)	(9.81)
95	188.9	0.31	4.94	-0.00	-13.31	13.12
Corrected		(0.04)	(1.14)	(0.21)	(4.09)	(2.33)
85	167.9	0.75	4.41	0.04	-11.02	12.34
Corrected		(0.14)	(0.61)	(0.18)	(2.18)	(1.42)
70	138.6	1.33	4.27	-0.02	-9.89	12.03
Corrected		(0.23)	(0.57)	(0.18)	(1.56)	(1.35)
50	100.7	1.97	3.86	-0.05	-8.24	11.24
Corrected		(0.37)	(0.39)	(0.17)	(1.64)	(1.11)

Table G.87: Means, Standard Deviations, Skewnesses, Minima and Maxima of *PC* # 2 Scores of the Background Data for Data Set Structure # 23

Background (%)	$\bar{n}$	Mean	Standard Deviation	Skewness	Minimum	Maximum
Sample	200.0	0.00	6.08	-0.11	-17.67	16.56
		(0.00)	(0.23)	(0.14)	(2.05)	(1.58)
95	188.9	0.00	5.98	-0.12	-17.40	16.42
		(0.00)	(0.17)	(0.15)	(2.23)	(2.38)
85	167.9	0.00	5.94	-0.13	-17.12	15.83
		(0.00)	(0.20)	(0.10)	(2.36)	(1.85)
70	138.6	0.00	5.94	-0.09	-16.47	15.60
		(0.00)	(0.23)	(0.09)	(2.29)	(1.77)
50	100.7	0.00	5.89	-0.13	-15.88	13.77
		(0.00)	(0.23)	(0.22)	(2.12)	(2.46)
95	188.9	0.00	5.97	-0.13	-17.39	16.37
Corrected		(0.00)	(0.18)	(0.14)	(2.16)	(2.30)
85	167.9	0.00	5.92	-0.15	-17.12	15.75
Corrected		(0.00)	(0.20)	(0.09)	(2.19)	(1.71)
70	138.6	0.00	5.90	-0.10	-16.39	15.52
Corrected		(0.00)	(0.24)	(0.10)	(2.25)	(1.65)
50	100.7	0.00	5.87	-0.10	-15.27	13.65
Corrected		(0.00)	(0.23)	(0.21)	(2.34)	(2.48)

Table G.88: Means, Standard Deviations, Skewnesses, Minima and Maxima of *PC* # 3 Scores of the Background Data for Data Set Structure # 23

Background (%)	$\bar{n}$	Mean	Standard Deviation	Skewness	Minimum	Maximum
Sample	200.0	0.00	3.72	0.04	-9.89	9.95
		(0.00)	(0.21)	(0.15)	(1.19)	(1.21)
95	188.9	0.00	3.41	-0.18	-9.45	7.58
		(0.00)	(0.22)	(0.13)	(1.29)	(0.71)
85	167.9	0.00	3.03	-0.38	-8.87	5.94
		(0.00)	(0.26)	(0.14)	(1.38)	(0.87)
70	138.6	0.00	2.69	-0.54	-8.23	4.80
		(0.00)	(0.22)	(0.14)	(1.46)	(0.49)
50	100.7	0.00	2.30	-0.72	-7.23	3.93
		(0.00)	(0.20)	(0.19)	(1.36)	(0.66)
95	188.9	0.00	3.43	-0.13	-9.46	8.08
Corrected		(0.00)	(0.22)	(0.13)	(1.22)	(0.85)
85	167.9	0.00	3.10	-0.24	-8.77	7.08
Corrected		(0.00)	(0.27)	(0.13)	(1.15)	(1.33)
70	138.6	0.00	2.84	-0.24	-8.07	6.52
Corrected		(0.00)	(0.25)	(0.15)	(1.17)	(1.08)
50	100.7	0.00	2.59	-0.20	-7.08	6.23
Corrected		(0.00)	(0.26)	(0.24)	(0.85)	(1.46)

Table G.89: Means, Standard Deviations, Skewnesses, Minima and Maxima of *PC* # 2 and # 3 Radial Distance of the Background Data for Data Set Structure # 23

Background (%)	$\bar{n}$	Mean	Standard Deviation	Skewness	Minimum	Maximum
Sample	200.0	0.29 (0.01)	0.15 (0.01)	0.71 (0.19)	0.03 (0.01)	0.82 (0.13)
95	188.9	0.30 (0.01)	0.16 (0.01)	0.67 (0.21)	0.02 (0.01)	0.86 (0.11)
85	167.9	0.33 (0.02)	0.18 (0.02)	0.82 (0.25)	0.03 (0.01)	1.01 (0.17)
70	138.6	0.36 (0.02)	0.20 (0.02)	1.01 (0.26)	0.03 (0.02)	1.16 (0.15)
50	100.7	0.40 (0.03)	0.24 (0.02)	1.27 (0.26)	0.03 (0.03)	1.38 (0.14)
95 Corrected	188.9	0.27 (0.01)	0.14 (0.01)	0.66 (0.23)	0.02 (0.01)	0.74 (0.10)
85 Corrected	167.9	0.26 (0.02)	0.13 (0.02)	0.75 (0.21)	0.02 (0.01)	0.74 (0.13)
70 Corrected	138.6	0.24 (0.02)	0.13 (0.02)	0.74 (0.22)	0.03 (0.01)	0.68 (0.12)
50 Corrected	100.7	0.24 (0.02)	0.13 (0.02)	0.84 (0.28)	0.03 (0.01)	0.67 (0.15)

Clifford R. Stanley

Graduate Awards :

Hugo E. Meilicke Graduate Fellowship (1987)
Doctoral University Graduate Fellowship (1985, 1986)
Aaro E. Aho Memorial Scholarship (1984)

Publications :

Stanley, Clifford R., and Sinclair, Alastair J. (1987): **Anomaly Recognition in Multi-Element Geochemical Data - A Background Characterization Approach**, Journal of Geochemical Exploration, Vol. 29, pp. 333-353.

Stanley, Clifford R. (1987): **PROBPLOT - An Interactive Computer Program to Fit Mixtures of Normal (or Log-Normal) Distributions using Maximum Likelihood Optimization Procedures**, Special Volume # 14, Association of Exploration Geochemists, Rexdale, Ont., 40 p., 1 diskette.

Stanley, Clifford R. (1987): **Hinsdalite and Other Products of Alteration at the Daisy Creek Stratabound Copper-Silver Prospect, Montana**, Canadian Mineralogist, Vol. 25, Pt. 2, June, pp. 213-220.

Stanley, Clifford R. and Sinclair, Alastair J. (1988): **Univariate Patterns in the Design of Multivariate Analysis Techniques for Geochemical Data Evaluation**, in "Quantitative Analysis of Mineral and Energy Resources", *NATO Advanced Study Institute Series C : Mathematical and Physical Sciences*, Chung, C.F., Fabbri, A.G. and Sinding-Larsen, R. - eds., Vol. 223, Reidel Publishing Co., Boston, pp. 113-130.

Stanley, Clifford R. and Sinclair, Alastair J. (1988): **Sedimentologic Setting and Stratigraphic Correlation of Stratabound Cu-Ag Deposits of the Bonner Quartzite, Belt Supergroup, Western Montana**, *(In Press)*, Special Volume on Stratiform Copper Deposits, Canadian Mineralogist.

Stanley, Clifford R. and Russell, J.K. (In Press): **PEARCE.PLOT : Interactive Graphics-Supported Software for Testing Petrologic Hypotheses with Pearce Element Ratio Diagrams.** American Mineralogist.

Stanley, Clifford R. and Russell, J.K. (In Press): **PEARCE.PLOT : A Turbo-Pascal Program for the Analysis of Rock Compositions with Pearce Element Ratio Diagrams.** Computers and Geosciences.

Stanley, Clifford R. and Russell, J.K. (In Review): **Petrologic Hypothesis Testing with Pearce Element Ratio Diagrams : Derivation of Diagram Axes.** Geochimica et Cosmochimica Acta.