MML ESTIMATORS AND ROBUST CLASSIFICATION AND LINEAR REGRESSION PROCEDURES

By

R.S. AMBAGASPITIYA, M. Sc. (McMaster University)

A Thesis
Submitted to the School of Graduate Studies
in Partial Fulfilment of the Requirements
for the Degree
Doctor of Philosophy

McMaster University

MML estimators and robust classification and linear regression procedures

DOCTOR OF PHILOSOPHY (1991) MCMASTER UNIVERSITY
(Mathematics) Hamilton, Ontario

TITLE: MML estimators and robust classification and linear regression procedures

AUTHOR: R.S. Ambagaspitiya
M. Sc. (McMaster University)

SUPERVISOR: Dr. N. Balakrishnan

NUMBER OF PAGES: xxi, 345

Contents

Abstract

Statistical classification and regression are two widely used statistical techniques. The former is concerned with the problem of classifying an object of unknown origin into one of two or more distinct groups of populations on the basis of observations made on it. The broad appeal of regression techniques results from the conceptually simple process of using an equation to express the relationship between a set of variables. These techniques are said to be resistant when the result is not greatly altered in the case a small fraction of the data is altered: techniques are said to be robust of efficiency when their statistical efficiency remains high for more realistic than the utopian cases of normal distributions. These properties are particularly important in the formative stages of model building when the form of the response is not known exactly. Techniques with these properties are proposed and discussed.

First, we present Tiku's (1967, 1970, 1980) univariate MML (Modified Maximum Likelihood) estimation method based on Type-II symmetrically censored normal samples. Next, we describe the extension of this method of estimation as given by Tiku (1988) for the bivariate case and by Tiku and Balakrishnan (1988) for the general multivariate case. Some asymptotic distributional properties of these multivariate MML estimators are also discussed. Then, we develop a robust multivariate linear two-way classification procedure based on the MML (Modified Maximum Likelihood) estimators and demonstrate that it performs overall more efficiently than the classical linear two-way classification procedure, for a fixed value of one of the two errors of misclassification. In the case when both the errors of misclassification are allowed to float, we also show that the robust procedure has a smaller error rate than the classical procedure and a much smaller error rate than non-parametric

classification procedures like the nearest neighbour method and the method based on density estimates. For this vivid comparative study examining the performance of these various classification procedures, we use the data obtained from a very elaborative and extensive anthropometric survey conducted by Majumdar (1941) in the United Provinces of India.

Next, the classification procedures where the classification has to be based on the observed value of a pair of variables, one being a dichotomous random variable (univariate or multivariate) and the other being univariate or multivariate continuous variable have been studied. For this study, we consider the existing classification procedures due to Chang and Afifi (1974) and Balakrishnan and Tiku (1988) and extended the latter method to the case when the data contain a multivariate dichotomous and an associated multivariate continuous variable. We illustrate the non-robust characteristic of the Chang and Afifi (1974) procedure through a study of the distribution functions and expected values of errors of misclassification when both errors are allowed to float. After showing that the linear classification procedure could be sensitive to departures from the homogeneity of variances, we propose quadratic and transformed linear classification procedures for the problem of classification based on dichotomous and continuous variables with the populations differing from each other not only in means but also in variances.

Next, we adopt the modified maximum likelihood approach and derive a robust method of estimation of the parameters in a simple linear regression model. We derive asymptotic variances and covariances of these estimators via the information matrix. Then we compare the performance of this procedure based on MML estimators with many prominent procedures under both normal as well as a wide range of non-normal models for the error variable and also under departures from linearity in the simple linear regression model. We present two examples to illustrate the various methods of estimation considered in this study. Finally, we extend these results to the case of multiple linear regression models. Once again, through a comparative study we show that the MML estimators derived in this chapter are quite robust to departures from normality and remain highly efficient under normal and several non-normal models for the error variable.

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. N.Balakrishnan, for his constant guidance, unfailing support and encouragement during the preperation of this thesis. I would also like to thank Dr. J. Chadam, Dr. P.D.M. Macdonald, Dr. A.C. Cohen at the University of Georgia, Dr. K. Fung at the University of Windsor and Dr. S. Panchapakesan at the Southern Illinois University for reading the first draft of the thesis and giving some useful comments.

I am also grateful to McMaster University for providing financial aid to carry out my graduate studies.

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Historical background

In the past, the idea of measurement was associated almost entirely with scientific pursuit, by which was meant a preoccupation with the physical sciences. These admit very strict and exactly reproducible laws, the determination of which requires very precise measuring instruments. Much effort was therefore expended in refining instrumentation to achieve the required degree of precision, to enable such laws to be detected. Scientists were nevertheless aware that, no matter how precise the instruments that were used, some error of measurement is inevitable, and hence there was a need to allow for such error in scientific calculation. From these beginnings, via the work of such pioneers as Gauss and Laplace on theories of errors, developed the subject of Statistics. With greater knowledge about the likely behaviour of errors, scientists became more confident in their ability to separate measurements from errors. The process of measurement and quantification thus spread into less exact sciences, gradually permeating through the Biological and then the Social Sciences. This led to obtaining numerical information of all kinds and it was supported by the increasing power of computer technology to process these information. We now live in an age where our day-to-day life is filled with numerical and statistical information such as polls, Nielsen ratings of TV programmes and sports news.

A consequence of this development is the application of statistical analysis by

researchers in nearly any discipline to analyze their numerical data. Classification and the regression analysis are two important topics in statistical data analysis.

1.2 Classification

Statistical classification is one of the widest used statistical techniques and it is concerned with the problem of classifying an object of unknown origin into one of two or more distinct groups of populations on the basis of observations made on it. In the literature we find many names for this general area of problems; for example, allocation, identification, prediction, pattern recognition, selection, besides the standard terms, such as, classification and discrimination. We elaborate on the successful usage of statistical classification in various fields such as medicine, anthropology and mining, using the examples given by Fatti (1982) and Hand (1981).

Some examples

1. Haemophilia is a sex-linked genetic disease which is transmitted only by females,but whose symptoms are manifest only in males. Under normal medical examination it is impossible to distinguish between female carrying the disease and those not. In order to try and identify female carriers, the levels of coagulant factor and its related antigen (Factor VIII and Factor VIII RA) in the blood have been suggested as possible discriminators between carriers and non-carriers. A pilot study was carried out by Gomperts et al (1976) to test how well Factor VIII and its related antigen discriminate between carriers and non-carriers. A sample of 26 white females, of which 11 were known, for genetic reasons, to be carriers and 15 were known to be non-carriers was selected and the Factor VIII and Factor VIII RA levels measured in each subject. Using the linear classification rule based on the logarithms of the data and the jackknife reclassification procedure, ten out of the eleven carriers and thirteen out of the fifteen non-carriers were classified correctly. In a parallel study on black females (whose Factor VIII and Factor VIII RA levels tend to

be different from the whites) all ten carriers and fourteen out of fifteen non-carriers were correctly classified. Thus it would appear that the carriers of haemophilia can be identified, reasonably reliably, on the basis of their levels of Factor VIII and Factor VIII RA.

2. A common problem occurring in anthropology is that of identifying the tribe, race or even sex of cranium excavated amongst the remains of an ancient civilization (for example De Villers, 1976). By comparing various measurements (lengths and angles) made on the skull with those made on large numbers of individuals, males and females, from the various tribes at present inhabiting the region, it may be classified to the tribe from which it most probably came. In this problem, we might also be interested in the possibility that it did not originate from any of these tribes, but from another unknown tribe (possibly now extinct). Another possibility is of its occupying some position intermediate between the tribes which could result from intermarriage between members of the different tribes.

   A problem arising frequently in this type of application is that of choosing, amongst the large number of possible measurements that can be made on the skull, that subset giving the best discrimination between the various populations.

3. One of the most recent applications has been to the classification of crops from high altitude photographs. It can be cheaper to estimate total acreage this way than by ground measurement, and it is also possible to identify incipient crop disease before it can be recognized from the ground. Similar methods have been applied to the detection of mineral deposits.

4. Medical diagnosis is a particularly fruitful area of application. The methods have been applied to the assessment of the prognostic value of tests of lung function in miners with pneumoconiosis, to selecting appropriate operations for breast cancer patients, to predicting ischaemic heart disease, to predicting relapse in pulmonary tuberculosis sufferers, etc. The list is virtually endless.

5. Amongst the recent applications which will have a very direct effect on our everyday life are those of speech recognition and optical character recognition. Speech recognition is an example of a case where the objects to be classified are waveforms. Other such examples are cardiac wave analysis, target recognition from vehicle noise or radar returns, and automatic encephalograph analysis.

6. Personnel classification is an area of increasing though perhaps controversial importance. The methods have been applied in vocational guidance, academic attainment, and even in recognizing high-risk individuals in psychiatric research.

7. Other areas of application include ecology, taxonomy, author verification, psychology, fingerprint recognition , etc. Lachenbruch (1975) lists 579 references.

The origin of statistical classification is fairly old, and its development reflects the same broad phases as that of general statistical inference, namely, a Pearsonian phase followed by Fisherian, Neyman-Pearsonian, and Waldian phases. Das Gupta (1973) has given a comprehensive review of this developments. For the clarity of presentation we include selected parts of Das Gupta's (1973) article in Section 1.2.1 to Section 1.2.6.

1.2.1 Early History

In the first survey of discriminatory analysis, Hodges (1950) aptly mentioned the following.

> In his invited address at the meeting of the Institute of Mathematical Statistics in Berkeley, California, June 16, 1949, Professor M. A. Girshick pointed out that the development of discriminatory analysis reflects the same broad phases as does the general history of statistical inference. We may distinguish a Pearsonian stage,..., followed by a Fisherian stage. Professor Girshick further notes a Neyman-Pearson stage and a contemporary Waldian stage...

In the early work, the classification problem was noted precisely formulated and often confounded with the problem of testing the equality of two or more distributions; the term "discriminatory analysis" was used for both. In practice, the following scheme was generally followed for the two-population classification problem. Suppose we have three distributions, F, F_1, F_2 and T_i is a test statistic designed to test the hypothesis $F = F_i$ $(i = 1, 2)$. The decision $F = F_i$ is taken if T_i is the smaller of T_1 and T_2; sometimes the critical values of T_i's are compared in order to take decision. Thus statistics for testing the equality of two distributions played an important role. Generally, such a test statistics may be considered as a measure of divergence between two distributions. Karl Pearson (in a paper by Tildesley (1921)) proposed one such measure, termed as the "coefficient of racial likeness (CRL)." This was modified by Morant in 1928 and by Mahalanobis in 1927 and 1930. Mahalanobis called his measure D^2 and suggested (1930) also some measure of divergence in variability, skewness and kurtosis and studied the distributions of these measures. In 1926, Pearson published the first considerable theoretical work on the CRL and suggested the following form for the coefficient when the variables are dependent:

$$\frac{n_1 n_2}{n_1 + n_2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), \tag{1.1}$$

where $\bar{\mathbf{x}}_i$ is the sample mean vector based on a sample of size n_i from the ith population $(i = 1, 2)$ and $\mathbf{S}$ is the pooled sample covariance matrix. In 1936, Mahalanobis gave the dependent-variate versions of his D^2- statistic in the classical and the studentized forms. The distributions of these statistics were studied by Bose (1936a, b), Bose (1936c), Bose and Roy (1938), and Bhattacharya and Narayana (1941). In 1931 Hotelling suggested a test statistic T^2 which is a constant multiple of the studentized Mahalanobis D^2 and obtained its null distribution.

Hodges remarked that "the first clear statement of the problem of discrimination, and the first proposed solution to that problem were given by Fisher in the middle of the 1930's... the ideas of Fisher first appeared in print in papers by other people (Barnard, 1935; Martin, 1936)." Earlier than this, Morant (1926) considered the problem of classifying a skull into Eskimo or modern English groups by two sets of tests. Fisher's own first work on the subject appeared in his paper in 1936. For

the univariate two-population problem Fisher suggested a rule which classifies the observation x into the ith population if $|x - \bar{x}_i|$ is the smaller of $|x - \bar{x}_1|$ and $|x - \bar{x}_2|$. For a $p-$ component observation vector ($p > 1$), Fisher reduced the problem to the univariate one by considering an "optimum" linear combination (called the "linear discriminant function") of the p components. For a given linear combination of Y of the p components, Fisher considered the ratio between the difference in the sample means of the $Y-$values and the standard error within samples of the Y-values and maximizes this ratio in order to define the optimum linear combination. It turns out that the coefficients of this optimum linear combination are proportional to $\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$. Incidently, Fisher (1936) suggested a test for the equality of two normal distributions with the same unknown covariance matrix and this test is the same as the one proposed by Hotelling (1931).

The next development was influenced by the pioneering fundamental work by Neyman and Pearson (1933, 1936). For the two-population problem, Welch (1939) derived the forms of Bayes rules and the minimax Bayes rule when the distributions are known; he illustrated the theory with multivariate normal distributions with the same covariance matrix. This example was also considered by Wald (1944), who further proposed some heuristic rules by replacing the unknown parameters by their respective (maximum likelihood) estimates. Wald studied the distribution of the proposed classification statistic. Von Mises (1944) obtained the rule which maximizes the minimum probability of correct classification. The problem of classification into two normal distributions with different covariance matrices was treated by Cavalli (1945) and Penrose (1947) when $p = 1$ and by Smith (1947) for general p. In a series of papers Rao (1946, 1947a, 1947b, 1948, 1949a, 1949b, 1950) suggested different methods of classification into two or more populations following the ideas of Neyman-Pearson and Wald; in particular, Rao suggested a measure of distance between two groups and considered the possibility of withholding decision (through "doubtful" regions) and preferential decisions. Rao's development is for the case when the distributions are all known. General theoretical results on the classification problem (as a special case) in the framework of decision theory are given in the book by Wald (1950) and in a paper by Wald and Wolfowitz (1950).

1.2.2 Classification into two multivariate normal distributions with the same covariance matrix

The distribution of X in Π_i is $p-$variate normal with mean μ_i and variance Σ, ie. $N_p(\mu_i, \Sigma)$, $i = 1, 2$. Suppose all the parameters μ_1 are known. The class of Bayes rules is the same as the class of Likelihood Ratio (LR) rules. Typically, a LR rule δ_c classifies X into $N_p(\mu_1, \Sigma)$, iff

$$T(x) = T(x; \mu_1, \mu_2, \Sigma) = ||x - \mu_1||^2_\Sigma - ||x - \mu_2||^2_\Sigma \leq c,$$

where the L^2 norm is defined by $||a - b||^2_\Sigma = (a - b)^T \Sigma^{-1} (a - b)$. The minimax Bayes rule (0-1 loss) is given by δ_0; it is also called the minimum distance (MD) rule (for Mahalanobis distance). The probability of misclassification of δ_c is given by

$$e_{12}(\delta_c) = \Phi\left(-\frac{c + \Delta^2}{2\Delta}\right), \quad e_{21}(\delta_c) = \Phi\left(\frac{c - \Delta^2}{2\Delta}\right),$$

where $\Delta^2 = ||\mu_1 - \mu_2||^2_\Sigma$, and Φ is the cdf of $N(0, 1)$. This classical case is treated in many papers and books, for example Welch (1939), Wald (1944), Rao (1952) and Anderson (1958) are a few of them. Recall that Fisher's linear discriminant function (LDF) is given by $(\mu_1 - \mu_2)^T \Sigma^{-1} x$ which maximizes $\frac{[a^T(\mu_1 - \mu_2)]^2}{a^T \Sigma a}$ among all vectors a. Penrose (1947) suggested to consider the best LDF in terms of two linear functions of X given by the sum and a linear contrast of the components of X expressed in terms of their standard deviations; he called them the "size" and the "shape" respectively. He discussed the case when all the correlations are equal. If the unknown parameters are structured in a special way, reasonable rules based on X can be found. For instance, Rao (1966) considered the following structure, relevant to growth models: $\mu_i = \nu_i + \beta\theta_i$ $(i = 1, 2)$, where ν_i, β are known but the vectors θ_i are unknown. By restricting to similar divisions of the sample space or by considering ancillary statistics, the problem is reduced to finding the usual LDF in terms of the projection of X on a space orthogonal to the column-space of β. This problem was originally posed by Burnaby (1966). Rao also treated the case when the covariance matrices are different.

Cochran (1961, 1964) studied the effects of the different components of X

on Δ^2, which determines the probability of misclassification (PMC) of a LR rule, especially when all the correlations are equal.

When all the parameters are not known, random samples of sizes n_1 and n_2 from $N_p(\mu_1, \Sigma)$ and $N_p(\mu_2, \Sigma)$, repectively, are used to get information on the parameters. The literature in this area consists of

1. suggestions of some heuristic rules, especially the rules obtained by substituting the estimators for mean and variance to LR rules,
2. distributions of classification statistics and expressions for PMC,
3. estimations of the PMC of a given rule, and
4. derivation of constructive rules.

The rules considered in the literature are usually of the type involving a classification statistic Z and a cut-off point c (i.e., classifies into Π_1, iff $Z < c$), where Z is function of X, $\bar{X}_1$, $\bar{X}_2$ and S, where $\bar{X}_1$ and $\bar{X}_2$ are the sample mean vectors and S is the sample pooled covariance matrix (the divisor being $n_1 + n_2 - 2 = r$). The plug-in version of δ_c, denoted by $\hat{\delta}_c$, is based on the statistic $W = T(X; \bar{X}_1, \bar{X}_2, S)$, when all the parameters are unknown. This statistic was proposed by Anderson (1951). More generally, one may consider a plug-in LR rule by replacing the unknown parameters in T by their respective estimates. Fisher (1936) and Wald (1944) suggested the plug-in LDF as the classification statistic, which is given by $U = (\bar{X}_2 - \bar{X}_1)^T S^{-1} X$. Anderson (1951, 1958) proposed the LR rules which have the following classification statistic:

$$\left[\frac{r + (1 + \frac{1}{n_1})^{-1} ||X - \bar{X}_1||_S^2}{r + (1 + \frac{1}{n_2})^{-1} ||X - \bar{X}_2||_S^2}\right].$$

When the cut-off point c is 1, this rule reduces to

$$V = (1 + \frac{1}{n_1})^{-1} ||X - \bar{X}_1||_S^2 - (1 + \frac{1}{n_2})^{-1} ||X - \bar{X}_2||_S^2 \lessgtr 0.$$

This is the same as $\hat{\delta}_0$, when $n_1 = n_2$. For known Σ, the LR rules involves the statistic V with S replaced by Σ. In the sequel, the rule $V \lessgtr 0$ will be called the same terminology when some of the known parameters are used instead of their

estimates. Rao (1954) derived some rules restricting to invariance and local optimal conditions; the classification statistic for his rule (Σ unknown) will be called R. Matusita's (1967) minimum distance rules and reduce to MD rules in this case; Matusita also considered the case when there are n observations to be classified and obtained some lower bounds for the PCC of MD rule.

Rao (1946) suggested to test the hypothesis $\mu = (\mu_1 + \mu_2)/2$ by Hotelling's T^2-test and use the MD rule when this test is significant. Brown (1947) considered a problem where $\mu_i = \alpha + \beta w_i$ $(i = 1, 2)$, w_i being the classificatory variable (e.g., age). α and β can be estimated from a training sample and using these estimates w is estimated for the observation to be classified; Brown extended this to more than 2 populations. Cochran (1964) posed the problem when the last q components of X have the same means in Π_1 and Π_2 and suggested a statistic W^* (similar in form to W) in terms of the residuals in the first $p-q$ components of X after eliminating their (linear) regression on the last q components. Each of the statistics, U, V, W and R can be expressed as linear function of the elements of a 2×2 random matrix

$$M = [m_{ij}] = (Y_1, Y_2)^T A^{-1} (Y_1, Y_2), \tag{1.2}$$

where Y_1, Y_2 are independent $N_p(., I_p)$ vectors, and $A \sim W_p(n_1 + n_2 - 2, I_p)$, Wishart distribution, independently of Y_1 and Y_2. In particular, for V, W and R the means of Y_1 and Y_2 are proportional , and moreover, V is a constant multiple of m_{12}. Wald (1943) gave a canonical representation of U, and Harter (1951) derived the distribution when $p = 1$. Sitgreaves (1952) derived the distribution of M when the means of Y_1 and Y_2 are proportional; Kabe (1963) derived it without this restriction. Bowker (1960) showed that W can be represented as a rational function of two independent 2×2 Wishart matrices one of which is noncentral. Bowker and Sitgreaves (1961) used this representation to find an asymptotic expansion of the cdf of W in terms of n_1^{-1} and Hermite polynomials, when $n_1 = n_2$. Sitgreaves (1961) derived the distribution of m_{12} and explicitly obtained an approximation to the cdf of W for large $n_1 = n_2$ and asymptotic expansion for the average PMC of the MD rule in powers of $(n_1 + n_2)/(n_1 n_2)$ and Hermite polynomials in Δ^2. Teichroew and Sitgreaves (1961) used an empirical sampling plan to obtain an estimate of the cdf

of W. Okamoto (1963) considered the statistic W where the degrees of freedom r of S is not necessarily $n_1 + n_2 - 2$, and gave asymptotic expansions of

$$P\left[\frac{W - \Delta^2/2}{\Delta} < k|\Pi_1\right] \quad \text{and} \quad P\left[\frac{W + \Delta^2/2}{\Delta} < k|\Pi_2\right]$$

in terms of n_1^{-1}, n_2^{-1} and r^{-1} as n_1, n_2, and r tend to ∞ and n_1/n_2 tend to finite positive constant.

Anderson (1972) obtained asymptotic expansions of the above probabilities with Δ^2 replaced by $D^2 = ||\bar{X}_1 - \bar{X} - 2||_S^2$. Memon and Okamoto (1971) obtained an asymptotic expression for the cdf of $\frac{V+\Delta^2}{2\Delta}$, when $\mu = \mu_1$. Cochran (1964) numerically compared the PMC's (computed from Okamoto-expansion) of the rules $W^* \lessgtr 0$ with those of $W \lessgtr 0$ when $n_1 = n_2$ is large. Memon and Okamoto (1970) derived an asymptotic expansion for the distribution of W^* and the PMC of the W^*-rule in terms of n_1^{-1}, n_2^{-1} and r^{-1}.

John (1959, 1960) derived the distributions of the statistics U, V, W and Rao's statistic (when Σ is known), S being replaced by Σ and obtained explicitly the PMC when the cut-off point is 0. Some bounds, for the PCC were also given by John. When Σ is unknown, and S is used for Σ, some approximations are given for the distributions and the PCC's.

For $p = 1, \mu_1 < \mu_2$, Friedman (1965) considered a rule: $X \gtrless \frac{\hat{\mu}_1 + \hat{\mu}_2}{2}$ and compared its PMC with that of the rule $X \leq \frac{\mu_1 + \mu_2}{2}$, with approximations for large sample size of the training sample.

Das Gupta (1972) proved that for a large class of rules (including the MD and the ML rules– Σ known and Σ unknown) the PCC's are monotonic increasing functions of Δ^2.

Let $\delta = \delta(.; \mu_1, \mu_2, \Sigma)$ be a decision rule when all the parameters are known. We shall denote the plug-in version of δ by $\hat{\delta}$ are replaced the unknown parameters by their respective (standard) estimates. The conditional error probabilities of δ, given $\bar{X}_1$, $\bar{X}_2$ and S, are given by

$$e_i(\hat{\delta}) = P[\hat{\delta}\text{classifies } X \text{ into } \Phi_j | \bar{X}_1, \bar{X}_2, S; \mu = \mu_i],$$

$i \neq j$; i, $j = 1$, 2. The unconditional error-probabilities of $\hat{\delta}$ are $\alpha_i(\hat{\delta}) = E(e_i(\hat{\delta}))$. An estimate of $e_i(\hat{\delta})$ is given by $\hat{e}_i(\hat{\delta}))$ which is obtained by replacing the unknown

parameters in $e_i(\hat{\delta})$ by their standard estimates. Similarly $\hat{\alpha}_i(\delta)$ and $\hat{\alpha}(\hat{\delta})$ are defined.

In the literature, the error-probabilities of the minimax rule δ_0 (parameters known) and its plug-in version $\hat{\delta}_0$ (the MD rule) are mostly considered. When Σ is known, John (1961) derived the distributions of $e_i(\hat{\delta}_0)$ and obtained their means; similar results are obtained when the cut-off point is not 0 and only approximations are given when Σ is not known and n_1 and n_2 are large. In (1964) John considered the similar problem except that μ may be different from μ_1 and μ_2. John (1963) studied the conditional PMC's of the rules defined by the classification statistics

$$\left(1+\frac{1}{n_1}\right)^{-1}||X-\bar{X}_1||_\Sigma^2-\eta\left(1+\frac{1}{n_2}\right)^{-1}||X-\bar{X}_2||_\Sigma^2$$

and Rao's statistic R, when Σ is known. Dunn and Varady (1966) empirically studied (Monte Carlo methods) $1-\hat{\alpha}_i(\delta_0)$, $1-e_i(\hat{\delta}_0)$ and $1-\hat{e}_i(\hat{\delta}_0)$ and derived a confidence interval for the conditional error probabilities of $\hat{\delta}_0$. Geisser (1967) considered a prior measure for the parameters whose (improper) density is proportional to $|\Sigma|^{(p+1)/2}$. Using the posterior distribution of the parameters (given $\bar{X}_1$, $\bar{X}_2$ and S) he used normal approximations. Several estimates of $e_1(\hat{\delta}_0)$, $\alpha_1(\hat{\delta}_0)$, $\alpha_1(\delta_0)$ are suggested in the literature of which the following are main types:

1. Smith's (1947) reallocation or counting estimates,
2. Lachenbruch's (1967) deletion-counting estimate,
3. Fisher's estimate $\hat{e}_i(\hat{\delta}_0)=\Phi(-\frac{D}{2})$ or the estimate obtained by replacing Δ in $\Phi(-\frac{\Delta}{2})$ by some other estimates,
4. the leading term in the Okamoto-expansions and replacing Δ^2 by its estimate,
5. estimates obtained from additional training sample.

It follows from Hills (1966) that $\alpha_i(\hat{\delta}_0)>\alpha_i(\delta_0)$ when $n_1=n_2$. For $p=1$, Hills (1966) obtained the distribution of $\hat{e}_1(\hat{\delta}_0)$ and compared the expectations of $e_1(\hat{\delta}_0)$, $\hat{e}_1(\hat{\delta}_0)$ and those of the counting estimate by exact expressions and numerical computations. In 1967, Lachenbruch proposed the deletion-counting method for estimation. Lachenbruch and Mickey (1968) suggested some estimates of Δ^2 and

studied empirically the behaviour of the estimates (1) - (4). Brofitt (1969) derived the uniformly minimum variance estimates of the mean values of Smith's and Lachenbruch's estimates and suggested some other estimators with smaller mean-square errors. Sorum (1971) obtained some estimates based on additional observations. For known Σ, she derived the means, the variances and approximation to the mean-square errors of most of the estimates and studied these estimates numerically when Σ is unknown (1972a, 1972b). Dunn (1971) studied the average PCC of $\hat{\delta}_0$ and Lachenbruch's estimates (using his estimate of Δ^2) for $n_1 = n_2$ by Monte Carlo methods. For $p = 1$, Sedransk and Okamoto (1971) obtained asymptotic expansions for the mean-square errors of several estimates. Das Gupta (1972) obtained some results on Fisher's and Smith's estimates which generalize Hill's (1966) results.

Chan and Dunn (1972) studied the effect of missing data on the PMC of $\hat{\delta}_0$ by Monte Carlo methods using several standard techniques of handling missing data. Srivastava and Zaatar (1972) derived the ML rule when Σ is known and the samples from the two populations are incomplete (all the p components are not available on each unit sampled) and showed that this rule is admissible Bayes. Lachenbruch (1966) posed the problem when the parent populations of the observations in the training sample are incorrectly identified. Mclachan (1972) derived asymptotic expressions for the mean and the variance of $e_i(\hat{\delta}_0)$ incorporating the possibility of incorrect identification of the training sample. Following Glick (1969) it can be shown that as n_1, $n_2 \to \infty$, $\hat{\alpha}_i(\delta) \to \alpha_i(\delta)$ a.s. and $\hat{\alpha}_i(\hat{\delta}) \to \alpha_i(\delta)$. For related results, see Glick (1969) and for slightly weaker results, see Fix and Hodges (1950), Bunke (1964). Kinderman (1972) suggested a measure of the relative asymptotic efficiencies of two rules by the limit of the ratio of minimum total sample sizes required by the two rules to achieve a maximum probability of error α, as $\alpha \to 0$. In particular, he illustrated this concept by comparing a two sample rule based on samples from Π_0 and Π_1 and a three sample rule using Anderson's statistic when the populations are univariate normal with variance 1 and $\Delta = |\mu_1 - \mu_2| > 0$.

There are many ad hoc methods for choosing "good" components of the vector X. Cochran (1961, 1964) studied the effect of the different components of X on Δ^2, especially when all the correlations are equal. Urbakh (1971) made a similar

study on Δ^2, as well as on Lachenbruch's estimate of Δ^2. Linhart (1961) made a numerical comparison of the effectiveness of selecting components by $\Phi(-\Delta/2)$ and the average PMC of $\hat{\delta}_0$. Weiner and Dunn (1966) also studied empirically three methods for selecting components.

In the normal case, Glick (1969) obtained some interesting results for the 'best-of-class' rules. Let C_{LD} be the class of all rules based on linear (discriminant) functions of X (i.e., partitioning the sample space into two half spaces). Let δ^* be a rule in C_{LD} which maximizes (in C_{LD}) the average (over some known prior or the standard estimates of the proportions in the mixture) of the proportions of the training sample correctly classified. Then this maximum value converges (a.s.) to the PCC of best (Bayes) rule and the risk of δ^* converges a.s. to the Bayes risk as the sample size in the training sample increases to ∞. When the training sample comes from a mixed population, different methods are available to estimate the parameters and the proportions in the mixture, if they are unknown.. For the supervised case, there is not much change in the theory and methods for the usual case discussed before. For some asymptotic results see Glick (1969). In the non-supervised case, there is good deal of literature; for this and relevant references, see Fu (1968), Patrick (1972); for an earlier work see Pearson (1894). Rao (1954) derived an optimal rule in the class of rules for which the probabilities of error depend only on Δ^* using the following criteria: (i) to minimize a linear combination of the derivatives of the error-probabilities with respect to Δ at $\Delta = 0$ subject to the condition that the error-probabilities at $\Delta = 0$ leave a given ratio. (ii) The above criterion with the additional restriction that the derivatives of the error-probabilities at $\Delta = 0$ bear a given ratio. Rao separately treated the problem according as Σ is known or unknown. When Σ is known, Kudo (1959, 1960) showed that the ML rule has the maximum PCC among all translation-invariant rules δ for which the error-probabilities depend on Δ^2, and

$$\alpha_1(\delta; \Delta^2 = \Delta_1^2) = \alpha_2(\delta; \Delta^2 = \Delta_2^2)$$

for all Δ_1 and Δ_2 such that $\frac{\Delta_1^2}{1+1/n_1} = \frac{\Delta_2^2}{1+1/n_2}$. He also showed that this rule is most stringent in the above class without the requirement of translation-invariance.

When Σ is known, Ellison (1962) obtained a class of admissible Bayes rules which includes the MD and ML rules. In this case, Das Gupta (1962, 1965) showed that the ML rule is admissible Bayes (with a different prior and a general loss function) and minimax (unique minimax under some mild conditions). When Σ is unknown, similar results were obtained by Das Gupta (1962, 1965), restricting to the class of rules invariant under translation and the full linear group. For $p = 1$, $n_1 = n_2$, Bhattacharya and Das Gupta (1964) obtained a class of Bayes rules and showed that the MD rule is minimax Bayes. Srivastava (1964) also obtained a class of Bayes rules when Σ is unknown. Geisser (1964) used a prior (improper) density which is proportional to $|\Sigma|^{v/2}$, $v \leq n_1 + n_2$ and $v = 0$ when Σ is known; he derived the (improper) Bayes rules for these priors which are the likelihood-ratio rules in respective cases. For similar analysis, see Geisser (1966). Kiefer and Schwartz (1965) indicated a method to obtain a broad class of Bayes rules which are admissible; in particular, they showed that the LR rules are admissible Bayes when Σ is unknown and $r + 1 > p$. Marshall and Olkin (1968) derived Bayes rules for normal distributions in their special set-up. When $p = 1$, $n_1 = n_2$ and the number of observations to be classified is $n(\geq 1)$, Kinderman (1972) characterized an essentially complete class of rules, invariant under translation and change of signs.

1.2.3 Classification into two multivariate normal populations with different covariance matrices

The distribution of X in Π_i is taken as $N_p(\mu_i, \Sigma_i)$, $i = 0, 1, 2$; furthermore, it is known that Σ_1 and Σ_2 are different. Generally, three cases are considered: (i) $(\mu_0, \Sigma_i) = (\mu_i, \Sigma_i)$ for some $i = 1, 2$, (ii) $\mu_0 = \mu_i$, for some $i = 1, 2$, (iii) $\Sigma_0 = \Sigma_i$ for some $i = 1, 2$. When the parameters are known, the LR statistic was studied by Cavalli (1945) ($p = 1$), Smith (1947), Okamoto (1963) ($\mu_0 = \mu_1 = \mu_2$), Cooper (1963, 1965), Bartlett and Please (1963) ($\mu_0 = \mu_1 = \mu_2 = 0$, $\Sigma_i = (1-\rho_i)I_p + \rho_i J_p, i = 1, 2$), Bunke (1964), Han (1968) ($\Sigma_i = (1 - \rho_i)I_p + \rho_i J_p, i = 1, 2$), Han (1969), ($\Sigma_1 = d\Sigma_2$, $d > 1$), Han (1970) (Σ_i's are of circular type).

Kullback (1952, 1958) suggested a rule based on the linear statistic which maximizes the divergence $J(1,2)$ between $N_p(\mu_1, \Sigma_1)$ and $N_p(\mu_2, \Sigma_2)$. He also obtained some partial results on deriving the optimum class of rules based on linear functions of X from Neyman-Pearson viewpoint (i.e, minimizing one PMC by controlling the other). Clunies-Ross and Riffenburgh (1960) studied this problem geometrically. Anderson and Bahadur (1962) derived the minimax rule and characterized the minimal complete class after restricting to the class of rules based on linear functions of X. Banerjee and Marcus (1965) studied the form of this minimax rule.

Gilbert (1969) derived the PMC of LR rule when the parameters are known and compared it with the PMC of the corresponding LR rule when $\Sigma_1 = \Sigma_2$. For the latter he obtained the optimum cut-off point for which the total PMC is minimized.

Lbov (1964) studied the PMC when p is large and the parameters are known. Grenander (1972) considered a similar problem.

Andersor (1964) studied the problem of choosing components by minimizing Bayes risk when the distributions are univariate normal.

When μ_0 equals either μ_1 or μ_2, and the covariance matrices are known, a class of admissible Bayes rules was obtained by Ellison (1962); in particular he showed that the MD and the ML rules are admissible Bayes.

Okamoto (1963) derived the minimax rule and the form of Bayes rule when the parameters are known; he studied some properties of the Bayes' risk function, and suggested when Σ_i's are unknown, and the common values of μ_i's may be known or unknown. The asymptotic distribution of the plug-in $log(LR)$ statistic was also obtained by Okamoto. Bunke (1964) derived the minimax rule and the form of a Bayes rule and proved that the plug-in minimax rule is consistent. Following the method of Kiefer and Schwartz (1965), Nishida (1971) obtained a class of admissible Bayes rules when the parameters are unknown.

Matusita (1967) considered a minimum distance rule and suggested its plug-in version by replacing the unknown parameters by their respective estimates; the distance between two distributions with p.d.f.' p_1 and p_2 with respect to a σ-finite

measure m was taken as

$$\left[\int(\sqrt{p_1(x)}-\sqrt{p_2(x)})dm\right]^{1/2}.$$

He separately treated the different cases according as the μ_i's and Σ_i's are known or unknown, and obtained some bounds for the PCC.

When $\Sigma_1 = d\Sigma_2 \quad (d > 1)$, the distributions of the $log(LR)$ statistic and its plug-in version (by replacing the mean vectors by their estimates) were derived by Han (1969). Similar results were obtained by Han (1970) when the Σ_i's are of "circular" type.

Chaadha and Marcus (1968) studied (mainly by simulation) the behaviour of some estimates of measure of divergence defined as $2(\mu_1-\mu_2)^T(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2)$.

Aoyoma (1950) considered rules of the form $X \gtrless x_0$ and found the optimum value of x_0 which minimizes the PCC when X is a mixture of two univariate normal distributions; the mixture ratio may be known or unknown.

1.2.4 Classification into two discrete and other non-normal distributions

The following discrete and non-normal distributions have been considered in the literature.

Multinomial distribution

The random variable X is distributed as a multinomial distribution with k cells in each of the populations. Matusita (1956) proposed a minimum distance rule based on samples. Chernoff (1956) considered the problem that the distribution of X in Π_1 is the multinomial with equal-cell probabilities and a multinomial with unknown cell-probabilities in Π_2 and derived some 'optimal' rules. Wesler (1959) assumed the distribution of X in Π_i is as a multinomial with cell probabilities being any permutation of a given probability vector $p^{(i)}$, $(i = 1,2)$ and considered the classification procedure which minimizes one error probability when the maximum of other error probability is held fixed. Cochran and Hopkins (1961) obtained the

form of the Bayes rules and considered, in particular, the 'maximum likelihood' rule. For this rule they discussed the effect of 'plug-in' on the PMC and suggested a correction for bias. Raiffa (1961) considered the multinomial distributions and, in general, discrete distributions are included in the development of theories. Hills (1966) paper contains some theoretical developments on the errors of misclassification for the 'ML' rule in the two-population case. Bunke (1966) studied a property of the estimated minimax rule for the multinomial distributions. Glick's (1969) development is for general discrete distributions but also specialized for multinomial distributions. This paper generalizes some of the results of Cochran and Hopkins (1961) and Hills (1966) and furnishes rigorous proofs.

Multivariate Bernoulli distributions

The random variable X is a $p \times 1$ vector and each component of X takes values 0 or 1. Bahadur's (1961) paper gives some approximations to the log likelihood-ratio, e.g., normal approximation and approximations using various truncations of Bahadur's series representation for the probability functions. Some approximations to Kullback-Leibler symmetric information measure J are also obtained. These approximations are useful when J is small and p is large. Solomon (1960, 1961) studied numerically the effectiveness (PMC) and relative comparisons among rules based on the sum of the components, Fisher's LDF, LR statistic, and some truncated functions obtained from Bahadur's series representation for the probability functions. Hills's (1967) paper is concerned with the problem of estimating $\log(LR)$ at a given point $X = x_0$. Elashoff et al. (1967) considered Fisher's LDF, two functions based on a logistic model, and a function based on the assumption of mutual independence of the components are considered as a possible classification statistics. The effectiveness of these statistics is studied numerically. Martin and Bradley (1972) takes the probability function of X in Π_i as

$$p_i(x) = f(x)[1 + h_s(a_i, x)],$$

where h_s is a linear function of the orthogonal polynomials on the sample space of X. This paper deals with the estimation of a_i and f subject to some constraints.

Parametric non-normal continuous-type distributions

Cooper (1962, 1963) considers the distribution of X in Π_i as a known multivariate distribution of Pearson Type II or Type VII. The LR statistic is studied. Bhattacharya and Das Gupta's (1964) paper considers the distribution of X in Π_i as a member of the one-parameter exponential family. A class of admissible Bayes rules is obtained. Cooper (1965) takes the pdf of X in Π_1 as

$$p_i(x) = A_i|\Sigma_i|^{-1/2} f_i[(Q_i(x))^{1/2}],$$

where Q_i is a positive definite quadratic form and $f_i(u)$ decreases as u increases from 0. The LR statistic is studied. Day and Kerridge (1967) take the PDF of X in Π_i as

$$p_i(x) = d_i \exp\left[-\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i)\right] f(x).$$

Two cases are considered, namely, (i) $f(x) = 1$, (ii) $\Sigma = I$ and $f(x) = 1$ if every component of x is either 0 or 1 and $f(x) = 0$, otherwise. The posterior probability of the hypothesis $H_i : \phi = \phi_i$, given $X = x$, is expressed as $\frac{\exp(x'b+c)}{[1+\exp(x'b+c)]}$. This paper mainly deals with the maximum likelihood estimates of b and c. For classification, it includes the idea of 'doubtful' decision.

Other cases

Kendall's (1966) paper suggests some heuristic rules based on categorization of data. Marshall and Olkin (1968) formulates the problem as X is a binomial random variable with probability of success Y which is distributed as the uniform distribution on (0,1). The form of a Bayes rule is obtained.

1.2.5 Classification with both dichotomous and continuous variables

Chang and Afifi's (1974) article derives the Bayes procedure for classifying an observation consisting of one dichotomous and k continuous variables for a point biserial model which postulates that the conditional distribution of the k continuous

variables given a specific value of the dichotomous variables is normal. Krzanowski (1975) generalizing the Chang and Afifi's (1974) procedure, proposes a model for the situation where q of the variables are dichotomous and k variables are conditionally normal with mean vector $\mu_i^{(m)}$ corresponding to the mth ($m = 1, \ldots, 2^q; i = 1, 2$), and common covariance matrix Σ in all cells for both populations. This "location" model is often appropriate in medical applications where some of the classification variables are symptoms of a presence/absence nature while others, such as weight and blood pressure, are continuous.

1.2.6 Nonparametric methods

The so-called nonparametric or distribution-free methods are used in statistical inference when one is concerned with a wide class of distributions which usually cannot be expressed as a parametric family with finite number of parameters. When a statement regarding the probability of certain statistical inference remains valid for every member in a given family of distributions, we call that a distribution-free inference with respect to that family; in particular, if the distribution of a statistic (used for inference) is the same for every member of a family of underlying distributions of the random variables involved, we say that the statistic is distribution-free with respect to that family. In the classification problem sometimes we face a similar situation when we devise rules for a broad class of underlying distributions whose structures cannot be expressed in simple parametric forms. However, unlike the problems of testing hypothesis or estimation, a "classification problem cannot be distribution-free" (Anderson, 1966) in the broad sense.

The available work in this area can be classified broadly into three main categories:

1. Consider a "good" rule (generally taken as a Bayes and/or an admissible minimax) assuming that the distributions are known.. In this rule, replace the cdf's or the pdf's by their respective sample estimates. The rule thus obtained will be called "plug-in" rule. References for this method may be found in Van Ryzin (1966), Fu (1968), Glick (1969) and Patrick (1972).

2. Use the statistics involved in devising some well-known test for the nonparametric two-sample or k-sample problems. Fix and Hodges (1951) , Parzen (1962), and Loftsgarden and Quesenberry's (1965) nearest neighbour rules are examples for this method.

3. Some ad-hoc methods which are typical for the classification problems, e.g., the "minimum distance" rule proposed by Matusita (1956)

In the literature, the main emphasis is (a) to study the asymptotic behaviour (e.g., consistency, efficiency in some sense) of the rules, (b) to obtain some bounds for the PCC of a given rule, and (c) to study the small sample performance.

1.2.7 Other approaches to classification

Recently, many advances have occurred in the applications of statistical classification techniques to pattern recognition, neural networks and behavioral sciences. In some of these newly found applications, researchers have taken a somewhat different approach to the traditional statistical classification. Following is a brief summary of a few recent research articles so that one could easily see how popular and useful is the area of statistical classification at the present time.

Duda and Hart (1973) have given an excellent introduction to statistical pattern classification and scene analysis. Jain (1987) has studied the dimensionality and sample size consideration in classification. Shoemaker (1991) has addressed the implication of learning based on minimization of sum-square classification error. Geva and Sitte (1991) have described a variant of nearest neighbor pattern classification. Traven (1991) has presented a method for designing a near optimal nonlinear classifier based on self-organizing technique for estimating probability density functions when only weak assumptions are made about the densities.

Mullen and Ennis (1987) have given a mathematical formulation of multivariate Euclidean models for discrimination methods. Takane (1987) has given an analysis of contingency tables by ideal point discriminant analysis. Tatsuoka and Tatsuoka (1987) have discussed the theoretical foundation of the model, which permits measuring cognitive skill acquisition, diagnosing cognitive errors, detecting

the weakness and strength of knowledge possessed by individuals, by introducing "bug distribution" and hypothesis testing for classifying subjects into their most plausible latent state of knowledge. Grayson (1988) has considered the two-group classification when a unidimensional latent trait, θ, is appropriate for explaining data.

1.2.8 Robust classification procedures

If the distribution of the population is not normal, the linear classification procedure is in general no longer optimal. In general the underlying distributions of populations are not known.

The robustness of the linear classification procedure to nonnormality has been examined by several authors. Subrahmaniam and Ching'anda (1978) and Ching'anda and Subrahmaniam (1979) derive the distributional properties of the misclassification probabilities in sampling from a number of nonnormal populations. The authors use a modification of the developments of John (1960a, 1960b) in the normal case. Amoh (1983) develops some similar results for the inverse Gaussian distribution. Balakrishnan et al (1985) examine the case of sampling from a mixture of two normal populations. Balakrishnan et al. (1986) derive the distribution and the expected values of the errors of misclassification in Chang and Afifi's (1974) classification procedure based on dichotomous and normal variables. Kocherlakota et al (1987) paper gives a unified development with examples of the expansion available for the asymptotic distribution of the errors of misclassification in the non normal situations. They also elaborate the inflation in the error of misclassification through a detailed study in the case of outliers in sampling from normal distribution. Balakrishnan et al (1988) paper gives a study of the robustness of the Chang and Afifi's (1974) procedure which includes the general results for the distribution functions of the error rates as well as their expected values. Then it gives a detailed study for the case of truncated normal and mixtures of normals (contamination model). From these published results we can make the following generalizations:

1. If the distributions have lighter tails than the normal, then the linear classification procedure (for homoscedastic data) or the quadratic classification procedure (for heteroscedastic data) should perform very adequately.

2. If the distributions are heavy tailed and skew, then the linear classification procedure and quadratic classification procedure will perform very poorly.

3. If the distributions are heavy tailed but symmetric and the training samples are very large, then the quadratic classification procedure may perform reasonably well in terms of overall error rate (though not the individual rates). However if the training samples are small, then the heavy distributional tails introduce excessively large sampling errors into the parameter estimates, and so lead to an unreliable discriminant function.

As these remarks imply, the departure from normality and departures from linearity are to be watched for heavy tailed distribution, especially if it is skew. Balakrishnan, Tiku and El Shaarawi (1985) proposed a robust linear classification procedure based on Tiku's (1967, 1970, 1980) MML (Modified Maximum Likelihood) estimators based on symmetrically Type-II censored samples from the normal distribution analogous to classical classification procedure. By comparing these two procedures they have shown that they are more efficient than the respective two-way classification procedures, when one of the two errors of misclassification is at a fixed level. A robust quadratic classification procedure analogous to the classical procedure has been obtained by Tiku and Balakrishnan (1985, 1989) by replacing the classical estimators of means and variances of two populations by the corresponding MML estimators. Through a simulation study for small samples and asymptotic argument for large sample they have shown that their procedure perform better than the corresponding classical procedure. Tiku and Balakrishnan (1985, 1989) have made transformations on the observations and proposed transformed linear classification procedures and studied the robustness of these procedures. Also they have extended this into the bivariate population case. Balakrishnan and Tiku (1988b) have proposed a robust classification procedure based on MML estimators for samples based on both dichotomous and continuous variables analogous to the Chang and Afifi's

(1974) procedure. They have studied the robustness of these two procedures and come to the conclusion that their procedure perform better. Tiku, Balakrishnan and Ambagaspitiya (1988) have also studied the errors of misclassification by using asymptotic arguments for large samples and simulation study for small samples under normal and wide range of non-normal models.

1.3 Linear regression

Regression models are used for several purposes, including data description, parameter estimation, prediction and estimation, and control. Engineers and scientists frequently use equations to summarize or describe a set of data. Regression analysis is helpful in developing such equations. For example, we may collect a considerable amount of delivery time, delivery volume data, and a regression model would probably be a much more convenient and useful summary of those data than a table or even a graph. Sometimes parameter estimation problems can be solved by regression methods. For example, suppose that an electrical circuit contains an unknown resistance of R ohms. Several different known currents are passed through the circuit and the corresponding voltages measured. The scatter diagram will indicate that voltage and current are related by a straight line through the origin with slope R (because voltage E and current I are related by Ohm's law $E = IR$). Regression analysis can be used to fit this model to the data producing an estimate of the unknown resistance. Many applications of regression involve prediction of the response variable. For example, we may wish to predict delivery time for a specified number of cases of soft drinks to be delivered. These predictions may be helpful in planning delivery activities such as routing and scheduling, or in evaluating the productivity of delivery operations. ' Regression models may be used for control purposes. For example, a chemical engineer could use regression analysis to develop a model relating the tensile strength of paper to the hardwood concentration in the pulp. This equation could then be used to control the strength to suitable values by varying the level of hardwood concentration. When a regression equation is used for control purposes, it is important that the variables be related in a casual manner.

Note that a cause and effect relationship may not be necessary if the equation is to be used only for prediction. In this case it is only necessary that the relationships that existed in the original data used to build the regression equation are still valid. For example, the daily electricity consumption during July in Hamilton, Ontario may be a good predictor for the maximum daily temperature in July. However, any attempt to reduce the maximum temperature by curtailing electricity consumption is clearly doomed to failure.

1.3.1 The need for robust estimation

When the observations y in the linear regression model $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ are normally distributed, the method of least squares works well in the sense that it produces an estimate of β that has good statistical properties. However, when the observations follow some nonnormal distribution, particularly one that has longer or heavier tails than the normal, the method of least squares may not be appropriate. Heavy-tailed distributions usually generate outliers, and these outliers may have a strong influence on the least squares estimate. In effect, outliers "pull" the least squares fit too much in their direction, and consequently the identification of these outliers is difficult because their residuals have been made artificially small. Skilful residuals analysis coupled with the use of techniques for identifying influential observations such as Daniel and Woods (1980) weighted sum of squared distance of the ith point from the center of the data i.e.

$$WSSD_i = \sum_{j=1}^{k} \left[\frac{\hat{\beta}_j(x_{ij} - \bar{x}_j)}{\sqrt{MS_E}} \right]^2, \quad i = 1, 2, \ldots, n$$

or the use of hat matrix suggested by Hoaglin and Welsch (1978) can help the analyst discover these problems. However, the successful use of these diagnostic procedures often requires abilities beyond those of the average analyst.

A number of authors have proposed robust regression procedures designed to dampen the effect of observations that would be highly influential if least squares were used. That is, a robust procedure tends to leave the residuals associated with outliers large, thereby making identification of influential points much easier.

In addition to insensitivity to outliers, a robust estimation procedure should be 90-95 percent as efficient as least squares when the underlying distribution is normal. Basic references in robust estimation include Andrews et al. (1972), Andrews (1974), Yale and Forsythe (1974), Hill and Holland (1977), Hogg (1974, 1979a,b) and Huber (1972, 1973, 1981). Reference may also be made to the works of Leone and Moussa-Hamouda (1973), Moussa-Hamouda and Leone (1974, 1977a, 1977b, 1977c), Moussa-Hamouda (1988), and Tiku (1981) regarding the robust methods of estimation (through censoring) for the regression problem with multiple measurements at each level of explanatory variables.

To motivate the discussion, and to demonstrate why it may be desirable to use an alternative to least squares when the observations are nonnormal, consider the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i, \qquad i = 1, 2, \ldots, n$$

where the errors are independent random variables that follow the double exponential distribution

$$f(e_i) = \frac{1}{2\sigma} exp\left(-\frac{|e_i|}{\sigma}\right), \qquad -\infty < e_i < \infty$$

The double exponential distribution is more pointed in the middle than the normal and tails off to zero as $|e_i|$ goes to infinity. However, since the density function goes to zero as $\exp(-|e_i|)$ goes to zero, and the normal density function goes to zero as $\exp(-e_i^2/2)$ goes to zero, we see that the double exponential distribution has heavier tails than the normal.

We will use the method of maximum likelihood to estimate β_0 and β_1. The likelihood function is

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \frac{1}{2\sigma} exp\left(-|e_i|/\sigma\right) = \frac{1}{(2\sigma)^n} \exp\left(-\sum_{i=1}^{n} |e_i|/\sigma\right)$$

Therefore maximizing the likelihood function would involve minimizing $\sum_{i=1}^{n} |e_i|$, the sum of the absolute errors. It is easy to show that the method of maximum likelihood applied to the regression model with normal errors leads to the least square criterion. Thus the assumption of an error distribution with heavier tails

than the normal implies that the method of least square is no longer an optimal estimation technique. Note that the absolute error criterion would weight outliers far less severely than would least squares. Minimizing the sum of the absolute errors is often called the L_1-norm regression problem. Least squares is the L_2-norm regression problem.

The L_1-norm regression problem can be formulated as a linear programming (LP) problem. Unfortunately, standard LP algorithm do not ensure that unbiased estimates of β are obtained. Hartley and Sielken (1973) have developed an efficient LP algorithm that produces an unbiased solution. For other references on L_1- norm regression, see Barrodale (1968), Barrodale and Roberts (1973), Book, Booker, Hartley and Sielken (1980), Gentle, Kennedy and Sposito (1977), and Wagner (1959).

The L_1-norm regression problem is a special case of L_p-norm regression, in which the model parameters are chosen to minimize $\sum_{i=1}^{n} |e_i|^p \quad (1 \leq p \leq 2)$. For $1 < p < 2$ this reduces to a nonlinear programming problem. Forsythe (1972) has studied this procedure extensively for the straight-line regression via Monte Carlo simulation, using several nonnormal error distributions. He notes that $p = 1.5$ is a good compromise choice leading to substantially better estimates than least squares when the errors are nonnormal. When the error distribution is normal, using $p = 1.5$ results in estimates that are at worst 90 percent as efficient as least squares.

1.3.2 M estimators

We have noted that L_1-norm regression problem arises naturally from the maximum likelihood approach with double exponential errors. In general we may define a class of robust estimators that minimize a function ρ of the residuals, for example

$$\frac{min}{\beta} \sum_{i=1}^{n} \rho(e_i) = \frac{min}{\beta} \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i'\beta) \tag{1.3}$$

where $\mathbf{x}_i'$ denotes the ith row of $\mathbf{X}$. An estimator of this type is called an M-estimator, where M stands for maximum likelihood. That is, the function ρ is related to the likelihood function for an appropriate choice of the error distribution. For

example, if the method of least squares is used (implying that the error distribution is normal), then $\rho(z) = \frac{1}{2}z^2, -\infty < z < \infty$. The M-estimator is not necessarily scale-invariant (that is, if the residuals $y_i - \mathbf{x}_i\beta$ were multiplied by a constant, the new solution to (1.3) might not be the same as the old one). To obtain a scale-invariant version of this estimator, we usually solve

$$\frac{min}{\beta}\sum_{i=1}^{n}\rho(e_i/s) = \frac{min}{\beta}\sum_{i=1}^{n}\rho\left[(y_i - \mathbf{x}_i'\beta)/s\right] \tag{1.4}$$

where s is a robust estimate of scale. A popular choice for s is

$$s = median|e_i - median(e_i)|/0.6745$$

The constant 0.6745 makes s an approximately unbiased estimator of σ if n is large and the error distribution is normal.

To minimize (1.4), equate the first partial derivatives of ρ with respect to β_j $(j = 0, 1, \ldots, k)$ equal to zero, yielding a necessary condition for a minimum. This gives the system of $\rho = k + 1$ equations

$$\sum_{i=1}^{n} x_{ij}\psi\left[(y_i - \mathbf{x}_i'\beta)/s\right] = 0, \quad j = 0, 1, \ldots, k \tag{1.5}$$

where $\psi = \rho'$ and x_{ij} is the ith observation on the jth regressor and $x_{i0} = 1$. In general the ψ function is nonlinear and (1.5) must be solved by iterative methods. While several nonlinear optimization techniques could be employed, iteratively reweighted least squares is most widely used. This approach is usually attributed to Beaton and Tukey (1974). To use iteratively reweighted least squares, suppose that an initial estimate $\hat{\beta}_0$ is available and that s is an estimate of scale. Then write the $p = k + 1$ equations in (1.5)

$$\sum_{i=1}^{n} x_{ij}\psi\left[(y_i - \mathbf{x}_i'\beta)/s\right] = \sum_{i=1}^{n} x_{ij}\frac{\psi\left[(y_i - \mathbf{x}_i'\beta)/s\right]}{(y_i - \mathbf{x}_i'\beta)/s}(y_i - \mathbf{x}_i'\beta)/s = 0, \quad j = 0, 1, \ldots, k$$

as

$$\sum_{i=1}^{n} x_{ij}w_{i0}(y_i - \mathbf{x}_i'\beta) = 0, \quad j = 0, 1, \ldots, k \tag{1.6}$$

where

$$w_{i0} = \begin{cases} \frac{\psi\left[(y_i - \mathbf{x}_i'\beta)/s\right]}{(y_i - \mathbf{x}_i'\beta)/s} & \text{if } y_i \neq \mathbf{x}_i'\hat{\beta}_0 \\ 1 & \text{if } y_i = \mathbf{x}_i'\hat{\beta}_0 \end{cases} \tag{1.7}$$

In matrix notation, (1.6) becomes

$$\mathbf{X}'\mathbf{W}_0\mathbf{X}\beta = \mathbf{X}'\mathbf{W}_0\mathbf{y} \tag{1.8}$$

where $\mathbf{W}_0$ is an $n \times n$ diagonal matrix of "weights" with diagonal elements $w_{10}, w_{20}, \ldots, w_{n0}$ given by (1.7). We recognize (1.8) as the usual weighted least squares normal equations. Consequently the one-step estimator is

$$\hat{\beta}_1 = (\mathbf{X}'\mathbf{W}_0\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}_0\mathbf{y}$$

At the next step, we recompute the weights from (1.7) but using $\hat{\beta}_1$ instead $\hat{\beta}_0$. Usually only a few iterations are required to achieve convergence. As Holland and Welsch (1977) note, the iteratively reweighted least squares procedure requires only a standard weighted least squares computer program.

A number of popular robust criterion functions such as Huber's (1964) t function, Andrews (1972, 1974) wave function, Hampel's (1977) 17A and Ramsay's (1977) E_a function, can be found in the literature.

1.3.3 R and L estimation

In addition to the M estimators discussed in the previous section, there are other approaches to robust regression. R estimation is a procedure based on ranks. To illustrate the general procedure, consider replacing one factor in the least squares objective function $S(\beta) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i'\beta)$ by its rank. Thus if R_i is the rank of $y_i - \mathbf{x}_i'\beta$, then we wish to minimize $\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\beta)R_i$. More generally, we could replace the ranks (which are integers $1, 2, \ldots, n$) by the score function $a(i)$, $i = 1, 2, \ldots, n$, so that the objective function becomes

$$\frac{min}{\beta}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\beta)a(R_i)$$

If we set the score function equal to ranks, that is, $a(i) = i$, the results are called Wilcoxon scores. Another possibility is to use median scores, that is, $a(i) = -1$ if $i < (n+1)/2$ and $a_i = 1$ if $i > (n+1)/2$. Important references on R estimation in regression include Adichie (1967), Hogg and Randles (1975), Jaeckel (1972), and Jureckova (1977).

L estimators are based on linear functions of order statistics. The sample median would be an $L-$ estimator, since it is a measure of location based on these order statistics. A number of other $L-$ estimators for the location problem are described in Andrews et al. (1972). The use of L estimation in the regression context is not as simple as M and R estimation. Denby and Larsen (1977) describe slice regression, which for one regressor divides the data into groups and fits the straight line using the centroids of the groups. A stepwise-type extension of this technique could be used for multiple regression. Moussa-Hamouda and Leone (1974, 1977a, 1977b, 1977c) and Moussa-Hamouda (1988) propose procedures for simple linear regression with repeated observations on y at each x that involve trimming or discarding remote values of y.

1.4 Scope of this thesis

In the last two sections, various developments that have taken place with regard to the issue of robustness in the areas of classification and linear regression have been described. In this section, we present a description of the contributions that are made in this thesis to these two broad and significant areas of research.

In Chapter 2, we first describe briefly in Section 2 the univariate MML estimators based on Type-II censored samples derived by Tiku (1967, 1970, 1980) for the sake of clarity, completeness and a better understanding of the latter developments in this thesis. In Section 3 we present next a discussion on the classical univariate two-way linear classification procedure and its robust analogue based on the MML estimators as developed by Balakrishnan, Tiku and El Shaarawi (1985). For facilitating the generalization of this robust univariate linear classification procedure based on MML estimators to the multivariate situation, we first present in Section 4 the extension of the univariate MML estimation method to the multivariate situation as given by Tiku (1988) and Tiku and Balakrishnan (1988). Some asymptotic distributional properties of these multivariate MML estimators are also discussed in this section. In Section 5, we make use of these multivariate MML

estimators to develop a robust multivariate linear classification procedure. We describe here the asymptotic determination of the cut-off point of the procedure when one of the two errors of misclassification is at a pre-fixed level. In Section 6, we assess the performance of the classical and the robust multivariate linear classification procedures by applying them to Majumdar's (1941) anthropometric data after one of the two errors of misclassification at a pre-specified level. Through this comparative study, we show that the robust procedure based on the MML estimators developed in this chapter is more efficient than the classical classification procedure. Further, by allowing both the errors of misclassification to freely float (that is, by not fixing either of the two errors at a specified level), we apply the classical and the robust procedures to this data and find the average errors of misclassification for both the procedures. By comparing these average error values with those of some well-known non-parametric procedures like the nearest-neighbour and density estimation methods, we show that the robust classification procedure based on the MML estimators developed in this chapter is better than the classical procedure and also is considerably more efficient than the non-parametric or distribution-free procedures.

In Chapter 3, we consider the classification problem for data consisting of a dichotomous variable and an associated continuous variable. We first present in Section 2 the basic model of the classification problem considered here and describe briefly the classical linear and the robust linear procedures as developed by Chang and Afifi (1974) and Balakrishnan and Tiku (1988b), respectively. A brief discussion of the asymptotic determination of the cut-off points of these two procedures, when one of the two errors of misclassification is at a pre-specified level, is also included in this section. In Section 3, we describe the two procedures when both the errors of misclassification are allowed to float and derive some asymptotic formulas for the errors of misclassification. We also carry out an extensive Monte Carlo study in this section through which we display that the robust procedure has a smaller average error rate than the classical procedure under normal and various non-normal models. The results presented in this chapter have been reported in a paper by Tiku, Balakrishnan and Ambagaspitiya (1989).

In Chapter 4, we consider the classification problem based on a dichotomous and a multivariate normal variable and describe the classical linear procedure. We also develop an analogous robust procedure based on the MML estimators and show that it is very efficient and robust to departures from normality. In Section 2, we first describe the basic model of the classification problem based on a dichotomous variable and an associated multivariate normal variable. In Section 3, we derive the likelihood ratio rule for classifying a new independent observation for the case when all the parameters of the model are known and derive exact formulas for the two errors of misclassification. In Section 4, we consider the case when all the parameters are unknown and explain the classical linear classification procedure due to Chang and Afifi (1974) and discuss its asymptotic determination of the cut-off point when one of the two errors of misclassification is at pre-fixed level. Next, by making use of the multivariate MML estimators developed in Chapter 2, we propose a robust linear classification procedure and discuss its asymptotic determination of the cut-off point when one of the errors of misclassification is at a pre-fixed level. We also discuss in this section some asymptotic properties of this robust linear classification procedure. In Section 5, we consider the bivariate case and carry out a Monte Carlo comparison of the two procedures when one of the two errors of misclassification is at a pre-fixed level and display that the robust linear classification procedure developed in this chapter is more efficient and quite robust to departures from bivariate normality as compared to the classical linear procedure. This work generalizes the univariate robust linear procedure of Balakrishnan and Tiku (1988b) described in Chapter 3 and also extends the multivariate robust linear procedure proposed in Chapter 2 to cover situations where the data have a dichotomous variable (like gender or age group) and an associated multivariate continuous variable.

In Chapter 5, we consider the classification problem based on a k-variate dichotomous and an univariate normal variable. The following is a typical situation where this model arises naturally: Two life insurance companies offer similar life insurance policies. Each insurance company keeps a record per policy holder consisting of gender (0 for female, 1 for male), marital status (0 for married, 1 for

single), smoking status (0 for non-smoking, 1 for smoking), etc., and also the premium. Now, given a sample of records from each company, one may be interested in classifying a record into one of the two companies. In Section 2, we first describe the basic model of the classification problem based on a k-variate dichotomous and an univariate continuous variable. In Section 3, we derive the classical classification procedure under the assumption of normal distribution for the continuous variable first for the case when all the parameters are unknown. By considering the case when both errors of misclassification are allowed to float, we derive the exact distribution functions, density functions, and expected values of the two errors of misclassification. In Section 4, we consider the outlier-normal model for the data on the continuous variable and derive the exact distribution functions, density functions, and expected values of the two errors of misclassification of the normal procedure when both the errors are allowed to float. In Section 5, we carry out a similar study by assuming the mixture-normal model for the continuous variable. In Section 6, by using all these results we assess the robustness features of the normal classification procedure under the outlier-normal and mixture-normal variables. Two dichotomous and one continuous variables model is made use of for this robustness study. Through this study, we show that the classical linear classification procedure is sensitive to departures from normality.

In Chapter 6, we consider the classification problem based on a k-variate dichotomous and an associated continuous variable. In Section 2, we present the classical linear procedure for this problem when all the parameters in the model are known. In Section 3, we present the classical linear classification procedure for the case when all the parameters are unknown and discuss its asymptotic determination of the cut-off point when one of the two errors of misclassification is at a pre-fixed level. In Section 4, we propose an analogous robust linear classification procedure and explain its asymptotic determination of the cut-off point when one of the two errors of misclassification is at a pre- specified level. We also discuss in this section some asymptotic optimal properties of this procedure. In Section 5, we derive the asymptotic formulas for the two errors of both the procedures and demonstrate why the procedure based on MML estimators will be robust to departures from normality

while the classical linear procedure will be non-robust in nature. In Section 6, we make a comparison of these two procedures in case of small samples using Monte Carlo simulations by considering normal as well as various non-normal models. Through this study, we show that the procedure based on the MML estimators proposed in this chapter is quite robust to departures from normality and also highly efficient as compared to the classical linear procedure. This work generalizes the results of Balakrishnan and Tiku (1988b) and Chang and Afifi (1974) to the case when the data contains a k-variate dichotomous variable and an associated univariate continuous variable.

In Chapter 7, we consider the classification problem in a very general set-up, namely, when the data contains a multivariate dichotomous and an associated multivariate normal variable. In Section 2, we first describe the basic model of the general classification problem based on a k-variate dichotomous variable and an associated $p-$ variate normal variable. In Section 3, we derive the classical likelihood-ratio classification procedure for the case when all the parameters in the model are known and derive exact and explicit expressions for the two errors of misclassification. In Section 4, we describe the classical linear classification procedure for the case when all the parameters in the model are unknown and explain its asymptotic determination of the cut-off point when one of the two errors of misclassification is at a pre-fixed level. In Section 5, we make use of the multivariate MML estimators developed in Chapter 2 to propose a robust linear procedure for this general classification problem. We also describe in this section the asymptotic determination of the cut-off point of this robust procedure when one of the two errors of misclassification is at a pre-specified level and discuss some asymptotic optimal properties of this procedure. The results presented in this chapter generalize the results developed in Chapter 3, 4 and 6 in different ways.

In Chapter 8, we consider the classical univariate linear classification procedure when the parameters are all unknown and study what effect the heterogeneity of variances of the two populations has on this classification procedure. In Section 2, by assuming normal distributions for the two populations with unequal variances

we derive the distribution functions and expected values of the errors of misclassification of the classical linear procedure. It should be mentioned here that Gilbert (1969) examined the effect of heterogeneity of variances on the linear classification procedure in the case when all the parameters are known. So, the results presented in Section 2 and in the rest of this chapter are of broader and more useful nature as they are for the case when all the parameters in the model are unknown. In Section 3, we next derive similar results by assuming outlier-normal models for the samples from the two populations. In Section 4, we make use of these results and examine the effects of the heterogeneity of variances on the classical linear classification procedure. We show that the errors of misclassification tend to become large as variances of the two populations get apart and if, in addition to the heterogeneity of variances, the samples also contains some outliers, the errors of misclassification tend to become very large.

In Chapter 9, we consider the classification problem based on a dichotomous and an univariate normal variable when the conditional variances of the two populations are unequal. In Section 2, we describe the basic model of this classification problem. In Section 3, we derive the likelihood-ratio classification procedure for the case when all the parameters of the model are known. We discuss some approximations to the distribution of the classification statistic in this case. In Section 4, by considering the case when all the parameters of the model are unknown we describe the classical quadratic classification procedure and develop an analogous robust quadratic classification procedure based on the MML estimators. The asymptotic determination of the cut-off points of these two procedures, when one of the two errors of misclassification is at a pre-specified level, is also discussed in this section. In Section 5, we develop a transformed linear classification procedure and a robust analogue of it based on the MML estimators. In Section 6, we make a comparison of all these four procedures under normal as well as various nonnormal models. Through this comparative study, we show that the robust transformed linear classification procedure proposed in Section 5 is the best of these four procedures as it has its e_{12} values to be quite close to the presumed level and has its $1 - e_{21}$ values to be almost the same as that of the corresponding quadratic classification procedure,

in addition to having a simple asymptotic distribution theory and also being robust to departures from normality. The developments in this chapter generalize the results of Tiku and Balakrishnan (1985, 1989) to the case when the data contain a dichotomous variable and an associated univariate continuous variable.

At this juncture, we turn our attention towards robust estimation of parameters in a linear regression model. In Chapter 10, by considering a simple linear regression model with single measurements at each level of the explanatory variable, we derive the MML estimators of the parameters in the simple linear regression model by directly adopting Tiku's (1967, 1970, 1980) MML approach and approximating the likelihood function based on Type-II symmetrically censored sample of normalized residuals. In Section 2, we derived the MML estimators from a Type-II symmetrically censored sample formed from the normalized residuals. We show that these estimators are of the same form as those of the classical least-squares estimators. In Section 3, we describe the determination of the estimated normalized residuals that need to be censored for this estimation method. This conforms with the methodology on the identification of outliers in a linear regression model. In Section 4, we derive the asymptotic variances and covariances of these MML estimators through the information matrix and discuss the asymptotic distribution of these estimators. In Section 5, we compare the MML estimators with the classical estimators under normal and scale mixture-normal models for the error variable for various sample sizes. We show that the MML estimators are slightly inferior to the classical estimators under the normal distribution but are considerably jointly more efficient than the classical estimators under the mixture-normal distribution. In Section 6, we describe some other prominent methods of estimation including Adichie's non parametric estimates, Andrews' M-estimates, Huber's M-estimates, Yale and Forsythe's Winsorized estimates, and Tan and Tabatabai's modified Winsorized estimates and compare the performance of all these methods of estimation (through their bias and mean square error) under normal as well as a wide range of non-normal models. Through this comparative study, we show that the MML estimators derived in this chapter are quite efficient and also robust to departures

from normality. In Section 7, we study these methods of estimation under departures from linearity by introducing a quadratic factor in the regression model. Once again, the MML estimation method turns out to be quite efficient and robust under departures from linearity. We also present some examples to illustrate all the methods of estimation discussed in this chapter. It should be mentioned here that the method of estimation of parameters discussed in this chapter is similar to a method given by Tan and Tabatabai (1988). But, the method presented in this chapter is based on a different method of identifying the normalized estimated residuals to be censored than the one used by Tan and Tabatabai (1988). Furthermore, in addition to deriving an estimator of σ the method presented in this chapter also enables one to determine the asymptotic variances and covariances of the estimators. Also, a very exhaustive comparison of this method of estimation with many other methods has been carried out under departures from normality as well as departures from linearity.

In Chapter 11, we extend the results of Chapter 10 to the multiple linear regression model. In Section 2, by forming a symmetrically Type-II censored sample from the normalized residuals, we derive the MML estimators of the parameters in the multiple linear regression model. These estimators turn out to be in the same form as the classical least-squares estimators. We describe the determination of the estimated normalized residuals that need to be censored for this estimation method. In Section 3, we derive the asymptotic variance-covariance matrix of these MML estimators through the information matrix and discuss the asymptotic distribution of these estimators. In Section 4, we compare the classical and the MML estimators to all other robust estimators mentioned earlier in the special case of the linear regression model with two explanatory variables. Through Monte Carlo simulations, we determine the bias and mean square error of all these estimators for various sample sizes under normal and various non-normal models for the error variable in the regression model. It is shown from this study that the MML estimators derived in this chapter are jointly highly efficient and are also quite robust to departure from normality. Finally, we consider an example given by Belsley, Kuh and Welsh (1980) and illustrate all the methods of estimation discussed in this chapter.

Finally, in Chapter 12 we explain briefly various contributions that have been made in this thesis in the areas of statistical classification and linear regression. Conclusions drawn from this work are also mentioned. Lastly, some problems of research that emerge from this thesis are also listed. These problems are of great interest theoretically as well as from a practical point of view.

Chapter 2

Multivariate Robust Linear Classification Procedure and Application

2.1 Introduction

As mentioned in the last chapter, after showing that the classical univariate linear classification procedure is quite sensitive to departures from normality Balakrishnan, Tiku and El Shaarawi (1985) used the Modified Maximum Likelihood (MML) estimators to develop a robust univariate linear classification procedure. In this chapter, we generalize the results of Balakrishnan, Tiku and El Shaarawi (1985) to the multivariate situation and for doing so we first extend the univariate MML estimators to the multivariate case. For the purpose of illustrating the classification procedure developed here, we consider the anthropometric data collected and given by Majumdar (1941). By applying the classical and the robust multivariate linear classification procedures to this data, we show that the robust procedure developed in this chapter performs better than the classical procedure. A similar comparison also reveals that the robust procedure performs much better than some of the known non-parametric or distribution-free procedures.

For the sake of clarity and a better understanding of the developments to

the multivariate situation that are to follow, we first describe briefly in Section 2 the univariate MML estimators based on Type-II censored samples derived by Tiku (1967, 1970, 1980). In Section 3 we present next a discussion on the classical univariate two-way linear classification procedure and the robust analogue based on the MML estimators as developed by Balakrishnan, Tiku and El Shaarawi (1985). In their work, they have shown that their procedure based on the MML estimators is quite robust and highly efficient (as compared to the classical classification procedure) under departures from normality. We consider two univariate data sets from Majumdar's (1941) anthropometric data and illustrate the classical and the robust univariate linear classification procedures.

For facilitating the generalization of the robust univariate linear classification procedure based on the MML estimators described in Section 3 to the multivariate situation, we first present in Section 4 the extension of the univariate MML estimation method to the multivariate situation as given by Tiku (1988) and Tiku and Balakrishnan (1988). Some asymptotic distributional properties of these multivariate MML estimators are also discussed in this section. In Section 5, we make use of these multivariate MML estimators to develop a robust multivariate linear classification procedure, analogous to the univariate linear classification procedure discussed already in Section 3. Here, we describe the asymptotic determination of the cut-off point when one of the errors of misclassification is at a prefixed level. Finally, in Section 6 we assess the performance of the classical and the robust multivariate linear classification procedures by applying them to Majumdar's anthropometric data after fixing one of the two errors of misclassification at a prefixed level. Through this comparison, we show that the robust procedure developed in Section 5 based on the MML estimators is more efficient than the classical classification procedure. Further, by not fixing either of the two errors of misclassification, i.e., allowing them to freely float, we apply the classical and the robust procedures on this data and determine the average error of misclassification for both the procedures. By comparing these values with the average error rates of some prominent nonparametric classification procedures, like the nearest-neighbour and the density estimation methods, we show that the robust procedure is once again better than the classical

procedure and also is considerably more efficient than the nonparametric methods.

As mentioned above, we have primarily used in this chapter the data obtained from an elaborative and extensive anthropometric survey conducted by Majumdar (1941) in the United Provinces of India in which about twelve to fourteen different physical measurements were made on each individual from 23 different castes and tribes. This data has earlier been statistically analyzed by Mahalanobis and Rao (1948) and also utilized by Rao (1948b) for illustrating the usefulness of the classical classification procedure in his extremely interesting, stimulating and pioneering paper that was read before the Research Section of the Royal Statistical Society. In his article, Rao has emphasized the main aim of anthropological classifications as, in the words of Morant (1939), "to unravel the course of human evolution, and it may be taken for granted to-day that the proper study of the natural history of man is concerned essentially with the mode and path of his descent".

2.2 Univariate Modified Maximum Likelihood (MML) Estimation

In this section, we briefly describe the MML estimators derived by Tiku (1967) based on symmetrically Type-II censored samples.

Let $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ be the order statistics obtained from a sample of size n from the normal $N(\mu, \sigma^2)$ population. Then, by considering the symmetrically Type-II censored sample $x_{(r+1)} \leq x_{(r+2)} \leq \ldots \leq x_{(n-r)}$ and writing down the likelihood function based on it as

$$L = \frac{n!}{(r!)^2}\sigma^{-(n-2r)} \exp\left\{-\frac{1}{2}\sum_{i=r+1}^{n-r} z_{(i)}^2\right\}\left\{\Phi(z_{(r+1)})\right\}^r\left\{1-\Phi(z_{(n-r)})\right\}^r, \qquad (2.1)$$

where $z_{(i)} = \frac{x_{(i)}-\mu}{\sigma}$, $\Phi(z) = \int_{-\infty}^{z}\phi(t)dt$, and $\phi(t) = \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$, Tiku (1967) has noted that the maximum likelihood equations

$$\frac{\partial \ln L}{\partial \mu} = 0 \quad \text{and} \quad \frac{\partial \ln L}{\partial \sigma} = 0$$

involve the functions

$$\psi_1(z_{(r+1)}) = \frac{\phi(z_{(r+1)})}{\Phi(z_{(r+1)})}$$

and

$$\psi_2(z_{(n-r)}) = \frac{\phi(z_{(n-r)})}{1 - \Phi(z_{(n-r)})}$$

and hence do not admit explicit solutions; determination of exact maximum likelihood estimates of μ and σ based on these Type-II censored samples has been discussed by Cohen (1959, 1961) and Harter and Moore (1966), and explained in great detail by Balakrishnan and Cohen (1990). Tiku (1967), therefore, has suggested the use of linear approximations

$$\psi_1(z_{(r+1)}) \approx \alpha - \beta z_{(r+1)} \quad \text{and} \quad \psi_2(z_{(n-r)}) \approx \alpha + \beta z_{(n-r)} \tag{2.2}$$

in order to simplify the maximum likelihood equations. The resulting equations

$$\frac{\partial \ln L^*}{\partial \mu} = 0 \quad \text{and} \quad \frac{\partial \ln L^*}{\partial \sigma} = 0$$

are referred to as the MML (Modified Maximum Likelihood) equations and they do admit explicit solutions. These solutions are referred to as the MML estimators of μ and σ and are given by

$$\hat{\mu} = \frac{1}{m}\left[\sum_{i=r+1}^{n-r} x_{(i)} + r\beta\left\{x_{(r+1)} + x_{(n-r)}\right\}\right] \tag{2.3}$$

and

$$\hat{\sigma} = \frac{B + (B^2 + 4AC)^{\frac{1}{2}}}{2\left\{A(A-1)\right\}^{\frac{1}{2}}}, \tag{2.4}$$

where

$$\begin{aligned} A &= n - 2r, \\ m &= n - 2r + 2r\beta, \\ B &= r\alpha\left\{x_{(n-r)} - x_{(r+1)}\right\}, \end{aligned}$$

and

$$C = \sum_{i=r+1}^{n-r} x_{(i)}^2 + r\beta\left\{x_{(r+1)}^2 + x_{(n-r)}^2\right\} - m\hat{\mu}^2. \tag{2.5}$$

The coefficients α and β depend only on the proportion of censoring, $q = \frac{r}{n}$, and are obtained from the equations

$$\beta = -\frac{\phi(t)\left\{t - \frac{\phi(t)}{q}\right\}}{q} \quad \text{and} \quad \alpha = \frac{\phi(t)}{q} - \beta t, \tag{2.6}$$

where $t = \Phi^{-1}(1 - q)$. Asymptotically, (2.2) are strict equalities (Tiku, 1968; Bhattacharyya, 1985) in which case $\hat{\mu}$ and $\hat{\sigma}$ in (2.3) and (2.4) are identical to the MLE of μ and σ, respectively.

The distributions of $\hat{\mu}$ and $\hat{\sigma}$ have been studied by Tiku (1980) in the sampling theory framework and by Tan (1985), Tan and Balakrishnan (1986, 1987), Balakrishnan (1988), and Balakrishnan and Tan (1988) from a Bayesian viewpoint. A book length account of the efficiency properties of $\hat{\mu}$ and $\hat{\sigma}$, and the robustness features of several inference procedures based on $\hat{\mu}$ and $\hat{\sigma}$, is available in Tiku, Tan and Balakrishnan (1986). One of the key results concerning the univariate MML estimators that is given by Tiku, Tan and Balakrishnan (1986) is presented in the following theorem.

Theorem 1 *Asymptotically (i.e. for large n)*

1. *$\hat{\mu}$ is distributed as $N_p(\mu, \frac{\sigma}{m})$;*
2. *$(A-1)\sigma^2$ is distributed as a Chi-squared with $A-1$ d.f.;*
3. *$\hat{\mu}$ and $n\hat{\sigma}$ are independent*

2.3 Univariate Classical and Robust Linear Classification Procedures and Application

In This section, we present a discussion on the classical univariate two-way linear classification procedure and the analogous robust two-way linear classification procedure based on the MML estimators proposed by Balakrishnan, Tiku and El Shaarawi (1985). We give a brief description of the determination of the asymptotic cut-off points of these two procedures when one of the two errors of misclassification is at a prefixed level. We also illustrate these two classification procedures with two univariate data sets taken from Majumdar's anthropometric data.

2.3.1 Classical Linear Classification Procedure

Let x_{1i}, $i = 1, 2, \ldots, n_1$, and x_{2i}, $i = 1, 2, \ldots, n_2$, be independent random samples from populations Π_1 and Π_2 assumed to be normal $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, respectively. Let us denote the sample means by

$$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i} \quad \text{and} \quad \bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i},$$

and the sample variances by

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 \quad \text{and} \quad s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2,$$

and the pooled sample variance by

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Then, to classify a new independent observation x_0, the classical linear classification procedure is to classify x_0 into Π_1 or Π_2 according as (Anderson, 1951)

$$V = \left\{x_0 - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)\right\}(\bar{x}_1 - \bar{x}_2)\frac{1}{s^2} \gtrless C, \tag{2.7}$$

where the cut-off point C is chosen such that one of the two errors of misclassification is at a pre-specified level.

Let us now denote the probability of wrongly classifying x_0 into Π_2 when x_0 actually belongs to Π_1 by e_{12} and similarly the probability of wrongly classifying x_0 into Π_1 when x_0 actually belongs to Π_2 by e_{21}. Note that $1 - e_{21}$ is simply the probability of correctly classifying x_0 into Π_2 when x_0 actually belongs to Π_2.

Even though $\bar{x}_1$ and $\bar{x}_2$ are independently distributed as $N(\mu_1, \frac{\sigma^2}{n_1})$ and $N(\mu_2, \frac{\sigma^2}{n_2})$, and $\frac{(n_1+n_2-2)s^2}{\sigma^2}$ is independently distributed as chi-square with $n_1 + n_2 - 2$ degrees of freedom, the distribution of V in (2.7) does not assume a manageable form; see, for example, Sinha and Giri (1975). It is, therefore, difficult to determine the cut-off point C in (2.7) for a specified e_{12}, for example. But, by using the first four moments of V worked out by Balakrishnan, Tiku and El Shaarawi (1985) one may fit a four-moment Pearson curve for the distribution of V by making use of the tables

of Johnson et al. (1963) and determine an approximate value for the cut-off point C. Alternatively, one may use the following asymptotic argument to approximate C.

Asymptotically (as $n_1, n_2 \to \infty$), $\bar{x}_1$, $\bar{x}_2$ and s tend to μ_1, μ_2 and σ, respectively, and hence the distribution of V in (2.7) is the same as that of Fisher's linear discriminant function

$$U = \left\{x_0 - \frac{1}{2}(\mu_1 + \mu_2)\right\}(\mu_1 - \mu_2)\frac{1}{\sigma^2} \tag{2.8}$$

The distribution of U is normal with mean

$$E(U|x_0 \in \Pi_1) = \frac{\delta^2}{2} = -E(U|x_0 \in \Pi_2) \tag{2.9}$$

and variance

$$Var(U|x_0 \in \Pi_1) = Var(U|x_0 \in \Pi_2) = \delta^2, \tag{2.10}$$

where $\delta^2 = \frac{(\mu_1 - \mu_2)^2}{\sigma^2}$ is the standardized squared distance between the two populations Π_1 and Π_2. Therefore, with e_{12} fixed as 0.05, for example, under the assumption of normality the cut-off point C in (2.7) may be determined approximately as

$$C = \frac{\delta^2}{2} - 1.64|\delta|. \tag{2.11}$$

Since δ is unknown, C can not be determined from (2.11); however, we may use an estimator of δ^2 given by (Lachenbruch and Mickey, 1968)

$$\hat{\delta}^2 = \left(\frac{n_1 + n_2 - 4}{n_1 + n_2 - 2}\right)\frac{(\bar{x}_1 - \bar{x}_2)^2}{s^2} \tag{2.12}$$

to obtain a suitable estimator of C from (2.11) to be

$$\hat{C} = \frac{\hat{\delta}^2}{2} - 1.64|\hat{\delta}|. \tag{2.13}$$

Then, the classical two-way linear classification procedure with e_{12} fixed as 0.05, is to classify the observation x_0 into Π_1 or Π_2 according as

$$V = \left\{x_0 - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)\right\}(\bar{x}_1 - \bar{x}_2)\frac{1}{s^2} \gtrless \hat{C}, \tag{2.14}$$

where the cut-off point $\hat{C}$ is as given in (2.13). It should be pointed out here that the estimator $\hat{\delta}^2$ in (2.12) is a biased estimator of δ^2 and the unbiased estimator of δ^2 given by $\hat{\delta}^2 - \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$ may take on inadmissible negative values.

2.3.2 Robust Linear Classification Procedure

A robust linear classification procedure analogous to the classical one in (2.14) has been proposed by Balakrishnan, Tiku and El Shaarawi (1985) by replacing $\bar{x}_1$, $\bar{x}_2$ and s by the corresponding MML estimators $\hat{\mu}_1$, $\hat{\mu}_2$ and $\hat{\sigma}$; here $\hat{\mu}_1$, $\hat{\sigma}_1$ are the MML estimators of μ_1 and σ obtained from the sample $x_{1i}, i = 1, 2, \ldots, n_1$, and $\hat{\mu}_2$, $\hat{\sigma}_2$ are the MML estimators of μ_2 and σ obtained from the sample $x_{2i}, i = 1, 2, \ldots, n_2$, and $\hat{\sigma}^2$ is the pooled MML estimator of σ^2 given by

$$\hat{\sigma}^2 = \frac{(A_1 - 1)\hat{\sigma}_1^2 + (A_2 - 1)\hat{\sigma}_2^2}{A_1 + A_2 - 2}. \tag{2.15}$$

Then the robust two-way linear classification procedure, with e_{12} fixed as 0.05, is to classify the observation x_0 into Π_1 or Π_2 according as

$$V_R = \left\{x_0 - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)\right\}(\hat{\mu}_1 - \hat{\mu}_2)\frac{1}{\hat{\sigma}^2} \gtrless \hat{C}_R, \tag{2.16}$$

where the cut-off point $\hat{C}_R$ is given by

$$\hat{C}_R = \frac{\hat{\delta}_R^2}{2} - 1.64|\hat{\delta}_R| \tag{2.17}$$

with $\hat{\delta}_R^2$ being the robust estimator of δ^2 given by

$$\hat{\delta}_R^2 = \left(\frac{A_1 + A_2 - 4}{A_1 + A_2 - 2}\right)\frac{(\hat{\mu}_1 - \hat{\mu}_2)^2}{\hat{\sigma}^2}. \tag{2.18}$$

Balakrishnan, Tiku and El Shaarawi (1985) have proved the following two theorems regarding the above classification procedure.

Theorem 2 *Under the assumption of normality, the distribution of V_R is exactly the same as that of V, for large $A_1 = n_1 - 2r_1$ and $A_2 = n_2 - 2r_2$.*

Theorem 3 *Under the assumption of normality, the $V-$ and V_R-procedures have the same $1 - e_{21}$ for a common e_{12}, for large A_1 and A_2 (hence large n_1 and n_2).*

Balakrishnan, Tiku and El Shaarawi (1985) have studied the robustness features of these two linear classification procedures under departures from normality. They have shown that the robust linear procedure has more stable e_{12} values and

considerably larger $1 - e_{21}$ values (the probability of correct classification) than the corresponding classical procedure under normal and a wide range of non-normal models. In the situation when neither of the two errors of misclassification is fixed at a pre-specified level, that is, when both the errors of misclassification are allowed to float freely, they have employed the two procedures with $C = 0$ and $C_R = 0$, respectively, and compared them with several well-known nonparametric procedures. Through this study, they have shown that the robust procedure based on MML estimators has the smallest average error rate $\frac{e_{12}+e_{21}}{2}$.

2.3.3 Illustrative Examples

In this section we consider two sets of data taken from the anthropometric data collected by Majumdar (1941) and illustrate the classical linear and the robust linear classification procedures described in Sections 2.3.1 and 2.3.2, respectively.

Example 2.1:

In Tables 2.1 and 2.2, we have given the measurements on nasal breadth of 106 individuals from the Agharia group and 180 individuals from the Bhil group, respectively.

Table 2.1: Measurements on nasal breadth of $n_1 = 106$ individuals from the Agharia group

i	x_{1i}	i	x_{1i}	i	x_{1i}	i	x_{1i}	i	x_{1i}	i	x_{1i}	i	x_{1i}
1	1683	17	1620	33	1683	49	1683	65	1671	81	1587	97	1748
2	1674	18	1655	34	1733	50	1558	66	1732	82	1707	98	1677
3	1663	19	1609	35	1738	51	1560	67	1645	83	1687	99	1750
4	1802	20	1645	36	1712	52	1705	68	1634	84	1647	100	1637
5	1765	21	1746	37	1664	53	1675	69	1599	85	1546	101	1715
6	1606	22	1690	38	1690	54	1674	70	1698	86	1730	102	1602
7	1709	23	1662	39	1673	55	1689	71	1539	87	1725	103	1698
8	1678	24	1613	40	1684	56	1650	72	1620	88	1674	104	1659
9	1647	25	1738	41	1624	57	1680	73	1713	89	1718	105	1530
10	1643	26	1730	42	1655	58	1687	74	1800	90	1748	106	1679
11	1754	27	1667	43	1743	59	1652	75	1770	91	1715		
12	1625	28	1674	44	1694	60	1808	76	1768	92	1607		
13	1590	29	1662	45	1637	61	1622	77	1585	93	1624		
14	1650	30	1666	46	1745	62	1636	78	1661	94	1814		
15	1630	31	1675	47	1699	63	1653	79	1735	95	1585		
16	1676	32	1632	48	1646	64	1689	80	1689	96	1680		

Table 2.2: Measurements on nasal breadth of $n_2 = 180$ individuals from the Bhil group

i	x_{2i}	i	x_{2i}	i	x_{2i}	i	x_{2i}	i	x_{2i}	i	x_{2i}	i	x_{2i}
1	1668	27	1678	53	1611	79	1670	105	1699	131	1716	157	1590
2	1597	28	1706	54	1640	80	1570	106	1616	132	1644	158	1617
3	1729	29	1620	55	1604	81	1562	107	1579	133	1614	159	1644
4	1606	30	1656	56	1543	82	1523	108	1546	134	1575	160	1632
5	1713	31	1760	57	1708	83	1591	109	1587	135	1664	161	1593
6	1686	32	1656	58	1552	84	1663	110	1600	136	1595	162	1762
7	1607	33	1697	59	1634	85	1618	111	1628	137	1631	163	1651
8	1522	34	1637	60	1687	86	1566	112	1636	138	1585	164	1619
9	1674	35	1628	61	1608	87	1580	113	1607	139	1682	165	1594
10	1610	36	1563	62	1673	88	1600	114	1645	140	1653	166	1634
11	1706	37	1653	63	1662	89	1632	115	1607	141	1656	167	1665
12	1608	38	1616	64	1511	90	1592	116	1642	142	1570	168	1613
13	1664	39	1657	65	1676	91	1700	117	1662	143	1708	169	1592
14	1567	40	1642	66	1598	92	1622	118	1574	144	1647	170	1646
15	1590	41	1656	67	1707	93	1631	119	1696	145	1624	171	1698
16	1711	42	1557	68	1690	94	1651	120	1695	146	1621	172	1592
17	1643	43	1707	69	1654	95	1510	121	1735	147	1626	173	1480
18	1603	44	1630	70	1615	96	1776	122	1608	148	1605	174	1672
19	1558	45	1619	71	1540	97	1588	123	1668	149	1605	175	1587
20	1755	46	1652	72	1654	98	1638	124	1540	150	1654	176	1615
21	1625	47	1628	73	1690	99	1560	125	1662	151	1713	177	1568
22	1675	48	1621	74	1605	100	1575	126	1644	152	1578	178	1580
23	1634	49	1540	75	1584	101	1624	127	1740	153	1635	179	1696
24	1668	50	1626	76	1565	102	1663	128	1739	154	1715	180	1586
25	1570	51	1555	77	1575	103	1541	129	1660	155	1608		
26	1563	52	1558	78	1760	104	1605	130	1616	156	1681		

From these two samples, we find:

$$\bar{x}_1 = 1674.6981, \quad \bar{x}_2 = 1630.1611, \quad s_1^2 = 3336.8381, \quad s_2^2 = 3060.0225,$$
$$\hat{\mu}_1 = 1675.2760, \quad \hat{\mu}_2 = 1629.9518, \quad \hat{\sigma}_1^2 = 3087.2207, \quad \hat{\sigma}_2^2 = 3151.6277,$$
$$\hat{C} = -0.98286 \qquad \text{and} \qquad \hat{C}_R = -0.99769.$$

So, with e_{12} (the probability of wrongly classifying an individual from the Agharia group into the Bhil group) fixed as 0.05, by applying the classical linear classification procedure in (2.14) and reclassifying each individual in Tables 2.1 and 2.2, we obtain estimates of e_{12} and $1 - e_{21}$ (the probability of correctly classifying an individual from the Bhil group into itself) to be 0.047 and 0.194, respectively. Similarly, by applying the robust linear classification procedure in (2.16) and reclassifying each individual in Tables 2.1 and 2.2, we obtain estimates of e_{12} and $1 - e_{21}$ to be 0.047 and 0.439, respectively. We observe that, for e_{12} fixed as 0.05, the robust procedure has a considerably larger $1 - e_{21}$ (correct classification into the Bhil group) value than the classical procedure

Example 2.2:

In Tables 2.3 and 2.4, we have given the measurements on nasal breadth of 85 individuals from the Basti-Brahmin group and 98 individuals from the Oraon group, respectively.

Table 2.3: Measurements on nasal breadth of $n_1 = 85$ individuals from the Basti-Brahmin group

i	x_{1i}	i	x_{1i}	i	x_{1i}	i	x_{1i}	i	x_{1i}	i	x_{1i}	i	x_{1i}
1	1564	14	1610	27	1617	40	1677	53	1584	66	1645	79	1690
2	1710	15	1675	28	1542	41	1681	54	1670	67	1578	80	1653
3	1639	16	1604	29	1589	42	1644	55	1575	68	1667	81	1580
4	1668	17	1717	30	1699	43	1700	56	1625	69	1622	82	1680
5	1675	18	1713	31	1599	44	1605	57	1695	70	1700	83	1652
6	1679	19	1703	32	1690	45	1732	58	1670	71	1666	84	1614
7	1598	20	1605	33	1694	46	1602	59	1716	72	1605	85	1742
8	1672	21	1637	34	1640	47	1717	60	1631	73	1688		
9	1620	22	1525	35	1678	48	1642	61	1565	74	1674		
10	1674	23	1609	36	1682	49	1642	62	1618	75	1678		
11	1652	24	1576	37	1593	50	1655	63	1597	76	1728		
12	1620	25	1630	38	1695	51	1622	64	1650	77	1668		
13	1639	26	1456	39	1692	52	1689	65	1667	78	1635		

Table 2.4: Measurements on nasal breadth of $n_2 = 98$ individuals from the Oraon group

i	x_{2i}	i	x_{2i}	i	x_{2i}	i	x_{2i}	i	x_{2i}	i	x_{2i}	i	x_{2i}
1	1655	15	1632	29	1627	43	1623	57	1588	71	1620	85	1572
2	1617	16	1600	30	1682	44	1644	58	1666	72	1619	86	1460
3	1553	17	1609	31	1572	45	1562	59	1592	73	1550	87	1628
4	1635	18	1670	32	1558	46	1580	60	1700	74	1657	88	1552
5	1652	19	1678	33	1533	47	1737	61	1560	75	1692	89	1588
6	1694	20	1536	34	1694	48	1702	62	1644	76	1557	90	1606
7	1688	21	1681	35	1611	49	1599	63	1608	77	1596	91	1638
8	1641	22	1647	36	1467	50	1563	64	1567	78	1540	92	1582
9	1520	23	1778	37	1591	51	1515	65	1635	79	1620	93	1566
10	1596	24	1666	38	1644	52	1532	66	1490	80	1590	94	1667
11	1646	25	1656	39	1660	53	1577	67	1640	81	1629	95	1689
12	1685	26	1516	40	1620	54	1700	68	1625	82	1600	96	1616
13	1614	27	1565	41	1670	55	1728	69	1630	83	1614	97	1654
14	1617	28	1627	42	1630	56	1640	70	1670	84	1480	98	1607

From these two samples, we find:

$$\bar{x}_1 = 1646.0706, \quad \bar{x}_2 = 1615.7041, \quad s_1^2 = 2610.4761, \quad s_2^2 = 3507.9587$$
$$\hat{\mu}_1 = 1647.6534, \quad \hat{\mu}_2 = 1617.0022, \quad \hat{\sigma}_1^2 = 2439.1426, \quad \hat{\sigma}_2^2 = 3396.6770,$$
$$\hat{C} = -0.74323 \qquad \text{and} \qquad \hat{C}_R = -0.76152.$$

So, with e_{12} (the probability of wrongly classifying an individual from the Basti-Brahmin group into the Oraon group) fixed as 0.05, by applying the classical linear classification procedure in (2.14) and reclassifying each individual in Tables 2.3 and 1.4, we obtain estimates of e_{12} and $1 - e_{21}$ (the probability of correctly classifying an individual from the Oraon group into itself) to be 0.035 and 0.143, respectively. Similarly, by applying the robust linear classification procedure in (2.16) and reclassifying each individual in Tables 2.3 and 2.4, we obtain estimates of e_{12} and $1 - e_{21}$ to be 0.035 and 0.163, respectively.

2.4 Multivariate Modified Maximum Likelihood Estimation

In this section, we present the generalization of the modified maximum likelihood estimators of μ and σ of a univariate normal distribution to the mean vector μ and the variance-covariance matrix Σ of a multivariate normal distribution as given by Tiku and Balakrishnan (1988). The extension to the bivariate case has earlier been carried out by Tiku (1988). We also present some asymptotic distributional properties of these multivariate MML estimators.

Theorem 4 *Let $\mathbf{x}_i, i = 1, 2, \ldots, n$, be a random sample from a $p-$variate normal distribution $N_p(\mu, \Sigma)$ and let $\mathbf{X}$ be the sample matrix of size $p \times n$. Then the MML estimators of μ and Σ are given by*

$$\hat{\mu} = \frac{1}{m} \sum_{i=r+1}^{n-r} w_i \mathbf{x}_{(i)}$$

$$\hat{\Sigma}^* = \frac{1}{n} \sum_{i=r+1}^{n-r} w_i (\mathbf{x}_{(i)} - \hat{\mu})(\mathbf{x}_{(i)} - \hat{\mu})^T$$

where $\mathbf{x}_{(i)}$ *is the ith column of the matrix obtained from arranging each row of* $\mathbf{X}$ *in an ascending order of magnitude an :*

$$w_i = \begin{cases} 1 + r\beta & for \quad i = r+1 \\ 1 & for \quad r+1 < i < n-r-1 \\ 1 + r\beta & for \quad i = n-r-1 \end{cases}$$

Proof: The likelihood function of the sample is

$$\begin{aligned} L &= \left\{ \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \right\}^n \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \mu)^T \Sigma^{-1} (\mathbf{x}_i - \mu) \right\} \\ &= \left\{ \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \right\}^n \exp\left\{ -\frac{1}{2|\mathbf{R}|} \sum_{k=1}^{p} R_{kk} \sum_{i=1}^{n} \frac{(x_{ik} - \mu_k)^2}{\sigma_k^2} \right. \\ &\quad \left. -\frac{1}{2|\mathbf{R}|} \sum_{k=1}^{p} \sum_{\substack{l=1 \\ l \neq k}}^{p} R_{kl} \sum_{i=1}^{n} \frac{(x_{ik} - \mu_k)(x_{il} - \mu_l)}{\sigma_k \sigma_l} \right\}, \end{aligned} \tag{2.19}$$

where $\mathbf{R}$ is the $p \times p$ correlation matrix and R_{kl} is the co-factor of the (k,l)th element in $\mathbf{R}$. Let us now denote $x_{(1)k} \leq x_{(2)k} \leq \ldots \leq x_{(n)k}$ $(k = 1, 2, \ldots, p)$ for the order statistics of the x_{ik} $(1 \leq i \leq n, 1 \leq k \leq p)$ observations. Then we have

$$\begin{aligned} \sum_{i=1}^{n} (x_{ik} - \mu_k)(x_{il} - \mu_l) &= \sum_{i=1}^{n} (x_{ik} - \hat{\mu}_k + \hat{\mu}_k - \mu_k)(x_{il} - \hat{\mu}_l + \hat{\mu}_l - \mu_l) \\ &= \sum_{i=1}^{n} (x_{ik} - \hat{\mu}_k)(x_{il} - \hat{\mu}_l) + n(\hat{\mu}_k - \mu_k)(\hat{\mu}_l - \mu_l) \\ &\quad + n(\bar{x}_k - \hat{\mu}_k)(\hat{\mu}_l - \mu_l) + n(\bar{x}_l - \hat{\mu}_l)(\hat{\mu}_k - \mu_k), \end{aligned} \tag{2.20}$$

where

$$\hat{\mu}_k = \frac{1}{m} \sum_{i=r+1}^{n-r} w_i x_{(i)k}.$$

But

$$\begin{aligned} &\sum_{i=1}^{n} (x_{ik} - \hat{\mu}_k)(x_{il} - \hat{\mu}_l) \\ &= \sum_{i=1}^{n} (x_{(i)k} - \hat{\mu}_k)(x_{(i)l} - \hat{\mu}_l) \\ &= C_{kl} - r\beta[(x_{(r+1)k} - \hat{\mu}_k)(x_{(r+1)l} - \hat{\mu}_l) + (x_{(n-r)k} - \hat{\mu}_k)(x_{(n-r)l} - \hat{\mu}_l)] \\ &\quad + \sum_{i=1}^{r} (x_{(i)k} - \hat{\mu}_k)(x_{(i)l} - \hat{\mu}_l) + \sum_{i=n-r+1}^{n} (x_{(i)k} - \hat{\mu}_k)(x_{(i)l} - \hat{\mu}_l), \end{aligned} \tag{2.21}$$

where

$$C_{kl} = \sum_{i=r+1}^{n-r} w_i(x_{(i)k} - \hat{\mu}_k)(x_{(i)l} - \hat{\mu}_l).$$

Therefore

$$\sum_{i=1}^{n}(x_{ik} - \mu_k)(x_{il} - \mu_l) \;=\; C_{kl} + m(\hat{\mu}_k - \mu_k)(\hat{\mu}_l - \mu_l) + \sigma_k\sigma_l D_{kl}, \qquad (2.22)$$

where

$$\begin{aligned}
\sigma_k\sigma_l D_{kl} \;=\; & -r\beta\left[(x_{(r+1)k} - \hat{\mu}_k)(x_{(r+1)l} - \hat{\mu}_l) + (x_{(n-r)k} - \hat{\mu}_k)(x_{(n-r)l} - \hat{\mu}_l)\right] \\
& + \sum_{i=1}^{r}(x_{(i)k} - \hat{\mu}_k)(x_{(i)l} - \hat{\mu}_l) + \sum_{i=n-r+1}^{n}(x_{(i)k} - \hat{\mu}_k)(x_{(i)l} - \hat{\mu}_l) \\
& + n(\bar{x}_k - \hat{\mu}_k)(\hat{\mu}_l - \mu_l) + n(\bar{x}_l - \hat{\mu}_l)(\hat{\mu}_k - \mu_k) \\
& + (n-m)(\hat{\mu}_k - \mu_k)(\hat{\mu}_l - \mu_l). \qquad (2.23)
\end{aligned}$$

Note that this relation holds for $l = k$ also. Tiku (1988) has shown that, for large n, $\frac{D_{kl}}{n}$ converges to its expected value $\rho_{kl}\left\{2q(1+\alpha t) + \frac{1}{n}(1 - \frac{n}{m})\right\}$. The modified likelihood function L^* is derived on lines exactly similar to the ones discussed by Tiku (1967) and Persson and Rootzen (1977) in the univariate case, by replacing D_{kk} and D_{kl} by their expected values, for large n. Thus,

$$L \;\approx\; L^*,$$

where

$$\begin{aligned}
L^* \;=\; & \left\{\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\right\}^n \exp\left\{-\frac{m}{2|\mathbf{R}|}\right. \\
& \left[\sum_{k=1}^{p} R_{kk}\frac{(\hat{\mu}_k - \mu_k)^2}{\sigma_k^2} + \sum_{k=1}^{p}\sum_{\substack{l=1\\ l\neq k}}^{p} R_{kl}\frac{(\hat{\mu}_k - \mu_k)(\hat{\mu}_l - \mu_l)}{\sigma_k\sigma_l}\right] \\
& \left. - \frac{1}{2|\mathbf{R}|}\left[\sum_{k=1}^{p}\frac{R_{kk}C_{kk}}{\sigma_k^2} + \sum_{k=1}^{p}\sum_{\substack{l=1\\ l\neq k}}^{p}\frac{R_{kl}C_{kl}}{\sigma_k\sigma_l}\right] - \frac{p}{2}\left[2r(1+\alpha t) + (1 - \frac{n}{m})\right]\right\} \\
\propto\; & \left\{\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\right\}^n \exp\left\{-\frac{m}{2|\mathbf{R}|}\right. \\
& \left[\sum_{k=1}^{p} R_{kk}\frac{(\hat{\mu}_k - \mu_k)^2}{\sigma_k^2} + \sum_{k=1}^{p}\sum_{\substack{l=1\\ l\neq k}}^{p} R_{kl}\frac{(\hat{\mu}_k - \mu_k)(\hat{\mu}_l - \mu_l)}{\sigma_k\sigma_l}\right]
\end{aligned}$$

$$-\frac{1}{2|\mathbf{R}|}\left[\sum_{k=1}^{p}\frac{R_{kk}C_{kk}}{\sigma_k^2}+\sum_{k=1}^{p}\sum_{\substack{l=1\\ l\neq k}}^{p}\frac{R_{kl}C_{kl}}{\sigma_k\sigma_l}\right]\Bigg\}, \tag{2.24}$$

since the term $\frac{p}{2}\left[2r(1+\alpha t)+(1-\frac{n}{m})\right]$ is free of the parameters. Using a simple algebraic manipulation L^* can be written as

$$\begin{aligned} L^* \;\propto\; & \left\{\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\right\}^n \\ & \exp\left\{-\frac{1}{2}\left[(\hat{\mu}-\mu)\left(\frac{\Sigma}{m}\right)^{-1}(\hat{\mu}-\mu)^T\right.\right. \\ & \left.\left.+trace\left\{\Sigma^{-1}\sum_{i=r+1}^{n-r}w_i(\mathbf{x}_{(i)}-\hat{\mu})(\mathbf{x}_{(i)}-\hat{\mu})^T\right\}\right]\right\}. \end{aligned} \tag{2.25}$$

We conclude the proof by following an argument similar to the one used for maximum likelihood estimators, see for example Anderson (1984). ∎

It should be mentioned here that the representation of the modified likelihood function L^* given in Eq. (2.25) is a much more elegant representation (in terms of trace) than the representation given by Tiku and Balakrishnan (1988).

It is of interest to mention here that $\frac{1}{n}\left[\ln L-\ln L^*\right]\to 0$ as $n\to\infty$. We also note that MML estimator of μ is unbiased but the MML estimator of Σ is biased for Σ. But, as pointed out by Tiku and Balakrishnan (1988) while discussing the bivariate case and by Tiku and Balakrishnan (1988) while discussing the multivariate case, the estimator obtained by multiplying the diagonal elements by $\frac{n}{A-1}$ and off-diagonal elements by $\frac{nA}{m(A-1)}$ of Σ^* is an asymptotically unbiased estimator of Σ. The variances and covariances of these MML estimators may be derived, as pointed out by Tiku and Balakrishnan (1988), from the information matrix as explained in the following theorem.

Theorem 5 *Asymptotically, that is, for large $A=n-2r$, the variance-covariances of $\hat{\mu}_k,\hat{\sigma}_k$ and $\hat{\rho}_{kl}$ are given by the elements of the matrix*

$$((J_{kl}))^{-1},$$

where $J_{11}=E\left(-\frac{\partial^2\ln L^*}{\partial\mu_1^2}\right), J_{12}=E\left(-\frac{\partial^2\ln L^*}{\partial\mu_1\partial\mu_2}\right)$, *etc.*,

Proof: Since the modified likelihood function L^* is almost identical to the likelihood function L, the large-sample variances and covariances of $\hat{\mu}_k, \hat{\sigma}_k$ and $\hat{\rho}_{kl}$ may be obtained via the information matrix of the modified likelihood function. Hence the theorem. ∎

In the context of robustness, r is chosen to be the integer value $r = [0.5 + 0.1n]$; see, for example, Tiku, Tan and Balakrishnan (1986). In this case, we have $q = \frac{r}{n} = 0.1, \beta = 0.8309, m = n - 2r + 2r\beta = n(1 - 2q + 2q\beta)$ and $A = n - 2r = n(1 - 2q)$, and hence $\sqrt{\frac{m}{A}} \approx 1.1$.

The following theorem, due to Tiku and Balakrishnan (1988), gives some important asymptotic distributional properties of the multivariate MML estimators discussed in this section.

Theorem 6 *Asymptotically (i.e. for large n)*

1. *$\hat{\mu}$ is distributed as $N_p(\mu, \frac{\Sigma}{m})$;*
2. *$n\Sigma^*$ is distributed as a Wishart random matrix with $A - 1$ d.f.;*
3. *$\hat{\mu}$ and $n\hat{\Sigma}^*$ are independent*

Proof: This theorem follows from the fact that the modified likelihood function can be factorized into two terms, one term being the multivariate normal density function and the other term being the density function of the Wishart matrix. ∎

2.5 Multivariate Classical and Robust Linear Classification Procedures

In this section we first describe the multivariate classical two-way linear classification procedure and explain the asymptotic determination of the cut-off point when one of the two errors of misclassification is at a prefixed level. Next, by using the multivariate MML estimators of μ and Σ of the multivariate normal distribution described in Section 2.4, we develop a multivariate robust two-way linear classification

procedure. By using the asymptotic properties of these estimators, we determine the cut-off point of this procedure (when one of the two errors of misclassification is at a prefixed level) through asymptotic arguments. This is a generalization of the univariate robust procedure based on the MML estimators given by Balakrishnan, Tiku and El Shaarawi (1985) and explained briefly in Section 3.

2.5.1 Classical Linear Classification Procedure

Let us assume $\mathbf{x}_{1k}$ $(k = 1, 2, \ldots, n_1)$ and $\mathbf{x}_{2k}$ $(k = 1, 2, \ldots, n_2)$ to be independent random samples from populations Π_1 and Π_2 which are taken to be $N_p(\mu_1, \Sigma)$ and $N_p(\mu_2, \Sigma)$, respectively. Let us denote the sample means by

$$\bar{\mathbf{x}}_1 = \frac{1}{n_1}\sum_{k=1}^{n_1}\mathbf{x}_{1k} \quad \text{and} \quad \bar{\mathbf{x}}_2 = \frac{1}{n_2}\sum_{k=1}^{n_2}\mathbf{x}_{2k}$$

and the sample variance-covariance matrices by

$$\mathbf{S}_1 = \frac{1}{n_1 - 1}\sum_{k=1}^{n_1}(\mathbf{x}_{1k} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1k} - \bar{\mathbf{x}}_1)^T$$

and

$$\mathbf{S}_2 = \frac{1}{n_2 - 1}\sum_{k=1}^{n_2}(\mathbf{x}_{2k} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2k} - \bar{\mathbf{x}}_2)^T$$

and the pooled sample variance-covariance matrix by

$$\mathbf{S} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}.$$

Then, to classify a new independent observation $\mathbf{x}_0$, the classical linear classification procedure is to classify $\mathbf{x}_0$ into Π_1 or Π_2 according as (Rao (1948a,b), Anderson (1951))

$$V = \left\{\mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)\right\}^T \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \gtrless C, \tag{2.26}$$

where, as before, the cut-off point C is determined such that one of the two errors of misclassification (say e_{12}) is at a pre-specified level. Although $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ are independently distributed as $N_p(\mu_1, \frac{1}{n_1}\Sigma)$ and $N_p(\mu_2, \frac{1}{n_2}\Sigma)$, respectively, and $(n_1 + n_2 - 2)\mathbf{S}$ is independently distributed as Wishart $W_p(n_1 + n_2 - 2, \Sigma)$, the distribution

of the statistic V in (2.26) does not assume a manageable form. Several authors have derived different representations and series expansions for the distribution of V; see, for example, Wald (1944), Anderson (1951), Sitgreaves (1952), Bowker (1960), Kabe (1963), Okamoto (1963), Dunn and Varady (1966), and Sinha and Giri (1975). Hence, it is very difficult to evaluate the cut-off point C exactly for a specified value of e_{12}. In the asymptotic situation, however, $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2$ and $\mathbf{S}$ tend to their respective expected values μ_1, μ_2 and Σ, and as a result the distribution of V in (2.26) is the same as that of Fisher's linear discriminant function

$$U = \left\{\mathbf{x}_0 - \frac{1}{2}(\mu_1 + \mu_2)\right\}^T \Sigma^{-1}(\mu_1 - \mu_2). \tag{2.27}$$

The statistic U in (2.27) is a linear function of $\mathbf{x}_0$ and hence has a $p-$variate normal distribution with mean

$$E(U|\mathbf{x}_0 \in \Pi_1) = \frac{\delta^2}{2} = -E(U|\mathbf{x}_0 \in \Pi_2) \tag{2.28}$$

and variance

$$Var(U|\mathbf{x}_0 \in \Pi_1) = Var(U|\mathbf{x}_0 \in \Pi_2) = \delta^2, \tag{2.29}$$

where $\delta^2 = (\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2)$ is the standardized squared distance between the two populations Π_1 and Π_2. Therefore, with e_{12} fixed as α, for example, under the assumption of normality the cut-off point C in (2.26) may be determined approximately as

$$C = \frac{\delta^2}{2} - z_{(\alpha)}|\delta|, \tag{2.30}$$

where $z_{(\alpha)}$ is the upper α percentage point of the standard normal distribution. Since δ^2 is unknown in practice, we replace C in (2.30) by an estimate

$$\hat{C} = \frac{\hat{\delta}^2}{2} - z_{(\alpha)}|\hat{\delta}|. \tag{2.31}$$

By realizing now that

$$\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$
$$\rightsquigarrow \text{ Hotelling's } T_p^2\left(n_1 + n_2 - 2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} \delta^2\right),$$

we may obtain an unbiased estimator of δ^2 as

$$\hat{\delta}^2 - \left(\frac{1}{n_1} + \frac{1}{n_2}\right) p, \tag{2.32}$$

where

$$\hat{\delta}^2 = \frac{n_1 + n_2 - p - 3}{n_1 + n_2 - 2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2). \tag{2.33}$$

But as mentioned by Lachenbruch and Mickey (1968), the unbiased estimator of δ^2 in (2.32) can take on inadmissible negative values and hence the biased estimator $\hat{\delta}^2$ in (2.33) itself serves as a suitable estimator of δ^2. Then, the classical two-way linear classification procedure, with e_{12} fixed as α, is to classify the observation $\mathbf{x}_0$ into Π_1 or Π_2 according as

$$V = \left\{\mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)\right\}^T \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \gtrless \hat{C}, \tag{2.34}$$

where the cut-off point $\hat{C}$ is as given in (2.31).

2.5.2 Robust Linear Classification Procedure

A robust linear classification procedure analogous to the classical procedure in (2.34) is developed here by making use of the multivariate MML estimators described in Section 2.4. Let $(\hat{\mu}_1, \hat{\Sigma}_1)$ and $(\hat{\mu}_2, \hat{\Sigma}_2)$ be the MML estimators of (μ_1, Σ) and (μ_2, Σ) obtained from the samples $\mathbf{x}_{1k}, k = 1, 2, \ldots, n_1$, and $\mathbf{x}_{2k}, k = 1, 2, \ldots, n_2$, respectively. Let $\hat{\Sigma}$ be the pooled MML estimator of Σ given by

$$\hat{\Sigma} = \frac{(A_1 - 1)\hat{\Sigma}_1 + (A_2 - 1)\hat{\Sigma}_2}{A_1 + A_2 - 2}. \tag{2.35}$$

Then the robust classification procedure for classifying a new observation $\mathbf{x}_0$ is to classify it into Π_1 or Π_2 according as

$$V_R = \left\{\mathbf{x}_0 - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)\right\}^T \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_2) \gtrless C_R, \tag{2.36}$$

where the cut-off point C_R is determined so that the error of misclassification e_{12} is fixed as α.

It may be noted from the modified likelihood function L^* in (2.24) that asymptotically $\hat{\mu}_1$ and $\hat{\mu}_2$ have independent normal distributions $N_p(\mu_1, \frac{1}{m_1}\Sigma)$ and $N_p(\mu_2, \frac{1}{m_2}\Sigma)$, respectively, and $(A_1 + A_2 - 2)\hat{\Sigma}$ is independently distributed as Wishart $W_p(A_1 + A_2 - 2, \Sigma)$. In the asymptotic situation, $\hat{\mu}_1$, $\hat{\mu}_2$ and $\hat{\Sigma}$ tend to their respective expected values μ_1, μ_2 and Σ and as a result the distribution of V_R in (2.36) is the same as that of Fisher's linear discriminant function U in (2.27). Consequently, we replace the cut-off point C_R in (2.36) by an estimate

$$\hat{C}_R = \frac{\hat{\delta}_R^2}{2} - z_{(\alpha)}|\hat{\delta}_R|, \tag{2.37}$$

with e_{12} fixed as α. A robust unbiased estimator of δ^2 is given by

$$\hat{\delta}_R^2 - \left(\frac{1}{m_1} + \frac{1}{m_2}\right)p, \tag{2.38}$$

where

$$\hat{\delta}_R^2 = \frac{A_1 + A_2 - p - 3}{A_1 + A_2 - 2}(\hat{\mu}_1 - \hat{\mu}_2)^T\hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_2). \tag{2.39}$$

As mentioned earlier in Section 2.5.1, the robust unbiased estimator of δ^2 in (2.38) can take on inadmissible negative values and hence the biased estimator $\hat{\delta}_R^2$ in (2.39) itself serves as a suitable estimator of δ^2. Then, the robust two-way linear classification procedure, with e_{12} fixed as α, is to classify the observation $\mathbf{x}_0$ into Π_1 or Π_2 according as

$$V_R = \left\{\mathbf{x}_0 - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)\right\}^T \mathbf{S}^{-1}(\hat{\mu}_1 - \hat{\mu}_2) \gtrless \hat{C}_R, \tag{2.40}$$

where the cut-off point $\hat{C}_R$ is as given in (2.37). It should be mentioned here that Tiku and Balakrishnan (1984) proposed robust linear classification procedure based on the univariate MML estimators for the bivariate situation. For that purpose, they had transformed the data by using regression coefficients and written the linear classification statistic for the bivariate case as a sum of two (conditionally independent) univariate linear classification statistics. The extension of Tiku and Balakrishnan's method to higher dimension, therefore, becomes quite complicated structurally and also will involve great deal of computation. But, the generalization presented in this section is direct extension of the univariate procedure presented in

Section 2.3, is analogous to the multivariate classical linear procedure, and is also based on the multivariate MML estimators described in Section 2.4. Hence, it becomes more appealing and useful in practical situations involving higher dimensions as will be illustrated in the next section.

2.6 Application to Majumdar's Anthropometric Data

In this section, we apply the multivariate classical and robust linear classification procedures discussed already in Section 2.5 to the anthropometric data of Majumdar (1941). Through an assessment of the performance of these two procedures, with one of the two errors of misclassification at a prefixed level, we show that the robust procedure performs better than the classical procedure. Further, by allowing both the errors of misclassification to freely float, we employ the classical and the robust procedures and then compare them with nonparametric methods like the nearest-neighbour and the density-estimation methods. We then show that the multivariate robust linear classification procedure derived in Section 2.5 (with both errors being free) has the smallest average error rate among all these procedures.

For the sake of completeness and also a better understanding of this application, we first present a brief description of the data and then regarding the validity of the assumption of multivariate normal distribution for this data.

2.6.1 Description of the characters

The total number of individual measurements on any individual reported by Majumdar (1941) did not exceed 16, but usually the first 12 measurements of the following 14 were taken: (1) Stature, (2) Sitting height, (3) Head length, (4) Head breadth, (5) Minimum frontal breadth, (6) Nasal length, (7) Nasal breadth, (8) Nasal depth, (9) Total facial length, (10) Upper facial length, (11) Bizygomatic breadth, (12) Bigonial breadth, (13) Head height, and (14) Weight. In some cases, measurements on (15) Orbito-nasal breadth and (16) Orbito-nasal arc were taken and also the (17)

Auricular height and (18) Span. A brief description of the characters measured may be secured from Majumdar (1941).

The various castes and tribes that have been included in the anthropometric survey of Majumdar (1941) are as follows: (1) Basti-Brahmins, (2) Other Brahmins, (3) Agharias, (4) Chattris, (5) Muslims, (6) Bhatus, (7) Habrus, (8) Bhils, (9) Doms, (10) Artisan Ahirs, (11) Artisan Kurmis, (12) Other Artisans, (13) Artisan Kahars, (14) Tharu males, (15) Chamars, (16) Cheros, (17) Majhwars, (18) Panikas, (19) Kharwars, (20) Oraons, (21) Rajwars, (22) Korwas, and (23) Tharu females. A brief report about the various castes and tribes listed above may be had from the works Risley (1891), Roy (1915), Majumdar (1941, 1944), or the supplementary ethnological note by Mahalanobis (1948). Should a more detailed account of the castes and tribes of the United Provinces considered above be required, one may refer to the volumes of Sir William Crooke (1896).

2.6.2 Testing for multivariate normality

The validity of the assumption of multivariate normality for the population distributions of all groups was first investigated by looking at the frequency distributions of individual characters with the help of β_1 and β_2, the coefficients of skewness and kurtosis, respectively. The plausible departure from normality in skewness was tested with the help of the statistic

$$Z_1 = \left\{\frac{(n+1)(n+3)}{6(n-2)}\right\}^{\frac{1}{2}} \sqrt{\beta}_1, \tag{2.41}$$

while the departure from normality in kurtosis was tested by using the statistic

$$Z_2 = \left\{\frac{(n+1)^2(n+3)(n+5)}{24n(n-2)(n-3)}\right\}^{\frac{1}{2}} \left[\beta_2 - 3 + \frac{6}{n+1}\right]; \tag{2.42}$$

for a large value of n, both these variables are referred to the standard normal distribution.

By using the statistic Z_1, we noticed that in most cases there was a general tendency towards negative skewness and a significant positive skewness only in one or two cases. Overall, the deviations from symmetry, even when significant, were

all small in magnitude, the largest values of $\sqrt{\beta_1}$ being for Head Length of Bhatu, Bizygomatic Breadth of Chattri, and Nasal depth Kurmi.

Similarly, by using the statistic Z_2, we observed that Nasal Depth, Sitting Height, Nasal Length, Head Length, Head Breadth and Bizygomatic Breadth had the largest number of deviations with respect to the kurtosis. With regard to the groups, we also observed that Charmar, Other Artisans, Bhatu, Chattri and Kharwar had the largest number of deviations. Moreover, Nasal Depth, Sitting Height and Head Breadth had the largest number of deviations when both skewness and kurtosis were considered jointly. Also, a similar comparison among the groups displayed that Other Artisans, Bhatu, Chattri, Kharwar and Korwa had the largest number of deviations.

For assessing the multivariate normality of the groups, we defined the generalized distance of each observation $\mathbf{x}$ from its sample mean vector $\bar{\mathbf{x}}$ as $(\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}})$, where $\mathbf{S}$ is the sample variance-covariance matrix of that group. We then know that these generalized distances all have an approximate chi-square distribution with p degrees of freedom. After ordering, we plotted the observed generalized distances against the expected values of order statistics from the chi-square distribution with p degrees of freedom. These 'probability plots' or 'Q-Q plots' would indicate departures from multivariate normality by departures from linearity and a detailed discussion of these plots has been provided by Everitt (1978) and Gnanadesikan (1977).

The expected values of order statistics from the chi-square distribution with p degrees of freedom, whose cdf is $F(x)$ and pdf $f(x)$ is given by

$$f(x) = \frac{1}{2^{p/2}\Gamma(p/2)} \exp(-x/2) x^{\frac{p}{2}-1}, \quad 0 \leq x < \infty, \tag{2.43}$$

were obtained as given below by using David and Johnson's (1954) approximation; one may refer to David (1981) and Arnold and Balakrishnan (1989) for more details on this method of approximation. Let $X_{i:n}(1 \leq i \leq n)$ denote the ith order statistic in a sample of size n from (2.43). Then the probability integral transformation $u = F(x)$ transforms $X_{i:n}$ into $U_{i:n}(1 \leq i \leq n)$, the ith order statistics from the

Uniform (0,1) distribution. Through a Taylor series expansion, we then obtain

$$
\begin{aligned}
E(X_{i:n}) \approx\ & G_i + \frac{1}{2(n+2)} p_i q_i G_i^{II} + \frac{1}{(n+2)^2} p_i q_i \left\{ \frac{1}{3}(q_i - p_i) G_i^{III} + \frac{1}{8} p_i q_i G_i^{IV} \right\} \\
& + \frac{1}{(n+2)^3} p_i q_i \left[-\frac{1}{3}(q_i - p_i) G_i^{III} + \frac{1}{4} \left\{ (q_i - p_i)^2 - p_i q_i \right\} G_i^{IV} \right. \\
& \left. + \frac{1}{6} p_i q_i (q_i - p_i) G_i^{V} + \frac{1}{48} p_i^2 q_i^2 G_i^{VI} \right. \qquad (2.44)
\end{aligned}
$$

where $p_i = \frac{i}{n+1}, q_i = 1 - p_i, G(u) = F^{-1}(u), G_i = G(p_i), G_i^I = \frac{dG(u)}{du} |_{u=p_i}$ andG_i^{II}, G_i^{III}, and so on, are the successive derivatives of G_i^I with respect to u evaluated at $u = p_i$. It may be easily worked out from (2.43) that

$$
\begin{aligned}
G_i^I &= \frac{1}{f(G_i)} = 2^{\frac{p}{2}} \Gamma\left(\frac{p}{2}\right) \exp(G_i/2) G_i^{1-\frac{p}{2}}, \\
G_i^{II} &= \frac{1}{2}\left[G_i^I\right]^2 + (1 - \frac{p}{2}) \frac{\left[G_i^I\right]^2}{G_i}, \\
G_i^{III} &= G_i^I G_i^{II} + (1 - \frac{p}{2}) \left[2\frac{G_i^I G_i^{II}}{G_i} - \frac{\left[G_i^I\right]^3}{G_i^2} \right], \\
G_i^{IV} &= \left[G_i^{II}\right]^2 + G_i^I G_i^{III} + (1 - \frac{p}{2}) \left[2\frac{\left[G_i^{II}\right]^2}{G_i} - 5\left[G_i^{II}\right]^2 \frac{G_i^{II}}{G_i^2} + 2\frac{\left[G_i^{II}\right]^2}{G_i^3} \right], \\
G_i^{V} &= 3G_i^{II} G_i^{III} + G_i^I G_i^{IV} + (1 - \frac{p}{2}) \left[6\frac{G_i^{II} G_i^{III}}{G_i} - 12\frac{G_i^I \left[G_i^{II}\right]^2}{G_i^2} \right. \\
&\quad \left. -7\left[G_i^I\right]^2 \frac{G_i^{III}}{G_i^2} + 18\left[G_i^I\right]^3 \frac{G_i^{II}}{G_i^3} + 2\frac{G_i^I G_i^{IV}}{G_i} - 6\frac{\left[G_i^I\right]^5}{G_i^4} \right]
\end{aligned}
$$

and

$$
\begin{aligned}
G_i^{VI} &= 3\left[G_i^{III}\right]^2 + 4G_i^{II} G_i^{IV} + G_i^I G_i^V (1 - \frac{p}{2}) \left[6\frac{\left[G_i^{III}\right]^2}{G_i} \right. \\
&\quad + 8\frac{G_i^{II} G_i^{IV}}{G_i} - 44\frac{G_i^I G_i^{II} G_i^{III}}{G_i^2} - 12\frac{\left[G_i^{II}\right]^3}{G_i^2} \\
&\quad + 76\frac{\left[G_i^I G_i^{IV}\right]}{G_i^3} - 9\left[G_i^I\right]^2 \frac{G_i^{IV}}{G_i^2} + 32\frac{\left[G_i^I\right]^3 G_i^{III}}{G_i^3}
\end{aligned}
$$

$$-84\left[G_i^I\right]^4\frac{G_i^{II}}{G_i^4}+2\frac{G_i^I G_i^V}{G_i}+24\frac{\left[G_i^I\right]^6}{G_i^5}\Bigg]. \tag{2.45}$$

Making use of these expressions, we computed the expected values of order statistics from the chi-square distribution with p degrees of freedom from equation (2.44) and plotted them against the ordered observed generalized distances for all the 23 groups. These $Q-Q$ plots of all the 23 groups are presented in pages 72-77. From these plots, we suspect the presence of at least one or two (in some cases, even more than two) outliers in each group. One would like to believe, however, that the presence of a very few outliers in a group of large size may not affect the classical classification procedure. Unfortunately, this does not seem to be the case as we display in the subsequent sections that the performance of the classical procedure is indeed impaired by the presence of these few outliers in the groups. Moreover, we show that the robust classification procedure based on the MML estimators performs better than the classical classification procedure.

2.6.3 Classification using classical procedure

We applied the classical classification procedure in (2.34) to classify each observation in Π_i into either Π_i or Π_j by fixing the error of misclassification e_{ij} as 0.05, for $1 \leq i < j \leq 23$. We then evaluated the estimates of the error of misclassification e_{ij} and $1-e_{ji}$ (the probability of correctly classifying an observation from Π_j into itself) for $1 \leq i < j \leq 23$. These are presented in Table 2.5, where the upper diagonal elements represent the estimates e_{ij} while the lower diagonal elements represent the estimates of $1-e_{ji}$. From this table, we computed the average of e_{ij} $(1 \leq i < j \leq 23)$ values as 0.054 and the average of $1-e_{ji}$ values as 0.578.

A careful examination of this table reveals the following points:

1. There seems to be a considerable overlap between Groups 1 and 2, viz., Basti-Brahmins and Other Brahmins, as the estimate of $1-e_{21}$ is only 0.326 even for an estimated value of e_{12} as high as 0.105.

2. Between Groups 10, 11, 12 and 13 (Ahir, Kurmi, Other and Kahar Artisans),

we once again observe a considerable overlap or a low discrimination. For example, among these four groups we have the average of the estimated e_{ij} values as 0.070 and the average of the estimated $1 - e_{ji}$ value as only 0.273.

3. Upon comparing Groups 14 and 23 (Tharu Males and Females), we observe a good difference between the two groups and, hence, a decent discrimination. For example, we have from Table 2.5 that the estimate of the error of misclassification, $e_{14,23}$, as 0.047 and the estimate of the probablity of correct classification, $1 - e_{23,14}$, to be 0.514.

4. By looking at the values corresponding to Group 22 (Korwa), we observe that there is a considerable difference between the Korwas and all other groups. The remarkable discrimination of the Korwas from all other castes and tribes may be, as pointed out by Majumdar (1944), due to their unwritten law prohibiting any inter-tribal intercourse and inter-tribal marriage and also their highly developed sense of solidarity and comradeship. We also note that the least discrimination of the Korwas occur against Group 19 (Kharwars). This is not surprising as the Kharwars display marked similarities with the Korwas as written earlier and rightly mentioned by Mjumdar (1944).

Finally, based on the first two points made above, we combined the two Brahmin groups (Groups 1 and 2) into one group and, similarly, the four Artisan groups (Groups 10, 11, 12 and 13) were also combined into one group. Then, we once again applied the classical classification procedure to classify each observation in Π_i^* into either Π_i^* or Π_j^* by fixing e_{ij}^* as 0.05, for $1 \leq i < j \leq 19$. As before, we evaluated the estimates of the error of misclassification e_{ij}^* and the probability of the correct classification of an observation from Π_j^* into itself, viz., $1 - e_{ji}^*$, for $1 \leq i < j \leq 19$. These are presented in Table 2.6, where the upper diagonal elements represent the estimates of e_{ij}^* while the lower diagonal elements represent the estimates of $1 - e_{ji}^*$. From this table, we computed the average of e_{ij}^* $(1 \leq i < j \leq 19)$ value as 0.052 and the average of $1 - e_{ji}^*$ values as 0.593 which is only a slight improvement over the previous result.

2.6.4 Classification using robust procedure

We used the robust classification procedure in (2.40) to classify each observation in Π_i into either Π_i or Π_j by fixing e_{ij} as 0.05 for $1 \leq i < j \leq 23$. Then we computed the estimates of e_{ij} and $1 - e_{ji}$ for $1 \leq i < j \leq 23$. These values are presented in Table 2.7, where the upper diagonal elements give the estimates of e_{ij} while the lower diagonal elements give the estimates of $1 - e_{ji}$. From this table, we calculated the average of e_{ij} $(1 \leq i < j \leq 23)$ values as 0.061 and the average of $1 - e_{ji}$ values (estimates of probability of correctly classifying an observation from Π_j into itself) as 0.635. From Table 2.7, we observe the following points:

1. As mentioned in the previous section, we note that there is a considerable overlap between Basti-Brahmins and other Brahmins, as for an estimated value of 0.151 for e_{12} the value of estimate of $1 - e_{21}$ is as low as 0.304.

2. We observe once again a considerable overlap between Ahir, Kurmi, Other and Kahar Artisans. Among these four groups, we have the average of the estimated e_{ij} values as 0.091 and the average of estimated $1 - e_{ji}$ value as 0.285 only.

3. We realize that the robust classification procedure discriminates the Tharu Males and Females remarkably well. From Table 2.7 we find the estimate of $e_{14,23}$ as 0.026 and the estimate of the probability of correct classification, $1 - e_{23,14}$, to be as large as 0.872.

4. As already noted in the previous section, we observe once again that there is a considerable distinction between the Korwas and all other groups. We also see that the least discrimination of the Korwas happens against the Kharwars.

By comparing these results with those of the classical procedure presented in Section 2.6.4, we see immediately that the robust classification procedure has a slightly larger average estimates e_{ij}- values (the difference being only 0.007) and, at the same time, also has a larger average estimated $1 - e_{ji}$ values (the difference being 0.057). This difference may not seem to be significant. However, if we take into

account the number of classifications performed to find the average of the estimated e_{ji} values and of the estimated $1 - e_{ji}$ values, respectively), we simply observe that the robust classification procedure wrongly classifies an individual from Π_i into Π_j ($1 \leq i < j \leq 23$) 217 times more than the classical classification procedure; furthermore, we observe that the robust procedure correctly classifies an individual from Π_j into itself ($2 \leq j \leq 23$) 1912 times more than the classical classification procedure.

Based on the first two points made above, the two Brahmin groups and the four Artisan groups were combined. We then applied the robust classification procedure to classify each observation in Π_i^* into either Π_i^* or Π_j^* by fixing e_{ij}^* as 0.05, for $1 \leq i < j \leq 19$. We evaluated the estimates of e_{ij}^* and $1 - e_{ij}^*$, for $1 \leq i < j \leq 19$. These values are presented in Table 2.8, where as before the upper diagonal elements represent the estimates of e_{ij}^* and the lower diagonal elements represent the estimates of $1 - e_{ji}^*$. From Table 2.8, we computed the average of e_{ij}^* ($1 \leq i < j \leq 19$) values as 0.059 and the average of $1 - e_{ji}^*$ values as 0.658. We note that this is only a slight improvement over the previous performance of the robust classification procedure. However, by comparing these results with those of the classical procedure presented in Section (2.6.4), we see that the robust classification procedure has a slightly larger average of estimated e_{ij}^* values (the difference being only 0.007) and, at the same time, also has a larger average of estimated $1 - e_{ij}^*$ values (the difference being 0.065). Now by taking into account the number of classification performed in order to calculate these values (that is, 27,033 and 25,815 classification performed to find the average of the estimated e_{ij}^* values and of the estimated $1 - e_{ji}^*$ values and of the estimated $1 - e_{ji}^*$ values, respectively), we observe that the robust classification procedure wrongly classifies an individual from Π_i^* into Π_j^* ($1 \leq i < j \leq 19$) 189 times more than the classical procedure; moreover, we observe that the robust procedure correctly classifies an individual from Π_j^* into itself ($2 \leq j \leq 19$) 1678 times more than the classical procedure.

2.6.5 Comparison with Distribution-free Procedures

Let us now consider the classical classification procedure for classifying an individual into either Π_1 or Π_2, without restricting the error of misclassification e_{12} to assume a pre-fixed level; that is, both e_{12} and e_{21} are given equal importance and therefore are allowed to float freely without placing any condition. The procedure for classifying a new observation $\mathbf{x}_0$ is based on the statistic V in (2.26), and is to classify x_0 into Π_1 or Π_2 according as

$$V = \left\{\mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)\right\}^T \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \gtrless 0 \tag{2.46}$$

We applied the above classical linear procedure to classify each observation in Π_i into either Π_i or Π_j for $1 \leq i, j \leq 23, i \neq j$, and then determined estimates of the errors of misclassification e_{ij} and e_{ji} for $1 \leq i < j \leq 23$. These values are presented in Table 2.9, where the (i,j)th entry gives the estimate of e_{ij}, the probability of wrongly classifying an individual from Π_i into Π_j. From this table, we found the average of the estimated e_{ij} values as 0.187 and the average of the estimated e_{ji} values as 0.178. We also observed a poor discrimination between Groups 1 and 2 (Basti-Brahmins and other Brahmins) as the estimates of e_{12} and e_{21} were as large as 0.384 and 0.413, respectively. Similarly, we observed a considerable overlap between the four Artisan groups (Groups 10, 11, 12 and 13) as the averages of the estimated e_{ij} and e_{ji} values were 0.360 and 0.361, respectively. Based on these two points, as done earlier in Section 2.6.3 and 2.6.4, we combined the two Brahmin groups and the four Artisan groups and then applied the linear classification procedure in (2.46) in order to classify each individual in Π_i^* into Π_i^* or Π_j^* for $1 \leq i, j \leq 19, i \neq j$. We then determined estimates of the errors of misclassification e_{ij}^* and e_{ji}^* for $1 \leq i < j \leq 19$, and these are presented in Table 2.10. From this table, we computed the average of the estimated e_{ij}^* values to be 0.177 and the average of the estimated e_{ji}^* values to be 0.168. We observe a slight improvement over the previous results based on all 23 groups.

In a similar way, the robust classification procedure for classifying a new observation $\mathbf{x}_0$ is based on the statistic V_R in (2.36) and is to classify $\mathbf{x}_0$ into Π_1 or

Π_2 according as

$$V_R = \left\{x_0 - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)\right\}^T \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_2) \gtrless 0 \tag{2.47}$$

We applied the above robust linear procedure to classify each observation in Π_i into either Π_i or Π_j for $1 \leq i, j \leq 23, i \neq j$, and then determined estimates of the errors of misclassification e_{ij} and e_{ji} for $1 \leq i < j \leq 23$, and these values are presented in Table 2.11. From this table, we found the average of the estimated e_{ij} values as 0.175 and the average of the estimated e_{ji} values as 0.159. We note that these average error rates are smaller than the corresponding values of the classical classification procedure. This means that the robust procedure in (2.47) has 1,010 fewer misclassifications than the classical procedure in (2.46). As done earlier, we once again combined the two Brahmin groups (Groups 1 and 2) and the four Artisan groups (Groups 10, 11, 12 and 13) and then applied the robust procedure in (2.47) to classify each individual in Π_i^* into either Π_i^* or Π_j^* for $1 \leq i, j \leq 19, i \neq j$. We then determined estimates of the errors of misclassification e_{ij}^* and e_{ji}^* for $1 \leq i < j \leq 19$, and these are presented in Table 2.12. From this table, the average of the estimated e_{ij}^* values to be 0.166 and the average of the estimated e_{ji} values to be 0.149, which is a slight improvement over the previous result. Both these average error rates are once again smaller than the corresponding values of the classical classification procedure. This means that the robust procedure in (2.47) has 787 fewer misclassifications than the classical classification procedure in (2.46).

In order to compare the above two procedures with some nonparametric or distribution-free procedures, we first considered the method based on density estimates which is as follows. Let $\hat{f}_{n_1}(x_0)$ and $\hat{f}_{n_2}(x_0)$ be consistent estimators of the density functions of the populations Π_1 and Π_2 at the point x_0. Then the procedure is simply to classify x_0 into Π_1 or Π_2 according as

$$\frac{\hat{f}_{n_1}(x_0)}{\hat{f}_{n_2}(x_0)} \gtrless 1. \tag{2.48}$$

The density estimator we use in this study is the one due to Loftsgaarden and

Quesenberry (1965) and is given by

$$\hat{f}_m(x_0) = \left\{ \frac{k(m)-1}{m} \right\} \frac{p\Gamma(\frac{p}{2})}{2\pi^{\frac{p}{2}} r_{k(m)}^p}, \tag{2.49}$$

where m is the size of the sample on which the density estimator is based, $r_{k(m)}$ is the distance from $\mathbf{x}_0$ to the $k(m)$'th closest observation in the sample to $\mathbf{x}_0$ as measured by the Euclidean distance function; $k(m)$ was chosen to be nearer to $\sqrt{m}$ and refer to Loftsgaarden and Quesenberry (1965) and Gessaman and Gessaman (1972) for details. It should be mentioned that Tiku and Balakrishnan (1984) have compared the performance of the classification procedure based on the above density estimator in the bivariate case to that based on the bivariate extension of Parzen's density estimator as given by Cacoullos (1964). We applied the above procedure based on (2.48) and (2.49) to classify each observation in Π_i into either Π_i or Π_j for $1 \leq i, j \leq 23$, $i \neq j$, and determined the estimates of the errors of misclassification e_{ij} and e_{ji} for $1 \leq i < j \leq 23$. These values are presented in Table 2.13. From this table, we first of all note that this classification procedure is inconsistent in the sense that some of the error rates are greater than 0.5. We also find the average of the estimated e_{ij} values ($1 \leq i < j \leq 23$) to be 0.246 and the average of the estimated e_{ji} values to be 0.220. It should be noted that both these values are considerably larger than the corresponding values of the robust classification procedure. This means that the robust procedure in (2.47) has 4,251 fewer misclassifications than the distribution-free classification procedure based on Loftsgaarden and Quesenberry's density estimator given above in (2.49).

Next, we consider the nearest neighbour classification procedure proposed by Fix and Hodges (1951) in its multivariate form which is as follows. First, calculate the Euclidean distances d_i ($i = 1, 2, \ldots, n_1 + n_2$), where $d_i = d(\mathbf{x}_0, \mathbf{x}_{1i}), i = 1, 2, \ldots, n_1$ and $d_i = d(\mathbf{x}_0, \mathbf{x}_{2i}), i = n_1 + 1, n_1 + 2, \ldots, n_1 + n_2$. Among the smallest m of these d_i's, let m_1 correspond to observations from the $\mathbf{x}_1$ sample and m_2 correspond to observations from the $\mathbf{x}_2$ sample. Then, Fix and Hodges have proposed to classify the observation x_0 into Π_1 if $m_1 > m_2$, Π_2 if $m_1 < m_2$, and arbitrarily into Π_1 or Π_2 if $m_1 = m_2$; refer to Gessaman and Gessaman (1972) for more details. Unfortunately, the choice of m seems to be very crucial for this procedure. In any

case, we applied this procedure with various choices of m to classify each observation in Π_i into either Π_i or Π_j for $1 \leq i, j \leq 23, i \neq j$, and determined the estimates of the errors of misclassification e_{ij} and e_{ji} for $1 \leq i < j \leq 23$ and are presented in Tables 2.14-2.16. From these tables, we observe that the nearest neighbour classification procedure is also inconsistent as some of the error rates were greater than 0.5. Further, we find the averages of the estimated e_{ij} and e_{ji} values $(1 \leq i < j \leq 23)$ to be 0.324 and 0.267, 0.358 and 0.298, and 0.295 and 0.247, corresponding to the cases $m = 10\%$ of $(n_i + n_j)$, $m = 20\%$ of $(n_i + n_j)$, and $m = \sqrt{n_i + n_j}$, respectively. From these values and also based on comparisons with some other choices of m, we observe that the choice of $m = \sqrt{n_i + n_j}$ yields the smallest error rates. However, all these average error rates are larger than the corresponding values of the procedure based on Loftsgaarden and Quesenberry's density estimator which in turn, as mentioned earlier, are considerably larger than the corresponding average error rates of the robust procedure. Thus, for example, in comparison to the nearest neighbour procedure with $m = \sqrt{n_i + n_j}$, the robust procedure in (2.47) has 6,677 fewer misclassifications. Yet another interesting distribution-free procedure based on 'statistically equivalent blocks' due to Anderson (1966) and Gessaman (1970) has not been considered in this comparative study as it requires unrealistically large sample sizes (of the order of 700) to achieve somewhat higher efficiency (or smaller error rates) than the Loftsgaarden and Quesenberry's procedure; see, for example, Gessaman and Gessaman (1972).

Figure 2.1: Q-Q Plots for groups 1-4

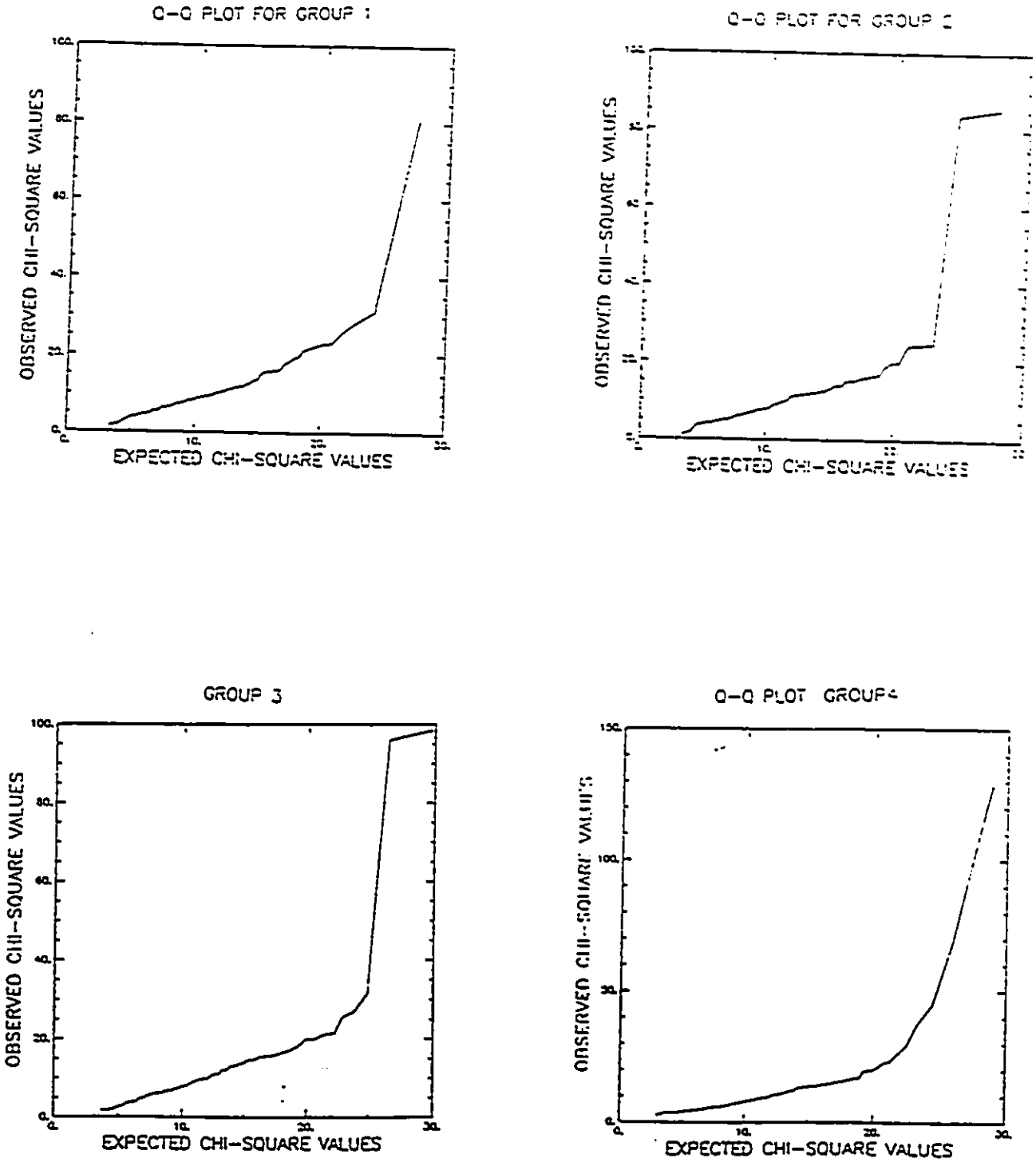


Figure 2.2: Q-Q Plots for groups 5-8

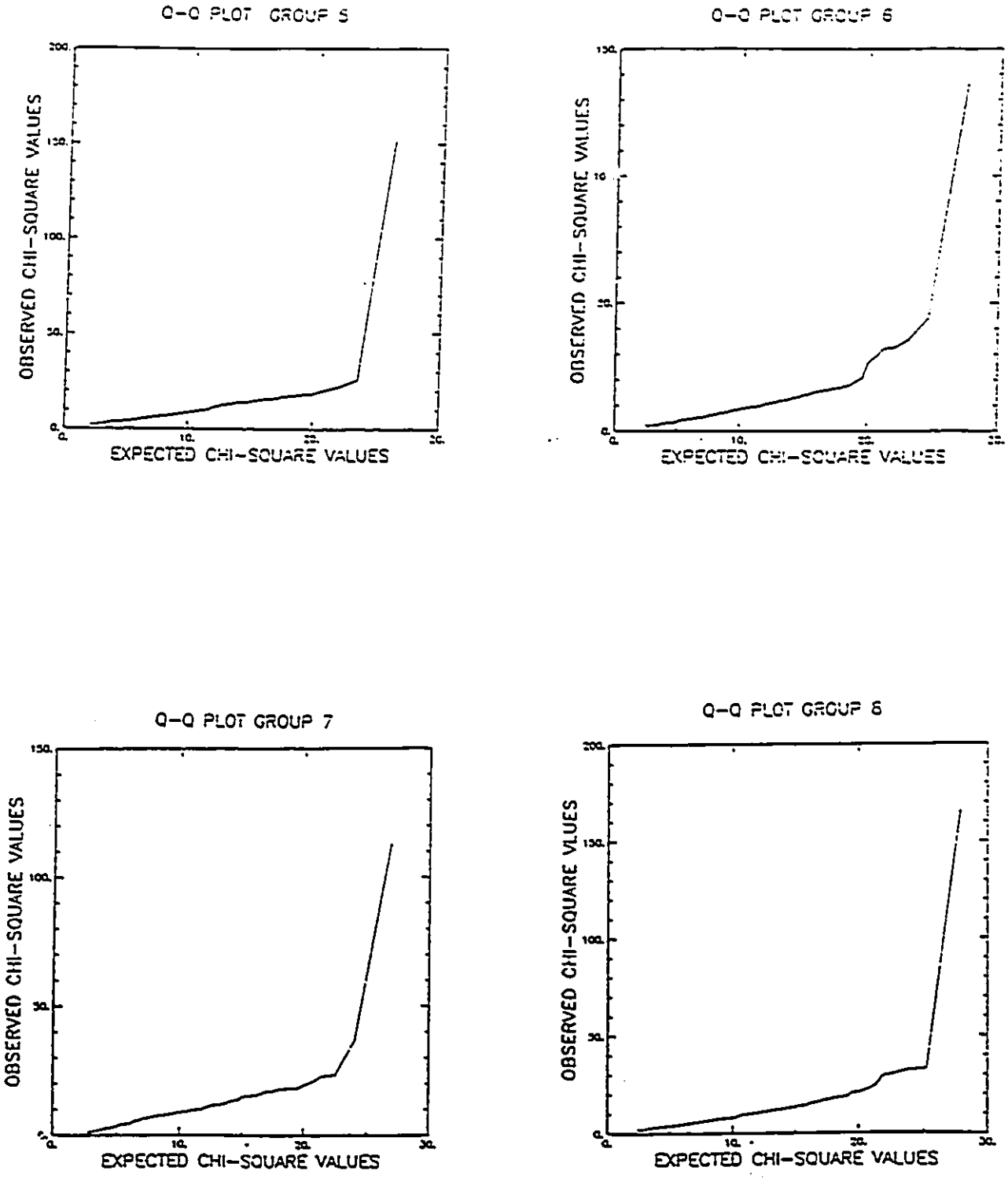


Figure 2.3: Q-Q Plots for groups 9-12

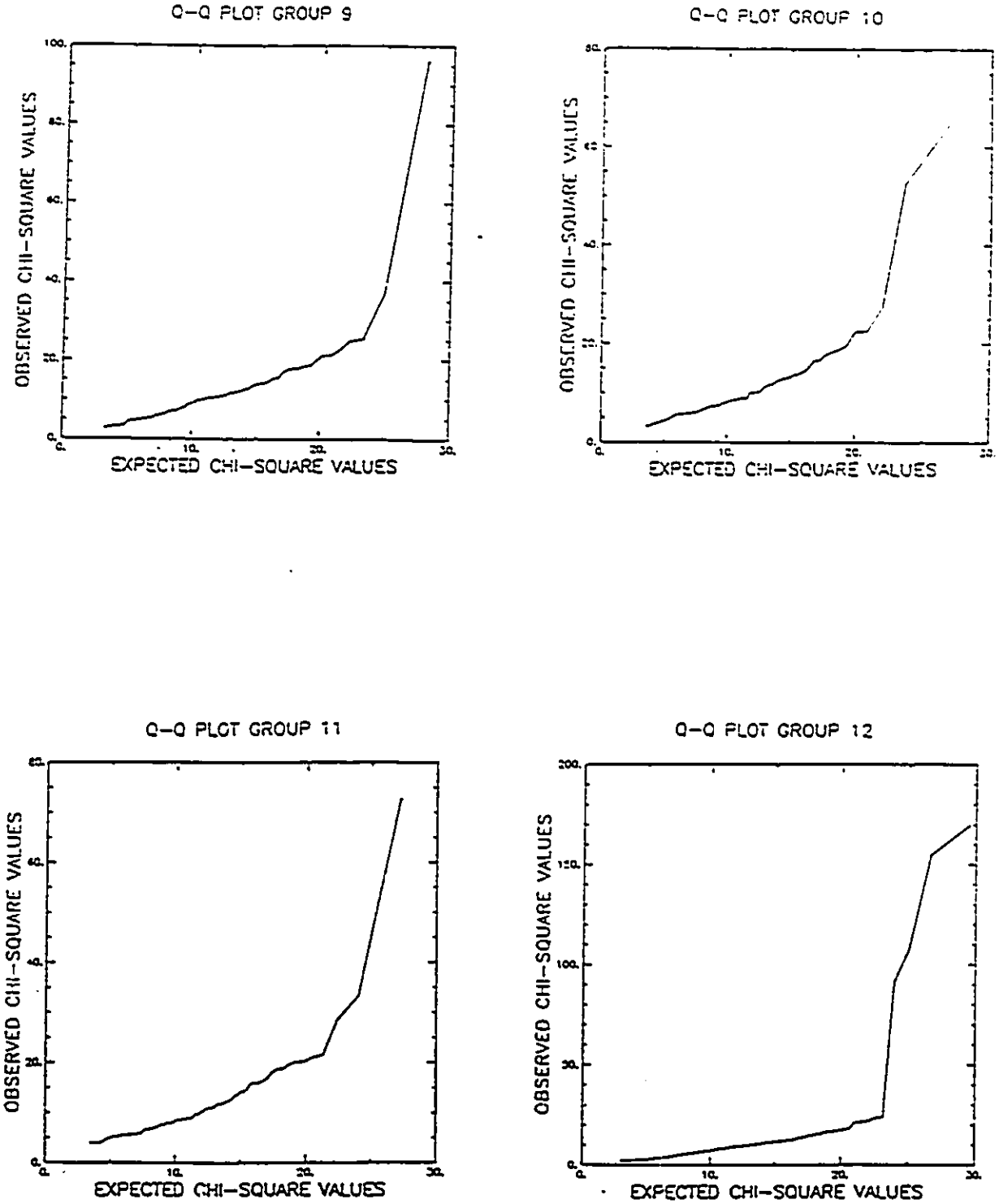


Figure 2.4: Q-Q Plots for groups 13-16

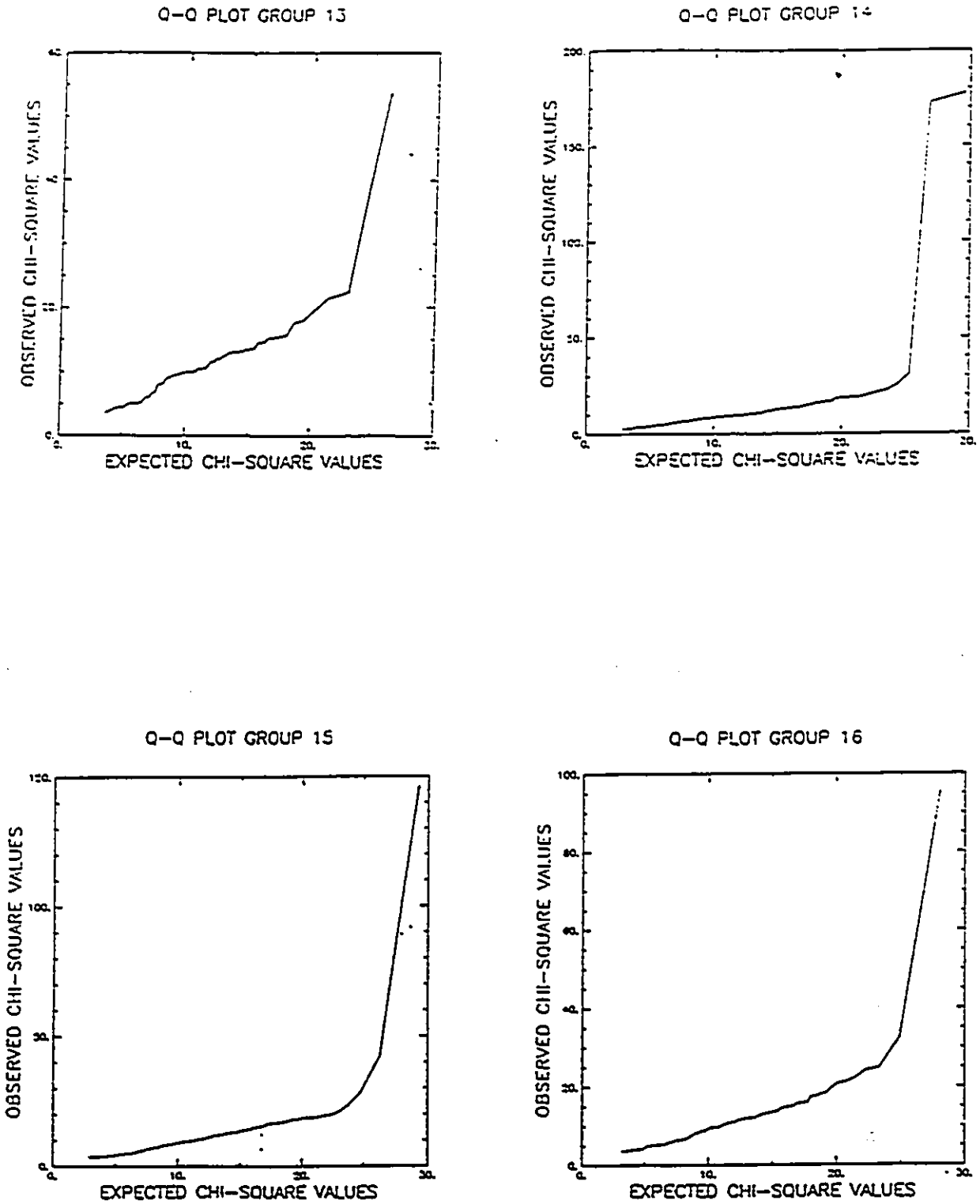


Figure 2.5: Q-Q Plots for groups 17-20

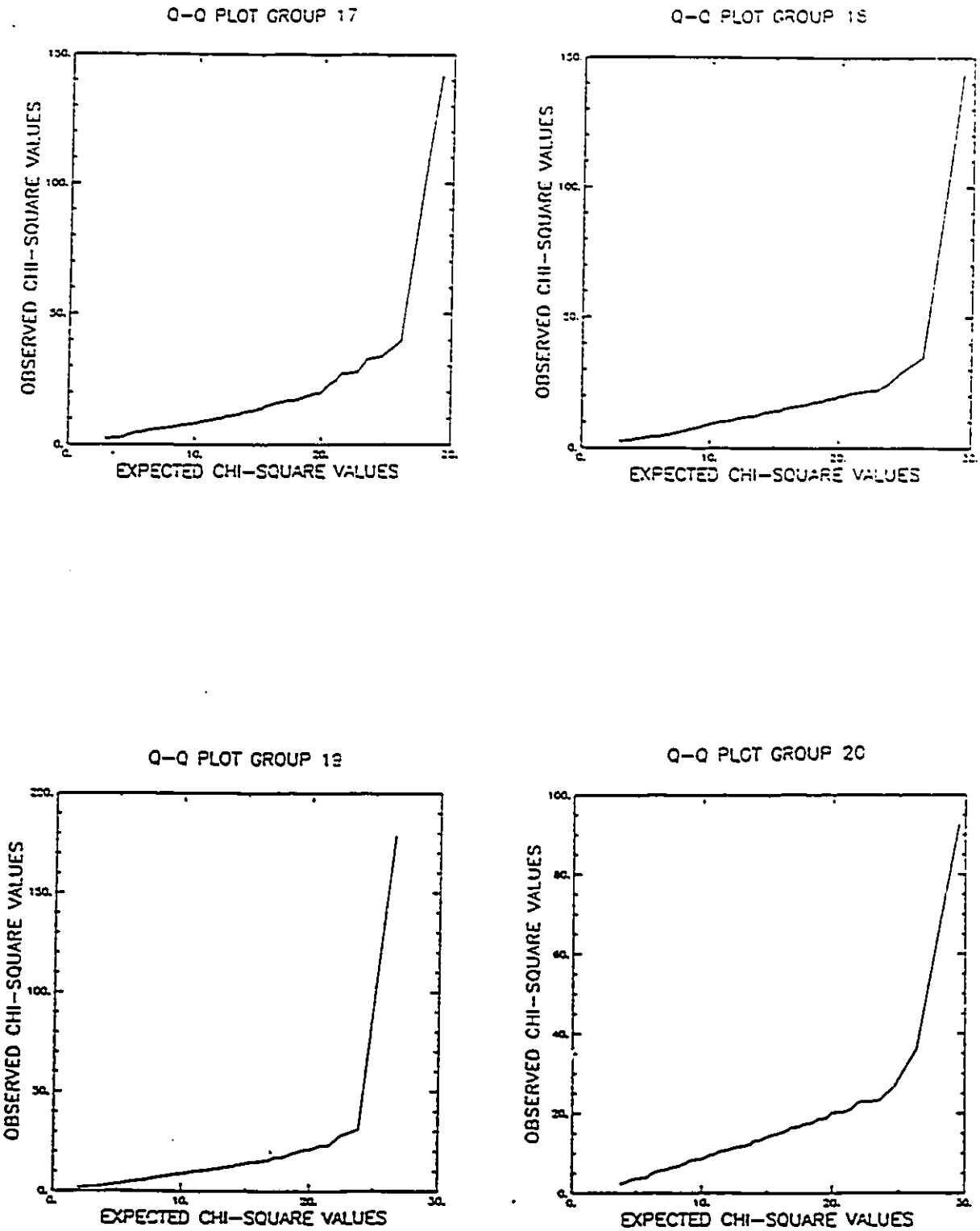
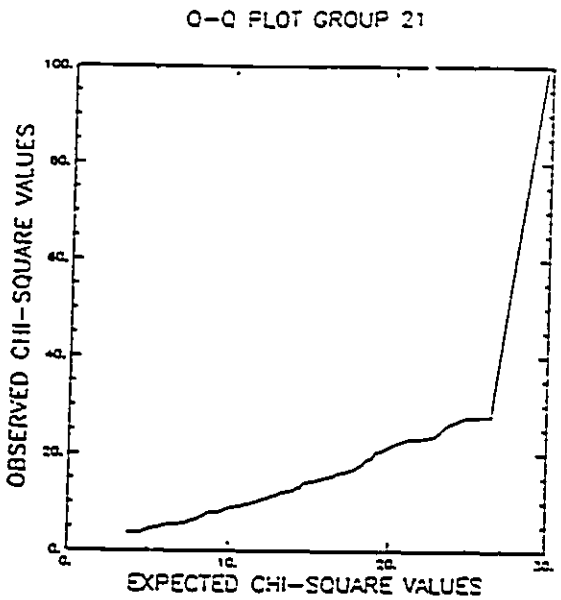
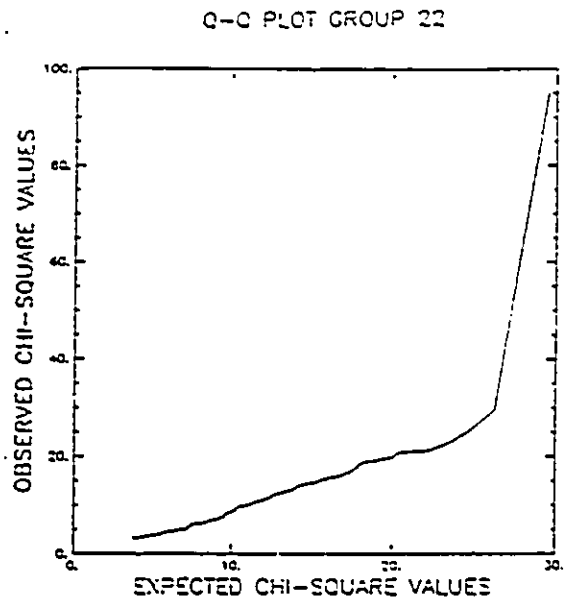


Figure 2.6: Q-Q Plots for groups 21-23





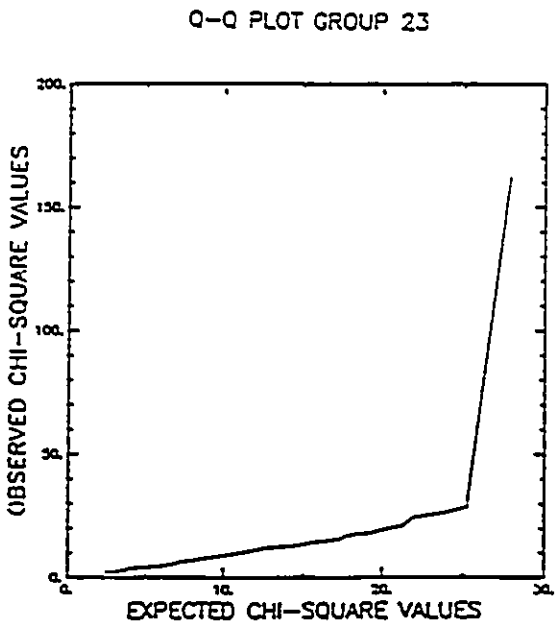


Table 2.5: Values of c_{ij} (fixed as 0.05) and $1 - c_{ji}$, $1 \leq i < j \leq 23$, for the classical procedure

(22)	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
(21)	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
(20)	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
(19)	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
(18)	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
(17)	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
(16)	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
(15)	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
(14)	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
(13)	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
(12)	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
(11)	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
(10)	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
(9)	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
(8)	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
(7)	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
(6)	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
(5)	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
(4)	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
(3)	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
(2)	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
(1)	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
	(22)	(21)	(20)	(19)	(18)	(17)	(16)	(15)	(14)	(13)	(12)	(11)	(10)	(9)	(8)	(7)	(6)	(5)	(4)	(3)	(2)	(1)		

Table 2.6: Values of ϵ^*_{ij} (fixed as 0.05) and $1 - \epsilon^*_{ji}$, $1 \leq i < j \leq 19$, for the classical procedure

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1)	****	.023	.023	.045	.034	.045	.051	.056	.045	.056	.011	.023	.040	.028	.051	.056	.040	.045	.062
(2)	.383	****	.076	.047	.028	.076	.047	.056	.037	.019	.047	.028	.047	.019	.038	.037	.037	.019	.037
(3)	.338	.454	****	.160	.060	.064	.064	.039	.093	.060	.036	.039	.038	.039	.060	.043	.036	.036	.036
(4)	.371	.434	.367	****	.042	.042	.054	.048	.066	.078	.078	.048	.072	.066	.090	.078	.060	.078	.078
(5)	.493	.490	.322	.315	****	.034	.074	.047	.080	.074	.047	.067	.074	.074	.080	.080	.047	.080	.064
(6)	.312	.236	.344	.312	.246	****	.041	.056	.062	.049	.074	.041	.049	.062	.066	.066	.066	.041	.066
(7)	.436	.447	.507	.619	.661	.463	****	.033	.038	.044	.033	.044	.050	.061	.060	.039	.061	.050	.066
(8)	.616	.587	.709	.587	.563	.663	.398	****	.038	.078	.038	.067	.097	.078	.067	.088	.049	.097	.078
(9)	.170	.368	.340	.199	.423	.313	.181	.482	****	.060	.067	.070	.055	.067	.049	.038	.036	.043	.040
(10)	.821	.643	.818	.826	.711	.674	.763	.858	.679	****	.058	.058	.042	.058	.037	.063	.037	.006	.047
(11)	.384	.263	.591	.333	.566	.377	.390	.535	.434	.711	****	.031	.038	.036	.044	.044	.057	.057	.066
(12)	.703	.624	.822	.646	.693	.605	.616	.634	.664	.752	.456	****	.058	.099	.089	.040	.089	.030	.049
(13)	.851	.652	.877	.669	.890	.618	.578	.792	.714	.812	.546	.247	****	.056	.026	.046	.046	.046	.032
(14)	.692	.603	.776	.481	.760	.647	.539	.782	.615	.776	.340	.224	.224	****	.077	.046	.064	.026	.030
(15)	.641	.564	.800	.616	.769	.686	.492	.733	.369	.262	.306	.267	.231	.210	****	.046	.056	.046	.046
(16)	.836	.808	.879	.737	.679	.748	.818	.939	.748	.394	.626	.666	.676	.476	.304	****	.071	.051	.061
(17)	.712	.686	.746	.651	.728	.623	.793	.896	.666	.628	.666	.717	.736	.860	.406	.396	****	.047	.067
(18)	.890	.860	.810	.780	.910	.820	.820	.880	.590	.430	.730	.680	.630	.480	.350	.480	.620	****	.080
(19)	.936	.827	.938	.860	.961	.911	.944	.983	.821	.614	.804	.749	.766	.642	.659	.352	.436	.692	****

Table 2.7: Values of c_{ij} (fixed as 0.05) and $1 - c_{ji}$, $1 \leq i < j \leq 23$, for the robust procedure

	(23)	(22)	(21)	(20)	(19)	(18)	(17)	(16)	(15)	(14)	(13)	(12)	(11)	(10)	(9)	(8)	(7)	(6)	(5)	(4)	(3)	(2)	(1)
(22)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(21)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(20)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(19)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(18)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(17)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(16)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(15)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(14)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(13)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(12)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(11)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(10)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(9)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(8)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(7)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(6)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(5)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(4)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(3)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(2)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(1)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

Table 2.8: Values of c^*_{ij} (fixed as 0.05) and $1 - c^*_{ji}$, $1 \leq i < j \leq 19$, for the robust procedure

Table 2.9: Values of e_{ij} and $1 - e_{ji}$, $1 \leq i < j \leq 23$, for the classical procedure

(22)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(21)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(20)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(19)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(18)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(17)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(16)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(15)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(14)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(13)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(12)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(11)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(10)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(9)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(8)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(7)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(6)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(5)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(4)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(3)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(2)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(1)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	

Table 2.10: Values of c^*_{ij} and $1 - c^*_{ji}$, $1 \leq i < j \leq 19$, for the classical procedure

Table 2.11: Values of c_{ij} and $1 - c_{ij}$, $1 \leq i < j \leq 23$, for the robust procedure

.000	.000	.021	.000	.020	.001	.020	.000	.010	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	(22)
.000	.000	.001	.001	.000	.022	.000	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	(21)
.101	.021	.000	.001	.000	.000	.000	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	(20)
.001	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	(19)
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	(18)
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	(17)
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	(16)
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	(15)
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	(14)
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	(13)
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	(12)
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	(11)
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	(10)
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	(9)
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	(8)
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	(7)
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	(6)
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	(5)
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	(4)
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	(3)
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	(2)
(22)	(21)	(20)	(19)	(18)	(17)	(16)	(15)	(14)	(13)	(12)	(11)	(10)	(9)	(8)	(7)	(6)	(5)	(4)	(3)	(2)	(1)			

Table 2.12: Values of c^*_{ij} and $1 - c^*_{ji}$, $1 \leq i < j \leq 19$, for the robust procedure

(1,1)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(2,1)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(3,1)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(4,1)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(5,1)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(6,1)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(7,1)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(8,1)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(9,1)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(10,1)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(11,1)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(12,1)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(13,1)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(14,1)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(15,1)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(16,1)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(17,1)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(18,1)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
(19,1)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)	(7,1)	(8,1)	(9,1)	(10,1)	(11,1)	(12,1)	(13,1)	(14,1)	(15,1)	(16,1)	(17,1)	(18,1)	(19,1)	

Table 2.13: Values of c_{ij} and $1 - c_{ji}$, $1 \leq i < j \leq 23$, for the Loftsgaarden-Quesenberry procedure

Table 2.14: Values of e_{ij} and $1 - c_{ji}$, $1 \leq i < j \leq 23$, for the nearest-neighbour procedure with $R = 10\%$ of $(n_i + n_j)$

Table 2.15: Values of e_{ij} and $1 - e_{ji}$, $1 \leq i < j \leq 23$, for the nearest-neighbour procedure with $R = 20\%$ of $(n_i + n_j)$

Table 2.16: Values of c_{ij} and $1 - c_{ji}$, $1 \leq i < j \leq 23$, for the nearest-neighbour procedure with $R = (n_1 + n_2)^{1/2}$

Chapter 3

Univariate Linear Classification Procedure Based On Dichotomous and Continuous Variables

3.1 Introduction

Consider the following situations:

1. In a clinical trial one might measure depression in blood pressure and presence or absence of various symptoms of two different groups (or races). Later one might want to classify a new blood sample to one of the two groups.

2. Consider sample of patients who had attempted suicide by ingesting overloads of barbiturates. The drug comes in two types: short acting (0) and long acting (1). Also patients' systolic blood pressure and stroke work whose clinical value plays a major role in predicting survival were recorded.For purposes of prognosis, the patients are to be classified as survivors or non survivors. (This example is from Chang and Afifi (1974)).

Above two examples illustrate that it is not uncommon to find some of the variables to be continuous while others to be discrete in classification problems. In the past most popular approaches to such problems seem to have been to treat

the discrete variables as continuous and to apply a straightforward linear decision surface- usually Fisher's method- or to group the continuous variables into categories. (see, for example, Cochran and Hopkins, 1961). Such approaches have obvious associated disadvantages.

Chang and Afifi (1974) introduced a classification procedure based on dichotomous and normal variables. They have shown that this procedure performs better than treating original data set as bivariate continuous data set or bivariate discrete data set. Krzanowski (1975) considered a classification procedure based on k binary variables and p-variate normal distribution and has shown that his procedure performs better than using the classical classification procedure after dichotomization of the continuous variables of original data or considering binary variables as if they were continuous variables. Both Chang and Afifi (1974) and Krzanowski (1975) have concentrated on the marginal distribution of the dichotomous components and the conditional distribution of the continuous components given the binary components, taking the continuous components collectively one at a time.

Balakrishnan and Tiku (1988b) have shown that Chang and Afifi's (1974) procedure is not robust to departures from normality. They have derived a procedure based on MML estimators and have shown that their procedure performs better, better in the sense that it has a larger probability of correctly classifying an observation from second population into itself when the error of misclassification of an observation from first population into second population is fixed. Balakrishnan and Tiku (1988b) have carried out a simulation study through which they have displayed the overall superiority of their procedure based on MML estimators when one of the errors of misclassification is at a prefixed level.

In this chapter, we consider the classical procedure of Chang and Afifi (1974) and the robust procedure of Balakrishnan and Tiku (1988b) when neither of the two errors of misclassification is at a prefixed level (that is, when both are allowed to float), and study the errors of misclassification. We derive in this case the asymptotic formulas for the errors of misclassification e_{12} and e_{21} for both procedures and show why the classical procedure is not robust to departures from normality

while the robust procedure is. Through Monte Carlo simulations, we show that the robust procedure based on MML estimators has a smaller average error rate than the classical procedure for a wide range of symmetric and skewed non-normal models.

For this purpose, we first present in Section 2 the basic model of the classification problem considered here and describe briefly the classical linear and the robust linear procedures as developed by Chang and Afifi (1974) and Balakrishnan and Tiku (1988b), respectively. We also discuss briefly the asymptotic determination of the cut-off points of these two procedures. This discussion of the past work on this problem is included in Section 2 for the following reasons:

1. For the sake of completeness of work on this problem,
2. For an understanding of the asymptotic arguments applied in studying the errors of misclassification in the rest of the chapter, and
3. For forming a basis for the generalizations of this problem considered in this and subsequent chapters.

In Section 3, we describe the two procedures when both the errors of misclassification are allowed to float (that is, neither is at a fixed level) and then derive some asymptotic formulas for the errors of misclassification. In this section, we also carry out an extensive Monte Carlo study through which we show that the robust procedure has a smaller average error rate than the classical procedure. The results presented in this chapter in their entirety have appeared recently in a paper by Tiku, Balakrishnan and Ambagaspitiya (1989).

3.2 Classification procedures when one of the errors is fixed

In this section, we first introduce the basic model and then describe the classical and the robust linear classification procedures when one of the errors of misclassification is at a pre- specified level.

3.2.1 The model

Let X be a Bernoulli variate with probability mass function

$$\Pr(X = x) = \theta^x(1-\theta)^{1-x}, \quad x = 0, 1,$$

and Y be a continuous variable such that the conditional distribution of Y, given $X = x$, is normal with mean $\mu(x) = \mu + x\delta$ and variance $\sigma^2(x) = \sigma^2 + x\gamma^2$. Let $\begin{pmatrix} x_{1i} \\ y_{1i} \end{pmatrix}$, $i = 1, 2, \ldots, n_1$, be a random sample of size n_1 from population Π_1 which is the model specified above with θ_1 and $\mu_1(x)$. Similarly, let $\begin{pmatrix} x_{2i} \\ y_{2i} \end{pmatrix}$, $i = 1, 2, \ldots, n_2$, be a random sample of size n_2 from population Π_2 which is the model specified above with θ_2 and $\mu_2(x)$. Let $\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ be an independent new observation that needs to be classified into either Π_1 or Π_2.

3.2.2 Classical Linear Classification Procedure

The classical estimates of the parameters are given by

$$\hat{\theta}_i = \frac{n_i(1)}{n_i} \quad (i = 1, 2),$$

where $n_i(1)$ is the number of observations in the ith sample that correspond to $x = 1$, and

$$\hat{\mu}_i(x) = \bar{y}_i(x)$$

and

$$\hat{\sigma}^2(x) = s^2(x) = \frac{(n_1(x)-1)s_1^2(x) + (n_2(x)-1)s_2^2(x)}{n_1(x) + n_2(x) - 2},$$

for $x = 0$ and 1, and $i = 1$ and 2. Then, as shown by Chang and Afifi (1974), Krzanowski (1975) and Balakrishnan and Tiku (1988b), the classical linear classification procedure is to classify $\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ into Π_1 or Π_2 according to

$$V(x_0) \;=\; \left\{y_0 - \frac{1}{2}(\bar{y}_1(x_0) + \bar{y}_2(x_0))\right\} \frac{(\bar{y}_1(x_0) - \bar{y}_2(x_0))}{s^2(x_0)}$$

$$+x_0 \ln\left[\frac{\hat{\theta}_1}{\hat{\theta}_2}\right] + (1-x_0)\ln\left[\frac{1-\hat{\theta}_1}{1-\hat{\theta}_2}\right] \gtrless C, \tag{3.1}$$

where the cut-off point C is determined so that the error of misclassification e_{12} (for example) is at a pre-specified level. By using asymptotic arguments, Balakrishnan and Tiku (1988b) have shown that (with e_{12} fixed as α) the cut-off point C in (3.1) is approximately the solution of the equation

$$(1-\theta_1)\Phi\left\{\frac{C-\frac{1}{2}\eta^2(0)-\ln\left(\frac{1-\theta_1}{1-\theta_2}\right)}{|\eta(0)|}\right\} + \theta_1\Phi\left\{\frac{C-\frac{1}{2}\eta^2(1)-\ln\left(\frac{\theta_1}{\theta_2}\right)}{|\eta(1)|}\right\} = \alpha, \tag{3.2}$$

where $\Phi()$ is the cumulative distribution function of a standard normal variable, and $\eta^2(x) = \frac{(\mu_1(x)-\mu_2(x))^2}{\sigma^2(x)}$ for $x = 0, 1$. Since all the parameters in Eq. (3.2) are unknown, the cut-off point C can not be determined from (3.2). However, by using the classical estimates of all the parameters, we may obtain an estimate of the cut-off point C as the solution $\hat{C}$ of the equation

$$(1-\hat{\theta}_1)\Phi\left\{\frac{\hat{C}-\frac{1}{2}\hat{\eta}^2(0)-\ln\left(\frac{1-\hat{\theta}_1}{1-\hat{\theta}_2}\right)}{|\hat{\eta}(0)|}\right\} + \hat{\theta}_1\Phi\left\{\frac{\hat{C}-\frac{1}{2}\hat{\eta}^2(1)-\ln\left(\frac{\hat{\theta}_1}{\hat{\theta}_2}\right)}{|\hat{\eta}(1)|}\right\} = \alpha, \tag{3.3}$$

where $\hat{\eta}^2(x) = \frac{(\bar{y}_1(x)-\bar{y}_2(x))^2}{s^2(x)}$, $x = 0, 1$.

3.2.3 Robust Linear Classification Procedure

Let $\hat{\theta}_i = \frac{n_i(1)}{n_i}$, $i = 1, 2$, where $n_i(1)$ is the number of observations in the ith sample that correspond to $x = 1$, be the estimate of θ_i as before. Further, let $(\hat{\mu}_1(x), \hat{\sigma}_1^2(x))$ and $(\hat{\mu}_2(x), \hat{\sigma}_2^2(x))$ be the MML estimators of $(\mu_1(x), \sigma_1^2(x))$ and $(\mu_2(x), \sigma_2^2(x))$ obtained from the two samples, respectively. Let $\hat{\sigma}^2(x)$ be the pooled MML estimator of $\sigma^2(x)$ defined by

$$\hat{\sigma}^2(x) = \frac{(A_1(x)-1)\hat{\sigma}_1^2(x) + (A_2(x)-1)\hat{\sigma}_2^2(x)}{A_1(x)+A_2(x)-2}$$

for $x = 0$ and 1. Then, the robust classification procedure as given by Balakrishnan and Tiku (1988b) is to classify the new independent observation $\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ into Π_1 or

Π_2 according as

$$\begin{aligned} V_R(x_0) &= \left\{y_0 - \frac{1}{2}(\hat{\mu}_1(x_0) + \hat{\mu}_2(x_0))\right\} \frac{(\hat{\mu}_1(x_0) - \hat{\mu}_2(x_0))}{\hat{\sigma}^2(x_0)} \\ &\quad + x_0 \ln\left[\frac{\hat{\theta}_1}{\hat{\theta}_2}\right] + (1 - x_0)\ln\left[\frac{1-\hat{\theta}_1}{1-\hat{\theta}_2}\right] \gtrless C_R, \end{aligned} \tag{3.4}$$

where the cut-off point C_R is determined so that the error of misclassification e_{12} is at a pre-fixed level. By using asymptotic arguments, Balakrishnan and Tiku (1988b) have shown that (with e_{12} fixed as α) the cut-off point C_R in (3.4) is approximately the solution of the equation

$$(1-\theta_1)\Phi\left\{\frac{C_R - \frac{1}{2}\eta^2(0) - \ln\left(\frac{1-\theta_1}{1-\theta_2}\right)}{|\eta(0)|}\right\} + \theta_1\Phi\left\{\frac{C_R - \frac{1}{2}\eta^2(1) - \ln\left(\frac{\theta_1}{\theta_2}\right)}{|\eta(1)|}\right\} = \alpha. \tag{3.5}$$

By using the MML estimates of all the parameters calculated from the two samples, we may obtain an estimate of the cut-off point C_R as the solution $\hat{C}_R$ of the equation

$$(1-\hat{\theta}_1)\Phi\left\{\frac{\hat{C}_R - \frac{1}{2}\hat{\eta}_R^2(0) - \ln\left(\frac{1-\hat{\theta}_1}{1-\hat{\theta}_2}\right)}{|\hat{\eta}_R(0)|}\right\} + \hat{\theta}_1\Phi\left\{\frac{\hat{C}_R - \frac{1}{2}\hat{\eta}_R^2(1) - \ln\left(\frac{\hat{\theta}_1}{\hat{\theta}_2}\right)}{|\hat{\eta}_R(1)|}\right\} = \alpha, \tag{3.6}$$

where $\hat{\eta}_R^2(x) = \frac{(\hat{\mu}_1(x)-\hat{\mu}_2(x))^2}{\hat{\sigma}^2(x)}, \quad x = 0, 1.$

Balakrishnan and Tiku (1988) have proved the following two theorems with regard to the distributions of $V(x_0)$ and $V_R(x_0)$.

Theorem 7 *Under the assumption of normality, $V(x)$ and $V_R(x)$ have identical distributions for large n_1 and n_2.*

Theorem 8 *Under the assumption of normality, $V(x)$ and $V_R(x)$ procedures have exactly the same e_{12} and e_{21} values for large n_1 and n_2.*

3.2.4 Illustrative Example

In this section we present a simulated data set in order to illustrate the classical and the robust classification procedures described above.

Table 3.1: Measurements on upper facial length of n_1=100 individuals from Group 1 along with their gender

i	x_{1i}	y_{1i}	i	x_{1i}	y_{1i}	i	x_{1i}	y_{1i}	i	x_{1i}	y_{1i}	i	x_{1i}	y_{1i}
1	0	44.5	21	1	34.1	41	1	34.8	61	0	40.5	81	0	40.8
2	0	23.0	22	1	44.0	42	1	38.5	62	1	43.6	82	0	40.5
3	0	37.6	23	1	37.6	43	0	36.4	63	0	39.0	83	0	28.6
4	0	45.1	24	1	37.7	44	0	41.2	64	0	38.6	84	1	44.0
5	0	42.1	25	0	36.7	45	0	47.8	65	0	34.3	85	0	35.0
6	0	39.2	26	0	40.5	46	0	48.4	66	1	46.3	86	0	46.8
7	1	42.8	27	1	36.4	47	0	40.9	67	1	38.0	87	1	51.2
8	0	43.7	28	1	41.7	48	0	38.6	68	0	43.4	88	0	29.4
9	1	37.4	29	1	41.8	49	0	41.1	69	1	49.9	89	0	39.7
10	1	49.9	30	0	39.2	50	0	43.2	70	0	39.1	90	1	42.7
11	1	33.9	31	0	30.4	51	0	40.8	71	0	41.0	91	0	43.2
12	0	35.4	32	0	38.5	52	1	48.2	72	1	54.5	92	1	44.3
13	0	45.2	33	1	40.4	53	0	35.8	73	0	37.6	93	0	41.5
14	0	40.2	34	0	41.5	54	1	50.5	74	0	45.8	94	0	49.1
15	1	39.2	35	0	36.3	55	1	41.7	75	0	35.3	95	0	39.4
16	1	44.1	36	0	43.7	56	0	35.9	76	1	42.2	96	1	41.9
17	0	44.3	37	0	38.3	57	0	35.5	77	0	39.7	97	1	42.3
18	0	44.7	38	0	40.6	58	1	37.7	78	1	42.2	98	1	43.5
19	1	39.6	39	0	42.9	59	1	45.5	79	1	-49.8	99	0	41.9
20	1	44.0	40	0	40.1	60	0	38.1	80	0	42.0	100	0	40.7

Example

In Tables 3.1 and 3.2, we have given a simulated data set in an anthropometric set-up with x denoting the gender (0 for female, 1 for male) and y the upper facial length for two different groups of people.

Table 3.2: Measurements on upper facial length of n_2=80 individuals from Group 2 along with their gender

i	x_{2i}	y_{2i}	i	x_{2i}	y_{2i}	i	x_{2i}	y_{2i}	i	x_{2i}	y_{2i}	i	x_{2i}	y_{2i}
1	0	52.8	17	0	38.1	33	1	39.9	49	1	41.6	65	1	35.7
2	0	36.1	18	1	32.2	34	0	31.2	50	1	44.3	66	0	40.6
3	0	24.6	19	0	26.5	35	1	50.0	51	0	29.6	67	0	31.4
4	0	40.7	20	0	35.2	36	0	38.7	52	1	46.4	68	1	47.4
5	0	37.6	21	1	31.8	37	0	34.6	53	1	44.0	69	1	39.6
6	1	35.8	22	1	38.4	38	1	36.5	54	1	41.2	70	1	37.6
7	0	41.5	23	0	33.5	39	0	37.7	55	0	40.5	71	1	38.1
8	0	44.6	24	0	37.9	40	0	37.0	56	0	34.6	72	0	32.6
9	0	36.2	25	0	35.7	41	1	41.0	57	1	36.4	73	1	39.7
10	1	32.2	26	0	38.3	42	1	37.6	58	1	32.9	74	0	39.9
11	1	30.2	27	0	30.8	43	0	36.7	59	1	34.4	75	1	47.3
12	0	33.5	28	1	32.0	44	1	42.7	60	1	38.7	76	0	42.8
13	0	40.7	29	1	48.4	45	1	49.0	61	1	45.0	77	1	36.9
14	1	41.0	30	1	38.7	46	0	29.4	62	0	35.4	78	1	42.6
15	1	37.4	31	0	33.0	47	0	32.0	63	0	40.8	79	1	37.6
16	1	39.6	32	1	43.8	48	0	37.8	64	0	34.6	80	1	43.6

From these two samples, we find:

$$\begin{array}{llll}
\bar{y}_1(0) = 39.8446, & \bar{y}_2(0) = 36.2827, & s_1^2(0) = 21.9430, & s_2^2(0) = 26.7565, \\
\hat{\mu}_1(0) = 40.1439, & \hat{\mu}_2(0) = 36.1609, & \hat{\sigma}_1^2(0) = 14.8989, & \hat{\sigma}_2^2(0) = 18.6869, \\
\bar{y}_1(1) = 42.5738, & \bar{y}_2(1) = 39.7397, & s_1^2(1) = 24.7587, & s_2^2(1) = 26.0199, \\
\hat{\mu}_1(1) = 42.5738, & \hat{\mu}_2(1) = 39.7397, & \hat{\sigma}_1^2(1) = 26.7490, & \hat{\sigma}_2^2(1) = 33.1631, \\
\hat{\theta}_1 = 0.3800, & \hat{\theta}_2 = 0.5125, & \hat{C} = -0.9024, & \hat{C}_R = -0.9796.
\end{array}$$

So, with e_{12} (the probability of wrongly classifying an individual from Group 1 into Group 2) fixed as 0.05, by applying the classical classification procedure in (3.1) and reclassifying each individual in Tables 3.1 and 3.2, we obtain estimates of e_{12} and $1 - e_{21}$ (the probability of correctly classifying an individual from Group 2 into itself) to be 0.060 and 0.150, respectively. Similarly, by applying the robust linear classification procedure in (3.4) and reclassifying each individual in Tables 3.1 and 3.2, we obtain estimates of e_{12} and $1 - e_{21}$ to be 0.060 and 0.200, respectively.

3.3 Classification procedures when both errors are allowed to float

In some instances, one may not be interested in pre-fixing the error of misclassification e_{12} and instead be interested in allowing both the errors of misclassification to float. This will be the case when committing one type of error is not considered to be any more serious than committing the other type of error. In this case, the classical linear classification procedure is to classify the independent observation $\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ into Π_1, Π_2 or arbitrarily into Π_1 or Π_2, according as

$$\begin{aligned}
V^*(x_0) &= \left\{y_0 - \frac{1}{2}(\bar{y}_1(x_0) + \bar{y}_2(x_0))\right\}(\bar{y}_1(x_0) - \bar{y}_2(x_0)) \\
&\quad + x_0 \ln\left[\frac{\hat{\theta}_1}{\hat{\theta}_2}\right] + (1 - x_0)\ln\left[\frac{1-\hat{\theta}_1}{1-\hat{\theta}_2}\right] \gtrless = 0,
\end{aligned} \tag{3.7}$$

and the corresponding robust linear classification procedure is to classify the observation $\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ into Π_1, Π_2 or arbitrarily into Π_1 or Π_2, according as

$$\begin{aligned} V^*(x_0) &= \left\{ y_0 - \frac{1}{2}(\hat{\mu}_1(x_0) + \hat{\mu}_2(x_0)) \right\} (\hat{\mu}_1(x_0) - \hat{\mu}_2(x_0)) \\ &\quad + x_0 \ln \left[\frac{\hat{\theta}_1}{\hat{\theta}_2} \right] + (1 - x_0) \ln \left[\frac{1 - \hat{\theta}_1}{1 - \hat{\theta}_2} \right] \mu_2(x_0)) \gtrless = 0. \end{aligned} \tag{3.8}$$

3.3.1 Results on errors of misclassification

In this section first we present expressions for the asymptotic values of e_{12} and e_{21}, i.e. for large sample sizes. Since it is very difficult to obtain exact analytical expressions for e_{12} and e_{21} we present results obtained by Monte Carlo methods by considering normal as well as various non-normal models.

3.3.2 Asymptotic values of e_{12} and e_{21}

As n_1 and n_2 become large ,$\frac{n_i(1)}{n_i}$, $\bar{y}_i(x)$ and $s^2(x)$ converge to their expected values θ_i, $\mu_i(x)$ and $\sigma^2(x)$, respectively. In that case, $V(x)$ is a linear function of y_0. The asymptotic distribution of $V(x)$ given $X = x$ is, therefore, the same as the distribution of y_0, other than its mean and variance. If we define

$$V_1 = \frac{V^*(x) - E(V^*(x))}{\sqrt{Var(V^*(x))}} \tag{3.9}$$

so that $E(V_1) = 0$ and $Var(V_1) = 1$, the asymptotic values of e_{12} and e_{21} associated with $V(x)$ are then given by

$$\begin{aligned} e_{12} &= (1 - \theta_1) \Pr[V^*(0) < 0 | w_0 \in \Pi_1] + \theta_1 \Pr[V^*(1) < 0 | w_0 \in \Pi_1] \\ &= (1 - \theta_1) \Pr \left\{ V_1 < \frac{-\frac{1}{2}\eta^2(0) - \ln\left(\frac{1-\theta_1}{1-\theta_2}\right)}{|\eta(0)|} \right\} + \\ &\quad \theta_1 \Pr \left\{ V_1 < \frac{-\frac{1}{2}\eta^2(1) - \ln\left(\frac{\theta_1}{\theta_2}\right)}{|\eta(1)|} \right\} \end{aligned} \tag{3.10}$$

and

$$
\begin{aligned}
e_{21} &= (1-\theta_2)\Pr[V^*(0) > 0 | w_0 \in \Pi_2] + \theta_2 \Pr[V^*(1) > 0 | w_0 \in \Pi_2] \\
&= (1-\theta_2)\Pr\left\{V_1 > \frac{\frac{1}{2}\eta^2(0) - \ln\left(\frac{1-\theta_1}{1-\theta_2}\right)}{|\eta(0)|}\right\} + \\
&\quad \theta_2 \Pr\left\{V_1 > \frac{\frac{1}{2}\eta^2(1) - \ln\left(\frac{\theta_1}{\theta_2}\right)}{|\eta(1)|}\right\} \qquad (3.11)
\end{aligned}
$$

The non-robust character of the $V^*(x)$ procedure is explained by equations (3.10) and (3.11). Since θ_i, $\eta(x)$ and C are fixed, the values of e_{12} and e_{21} given by (3.10) and (3.11) will clearly change with the distribution of V_1 (or y_0). Consider now the $V_R^*(x)$ procedure. If y_0 has a symmetric distribution, $\hat{\mu}_i(x)$ and $\hat{\sigma}(x)$ converge to their expected values $\mu_i(x)$ and $h\sigma(x)$, respectively, where $h = \frac{E(\hat{\sigma}(x))}{\sigma(x)}$; see Tiku (1980,1982) and Tiku, Tan and Balakrishnan (1986). In that case $V_R^*(x)$, like $V^*(x)$, is a linear function of y_0. Thus, for the $V_R^*(x)$ procedure

$$
\begin{aligned}
E(V_R^*(x)) &= (-1)^{i-1}\frac{\eta^2(x)}{2h^2} + (1-x)\ln\left(\frac{1-\theta_1}{1-\theta_2}\right) + \\
&\quad x\ln\left(\frac{\theta_1}{\theta_2}\right) \quad w_0 \in \Pi_i \ (i = 1,2) \qquad (3.12) \\
Var(V_R^*(x)) &= \frac{\eta^2(x)}{h^4} \qquad (3.13)
\end{aligned}
$$

Define

$$
V_2 = \frac{V_R^*(x) - E(V_R^*(x))}{\sqrt{Var(V_R^*(x))}} \qquad (3.14)
$$

The asymptotic distribution of V_2 is exactly the same as that of V_1 since both are linear in y_0 (asymptotically). Consequently, the asymptotic e_{12} and e_{21} values of $V_R^*(x)$ procedure are given by

$$
\begin{aligned}
e_{12} &= (1-\theta_1)\Pr[V_R^*(0) < 0 | w_0 \in \Pi_1] + \theta_1 \Pr[V_R^*(1) < 0 | w_0 \in \Pi_1] \\
&= (1-\theta_1)\Pr\left\{V_2 < \frac{-\frac{1}{2h^2}\eta^2(0) - \ln\left(\frac{1-\theta_1}{1-\theta_2}\right)}{\frac{|\eta(0)|}{h^2}}\right\} + \\
&\quad \theta_1 \Pr\left\{V_1 < \frac{-\frac{1}{2h^2}\eta^2(1) - \ln\left(\frac{\theta_1}{\theta_2}\right)}{\frac{|\eta(1)|}{h^2}}\right\}
\end{aligned}
$$

$$
\begin{aligned}
&= (1-\theta_1)\Pr\left\{V_2 < \frac{-\frac{1}{2}\eta^2(0) - h^2\ln\left(\frac{1-\theta_1}{1-\theta_2}\right)}{|\eta(0)|}\right\} + \\
&\quad \theta_1 \Pr\left\{V_2 < \frac{-\frac{1}{2}\eta^2(1) - h^2\ln\left(\frac{\theta_1}{\theta_2}\right)}{|\eta(1)|}\right\} \qquad (3.15)
\end{aligned}
$$

and

$$
\begin{aligned}
e_{21} &= (1-\theta_2)\Pr[V^*(0) > 0 | w_0 \in \Pi_2] + \theta_2 \Pr[V^*(1) > 0 | w_0 \in \Pi_2] \\
&= (1-\theta_2)\Pr\left\{V_2 > \frac{\frac{1}{2h^2}\eta^2(0) - \ln\left(\frac{1-\theta_1}{1-\theta_2}\right)}{\frac{|\eta(0)|}{h^2}}\right\} + \\
&\quad \theta_2 \Pr\left\{V_2 > \frac{\frac{1}{2h^2}\eta^2(1) - \ln\left(\frac{\theta_1}{\theta_2}\right)}{\frac{|\eta(1)|}{h^2}}\right\}; \\
&= (1-\theta_2)\Pr\left\{V_2 > \frac{\frac{1}{2}\eta^2(0) - h^2\ln\left(\frac{1-\theta_1}{1-\theta_2}\right)}{|\eta(0)|}\right\} + \\
&\quad \theta_2 \Pr\left\{V_2 > \frac{\frac{1}{2}\eta^2(1) - h^2\ln\left(\frac{\theta_1}{\theta_2}\right)}{|\eta(1)|}\right\} \qquad (3.16)
\end{aligned}
$$

Since h assumes different values for different distributions, the points

$$
\frac{\pm\frac{1}{2}\eta^2(0) - h^2\ln\left(\frac{1-\theta_1}{1-\theta_2}\right)}{|\eta(0)|} \quad \text{and} \quad \frac{\pm\frac{1}{2}\eta^2(1) - h^2\ln\left(\frac{\theta_1}{\theta_2}\right)}{|\eta(1)|} \qquad (3.17)
$$

change with the distribution of V_2 (or y_0). Suppose that V_2 has the following symmetric family of distributions with probability density function

$$
f(z) \propto \left[1 + \frac{z^2}{k}\right]^{-p} \quad -\infty < z < \infty \quad (k = 2p-3, p \geq 2) \qquad (3.18)
$$

This family represents a wide class of non-normal distributions with kurtosis ranging between 3 ($p = \infty$) and ∞ (p=2) and has been extensively used in robustness studies; see Tiku , Tan and Balakrishnan (1986). For the family (3.18) the values of h are also given in Tiku, Tan and Balakrishnan (1986). Using these, the asymptotic values of e_{12} and e_{21} can be calculated from (3.15) and (3.16).

3.3.3 Monte Carlo results

It is difficult to evaluate the e_{12} and e_{21} values analytically for small sample sizes n_1 and n_2. So, to compare the $V^*(x_0)$ and $V_R^*(x_0)$ procedures, we carried out an extensive Monte Carlo study. We simulated their e_{12} and e_{21} values for the following distributions which in addition to including normal also has a wide range of non-normal distributions:

Symmetric Distributions: $z = \frac{y-\mu}{\sigma}$

(1) $N(\mu, \sigma^2)$: $\frac{1}{\sqrt{2\pi}} \exp -\frac{z^2}{2}$ $(-\infty < z < \infty)$

(2) Logistic: $\frac{\exp(-z)}{[1+\exp(-z)]^2}$ $(-\infty < z < \infty)$

(3) Double exponential: $\frac{1}{2}\exp(-|z|)$ $(-\infty < z < \infty)$

(4) Student's t_4 : $\frac{3}{8}\left\{1 + \frac{z^2}{4}\right\}^{-\frac{5}{2}}$ $(-\infty < z < \infty)$

Dixon's outlier models (see Tiku 1977):

(5) $(n-2)$ come from $N(0, \sigma^2)$ and 2 from $N(0, (4\sigma)^2)$ $(-\infty < z < \infty)$

(6) $(n-5)$ come from $N(0, \sigma^2)$ and 5 from $N(0, (4\sigma)^2)$ $(-\infty < z < \infty)$

(7) $(n-2)$ come from $N(0, \sigma^2)$ and 2 from $N(0, (10\sigma)^2)$ $(-\infty < z < \infty)$

(8) $(n-5)$ come from $N(0, \sigma^2)$ and 5 from $N(0, (10\sigma)^2)$ $(-\infty < z < \infty)$

Mixture Model (see also Balakrsihnan and Tiku (1988b)

(9) $0.90N(0, \sigma^2) + 0.10N(0, (4\sigma)^2)$

(10) $0.90N(0, \sigma^2) + 0.10N(0, (10\sigma)^2)$

Skew distributions

(11) χ_6^2 : $\frac{1}{16}z^2 \exp(-\frac{z}{2})$ $(0 < z < \infty)$

(12) Exponential: $exp(-z)$ $(0 < z < \infty)$

(3.19)

A location-shift between Π_1 and Π_2 was created by adding a constant d to the sample observations generated from Π_2. Thus, Π_1 has mean $\mu_1 = \mu$ and Π_2 has mean $\mu_2 = d + \mu$; Π_1 and Π_2 are assumed to have the same variance as said earlier. Since $V^*(x)$ and $V_R^*(x)$ are location and scale invariant, μ and σ in (3.19) can be taken to be equal to 0 and 1, respectively, without any loss of generality. The simulated values of e_{12} and e_{21} were obtained as follows:

From the populations Π_i $(i = 1, 2)$, with specified values of the probabilities

$\Pr[X = 1] = \theta_i$ $(i = 1,\ 2)$, and specified values of the means $\mu_i(x)$ $(i = 1,\ 2)$ and a specified value of the common variance $\sigma^2(x)$ of the conditional distributions of Y given $X = x$, random samples of sizes n_1 and n_2 were generated using the IMSL (1977) routines. From these samples, the estimates $\hat{\theta}_i$ $(i = 1,\ 2)$, $\bar{y}_i(x)$ and $\hat{\mu}_i(x)$ $(i = 1,\ 2)$ and $s^2(x)$ and $\hat{\sigma}^2(x)$ were computed and incorporated in expressions (3.7) and (3.8). A new random observation from Π_1

$$\mathbf{w}_0 = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}$$

was then generated and classified based on the calculated values of $V^*(x_0)$ and $V_R^*(x_0)$. The above computations were repeated $N = 50,000/n$, $n = \min(n_1, n_2)$, times. The simulated values of e_{12} of the $V^*(x)$ and $V_R^*(x)$ procedures (given in Table 3.3-2.5) are then simply the number of negative values they assume divided by N. The simulated values of e_{21} were similarly obtained by generating a new observation from Π_2. The simulated values of e_{12} and e_{21} of the $V^*(x)$ and $V_R^*(x)$ are given in Tables 3.3-2.4 for $n_1 = n_2 = 50$. The standard errors in these values are well within $\pm$ 0.006. As expected, the e_{12} and e_{21} values, and the average error rates $(e_{12} + e_{21})/2$, of the two procedures are exactly the same for normal populations. For non-normal populations, the $V_R^*(x)$ procedure has smaller average error rates, particularly in situations when the samp' : contains outliers. Besides, the e_{12} and e_{21} values of the $V_R^*(x)$ procedure are more stable (robust) from distribution to distribution than those of the $V(x)$ procedure. Clearly, we observe the $V_R^*(x)$ procedure to be superior.

We also simulated the e_{12} and e_{21} values for smaller sample sizes and arrived at exactly the same conclusions as above. For $n_1 = n_2 = 25$ and $\theta_1 = \theta_2 = 0.5$, for example, we have the values given in Table 3.5.

Difficulties arise in case of small samples (less than 20, say), e.g., $\hat{\theta}_1$ or $\hat{\theta}_2$ may work out to be zero (or 1) in which case some of the estimates $\bar{y}_i(x)$ and $\hat{\mu}_i(x)$ and $s_i^2(x)$ and $\hat{\sigma}_i^2(x)$ may not be available for the computation of $V^*(x)$ and $V_R^*(x)$. So, one needs reasonably large sample sizes for the implementation of the classification procedures described in this chapter as only we could get estimates of

Table 3.3: Simulated values of e_{12} and e_{21} case I
Simulated values of e_{12} and e_{21} for the classical and the robust linear classification procedures for the models in (3.19) when $\mu_1 = 0, \sigma^2 = \gamma^2 = 1, \delta = 1$ and $n_1 = n_2 = n = 50, \theta_1 = \theta_2 = 0.5$

Model	e_{12} values						e_{21} values					
	$\mu_2 = 1.0$		$\mu_2 = 2.0$		$\mu_2 = 3.0$		$\mu_2 = 1.0$		$\mu_2 = 2.0$		$\mu_2 = 3.0$	
	V^*	V_R^*	V^*	V_R^*	V^*	V_R^*	V^*	V_R^*	V^*	V_R^*	V^*	V_R^*
(1)	0.29	0.29	0.17	0.17	0.06	0.06	0.33	0.33	0.16	0.16	0.06	0.06
(2)	0.28	0.27	0.16	0.15	0.09	0.09	0.31	0.31	0.14	0.14	0.06	0.06
(3)	0.28	0.28	0.13	0.12	0.08	0.07	0.25	0.24	0.12	0.12	0.06	0.06
(4)	0.29	0.29	0.12	0.12	0.05	0.05	0.27	0.26	0.12	0.12	0.05	0.05
(5)	0.32	0.31	0.16	0.16	0.05	0.05	0.35	0.34	0.18	0.17	0.07	0.07
(6)	0.34	0.32	0.17	0.15	0.09	0.08	0.37	0.33	0.19	0.18	0.06	0.06
(7)	0.33	0.29	0.22	0.17	0.09	0.07	0.35	0.30	0.24	0.18	0.10	0.08
(8)	0.42	0.33	0.26	0.16	0.15	0.07	0.44	0.34	0.29	0.19	0.14	0.06
(9)	0.31	0.30	0.19	0.18	0.07	0.07	0.35	0.34	0.18	0.16	0.07	0.06
(10)	0.38	0.30	0.28	0.18	0.14	0.07	0.41	0.33	0.27	0.19	0.14	0.08
(11)	0.29	0.30	0.16	0.17	0.08	0.09	0.37	0.35	0.14	0.11	0.01	.003
(12)	0.20	0.22	0.15	0.16	0.08	0.09	0.38	0.33	0.06	0.02	.004	.002

various parameters involved in the model.

Table 3.4: Simulated values of e_{12} and e_{21} case II
Simulated values of e_{12} and e_{21} for the classical and the robust linear classification procedures for the models
in (3.19) when $\mu_1 = 0, \sigma^2 = \gamma^2 = 1, \delta = 1$ and $n_1 = n_2 = n = 50, \theta_1 = 0.75, \theta_2 = 0.25$

Model	e_{12} values						e_{21} values					
	$\mu_2 = 1.0$		$\mu_2 = 2.0$		$\mu_2 = 3.0$		$\mu_2 = 1.0$		$\mu_2 = 2.0$		$\mu_2 = 3.0$	
	V^*	V_R^*	V^*	V_R^*	V^*	V_R^*	V^*	V_R^*	V^*	V_R^*	V^*	V_R^*
(1)	0.22	0.22	0.14	0.14	0.05	0.05	0.22	0.22	0.14	0.14	0.06	0.06
(2)	0.22	0.21	0.13	0.13	0.05	0.05	0.22	0.22	0.13	0.12	0.06	0.06
(3)	0.24	0.24	0.12	0.11	0.06	0.06	0.22	0.21	0.12	0.11	0.06	0.06
(4)	0.24	0.23	0.12	0.12	0.06	0.06	0.20	0.20	0.14	0.12	0.06	0.06
(5)	0.32	0.22	0.14	0.13	0.05	0.05	0.23	0.23	0.16	0.16	0.08	0.06
(6)	0.23	0.22	0.18	0.14	0.10	0.06	0.22	0.21	0.29	0.15	0.22	0.07
(7)	0.22	0.20	0.21	0.15	0.14	0.06	0.24	0.23	0.19	0.14	0.13	0.06
(8)	0.24	0.23	0.25	0.15	0.22	0.07	0.23	0.21	0.29	0.15	0.22	0.07
(9)	0.25	0.23	0.18	0.14	0.09	0.06	0.22	0.22	0.19	0.14	0.09	0.04
(10)	0.23	0.20	0.24	0.16	0.21	0.08	0.25	0.24	0.23	0.15	0.22	0.08
(11)	0.26	0.27	0.13	0.14	0.08	0.08	0.20	0.19	0.12	0.10	0.03	0.02
(12)	0.22	0.23	0.13	0.14	0.08	0.08	0.18	0.16	0.12	0.08	0.01	.003

Table 3.5: Simulated values of e_{12} and e_{21} case III
Simulated values of e_{12} and e_{21} for the classical and the robust linear classification procedures for the models in (3.19) when $\mu_1 = 0, \sigma^2 = \gamma^2 = 1, \delta = 1$ and $n_1 = n_2 = n = 25, \theta_1 = \theta_2 = 0.5$

Model	e_{12} values						e_{21} values					
	$\mu_2 = 1.0$		$\mu_2 = 2.0$		$\mu_2 = 3.0$		$\mu_2 = 1.0$		$\mu_2 = 2.0$		$\mu_2 = 3.0$	
	V^*	V_R^*	V^*	V_R^*	V^*	V_R^*	V^*	V_R^*	V^*	V_R^*	V^*	V_R^*
(1)	0.31	0.31	0.19	0.19	0.06	0.06	0.34	0.34	0.16	0.16	0.07	0.07
(2)	0.30	0.29	0.17	0.16	0.06	0.06	0.32	0.33	0.15	0.15	0.06	0.06
(3)	0.29	0.29	0.13	0.14	0.08	0.07	0.28	0.26	0.13	0.13	0.07	0.07
(4)	0.29	0.30	0.13	0.12	0.05	0.05	0.29	0.28	0.14	0.13	0.06	0.06
(5)	0.34	0.32	0.17	0.16	0.06	0.06	0.38	0.36	0.19	0.17	0.09	0.08
(6)	0.41	0.38	0.23	0.20	0.11	0.10	0.40	0.37	0.23	0.20	0.11	0.08
(7)	0.41	0.33	0.27	0.18	0.15	0.08	0.42	0.33	0.29	0.20	0.17	0.09
(8)	0.47	0.42	0.38	0.26	0.29	0.17	0.46	0.42	0.36	0.25	0.29	0.16
(9)	0.36	0.33	0.21	0.19	0.08	0.07	0.38	0.36	0.19	0.17	0.08	0.07
(10)	0.41	0.33	0.27	0.21	0.16	0.08	0.43	0.36	0.30	0.22	0.19	0.11
(11)	0.31	0.32	0.16	0.17	0.08	0.08	0.36	0.35	0.14	0.12	0.03	0.02
(12)	0.23	0.25	0.15	0.17	0.08	0.09	0.37	0.31	0.09	0.05	.004	001

Chapter 4

Multivariate Linear Classification Procedure Based On Dichotomous and Continuous Variables

4.1 Introduction

In the last chapter, we considered the classification problem based on a dichotomous and an univariate normal variable and described in Section 3.2 the classical linear procedure of Chang and Afifi (1974) and the analogous robust procedure of Balakrishnan and Tiku (1988b). We also discussed there the asymptotic determination of the cut-off points of these two procedures, when one of the errors of misclassification is fixed at a specified level.

In this chapter, we consider the classification problem based on a dichotomous and a multivariate normal variable and discuss the classical linear and the analogous robust linear classification procedures. We discuss these two procedures when one of the two errors of misclassification is at a pre-fixed level and also when both the errors are allowed to float (that is, when neither of the two errors are fixed). By considering the bivariate case and simulating the values of the errors of misclassification e_{12} and e_{21} of the two procedures, when e_{12} is fixed at a specified level, under the bivariate normal and various non-normal models, we show that the

robust procedure developed in this chapter is quite efficient and robust to departures from normality.

In Section 2, we first describe the basic model of the classification problem based on a dichotomous and a multivariate normal variable. In Section 3, by considering the case when all the parameters of the model are known we derive the likelihood ratio rule for classifying a new observation and derive exact formulas for the two errors of misclassification. In Section 4, we consider the case when all the parameters of the model are unknown and describe the classical linear classification procedure due to Chang and Afifi (1974) and discuss its asymptotic determination of the cut-off point when one of the errors of misclassification is at a prefixed level. By making use of the multivariate MML estimators of μ and Σ derived in Chapter 2, we propose an analogous robust linear classification procedure and discuss its asymptotic determination of the cut-off point when one of the errors of misclassification is at a pre-fixed level. This work generalizes the univariate linear classification procedure of Balakrishnan and Tiku (1988b) described in the last chapter and also extends the robust multivariate linear classification procedure proposed in Chapter 2 to cover situations where the data have multivariate continuous observations with an associated binary variable (like gender or social class or age group). We discuss some asymptotic properties of this robust linear procedure. In Section 5, we consider the bivariate case and carry out Monte Carlo comparisons of the two procedures when one of the two errors of misclassification is at a pre-fixed level and show the robust linear classification procedure developed in this chapter is more efficient and robust to departures from bivariate normality as compared to the classical linear classification procedure. Finally, we present an example through which we illustrate these two procedures.

4.2 The model

Let X be a random Bernoulli variable with probability mass function

$$\Pr(X = x) = \theta^x(1-\theta)^{1-x}, \quad x = 0, 1,$$

and $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}$ be a p-variate continuous variable such that the conditional distribution of $\mathbf{Y}$, given $X = x$, is p-variate normal with mean

$$\mu(x) = \begin{pmatrix} \mu_1(x) \\ \mu_2(x) \\ \vdots \\ \mu_p(x) \end{pmatrix} = \mu + x\Delta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} + x \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_p \end{pmatrix}$$

and variance-covariance matrix

$$\Sigma(x) = \Sigma + x\Gamma.$$

Then the joint density function of $\mathbf{w} = \begin{pmatrix} x \\ \mathbf{y} \end{pmatrix}$

$$f(\mathbf{w}) = \theta^x(1-\theta)^{1-x}\phi_p(\mu(x), \Sigma(x)),$$

where $\phi_p(\mu(x), \Sigma(x))$ is the conditional density function of $p-$variate normal distribution with mean $\mu(x)$ and variance-covariance matrix $\Sigma(x)$. The expected value of the random vector $\mathbf{W}$ is

$$\begin{aligned} E(\mathbf{W}) &= E(E(\mathbf{W}|X = x)) \\ &= E\begin{pmatrix} x \\ \mu + x\Delta \end{pmatrix} \\ &= \begin{pmatrix} \theta \\ \mu + \theta\Delta \end{pmatrix} \end{aligned} \tag{4.1}$$

and the variance-covariance matrix of the random vector $\mathbf{W}$ is

$$Cov(\mathbf{W}) = E\left(E\left(\mathbf{W}\mathbf{W}^T|X = x\right)\right) - E(\mathbf{W})E(\mathbf{W})^T,$$

but

$$E\left(\mathbf{W}\mathbf{W}^T|X = x\right) = E\left(\begin{pmatrix} X \\ \mathbf{Y} \end{pmatrix}\left(X\ \mathbf{Y}^T\right)|X = x\right)$$

$$
\begin{aligned}
&= E\left(\begin{pmatrix} X & X\mathbf{Y}^T \\ X\mathbf{Y} & \mathbf{Y}\mathbf{Y}^T \end{pmatrix} | X = x\right) \\
&= \begin{pmatrix} x & x(\mu^T + x\Delta^T) \\ x(\mu + x\Delta) & \Sigma + \Gamma x + (\mu + x\Delta)(\mu^T + x\Delta^T) \end{pmatrix} \quad (4.2)
\end{aligned}
$$

Therefore

$$
E\left(\mathbf{W}\mathbf{W}^T\right) = \begin{pmatrix} \theta & \theta\mu^T + \theta\Delta^T \\ \theta(\mu + \Delta) & \Sigma + \Gamma\theta + \mu\mu^T + \theta\Delta\mu^T + \theta\mu\Delta^T + \theta\Delta\Delta^T \end{pmatrix}, \quad (4.3)
$$

also

$$
E(\mathbf{W})E(\mathbf{W})^T = \begin{pmatrix} \theta^2 & \theta(\mu^T + \theta\Delta^T) \\ \theta(\mu + \theta\Delta) & +\mu\mu^T + \theta\Delta\mu^T + \theta\mu\Delta^T + \theta^2\Delta\Delta^T \end{pmatrix}, \quad (4.4)
$$

Hence,

$$
Cov(\mathbf{W}) = \begin{pmatrix} \theta(1-\theta) & \theta(1-\theta)\Delta^T \\ \theta(1-\theta)\Delta & \Sigma + \Gamma\theta + \theta(1-\theta)\Delta\Delta^T \end{pmatrix}. \quad (4.5)
$$

The marginal distribution of $\mathbf{Y}$ is

$$
\begin{aligned}
g(\mathbf{y}) &= \sum_{x=0}^{1} \theta^x (1-\theta)^{1-x} \phi_p(\mu(x), \Sigma(x)) \\
&= (1-\theta)\phi_p(\mu(0), \Sigma(0)) + \theta\phi_p(\mu(1), \Sigma(1)), \quad (4.6)
\end{aligned}
$$

i.e. two-component multivariate mixed normal. It is noted that X and $\mathbf{Y}$ are independent if and only if $\mu(0) = \mu(1)$ and $\Sigma(0) = \Sigma(1)$. For the problem of classifying an observation $\mathbf{w}$ into one of two populations, Π_1 and Π_2, it is assumed that $\mathbf{w}$ is from Π_i, then its density function is

$$
f_i(\mathbf{w}) = \theta_i^x (1-\theta_i)^{1-x} \phi_p(\mu_i(x), \Sigma(x)), \quad i = 1, 2. \quad (4.7)
$$

It should be noted that the covariance matrix of $\mathbf{W}$ is not the same in Π_1 and Π_2. However, the conditional covariance matrices of $\mathbf{Y}$ given X are equal for the two populations. Thus Σ and Γ do not change with i, but θ_i and $\mu_i(x)$ do.

4.3 Classification when the population parameters are known

In this section, we assume that the parameters of the two populations are known and proceed to derive the likelihood ratio classification procedure and the associated probabilities of misclassification for this model.

4.3.1 Likelihood ratio classification procedure

For the model specified by (4.7), the likelihood ratio is

$$\begin{aligned}
V(x) &= \ln\left[\frac{f_1(\mathrm{w})}{f_2(\mathrm{w})}\right] \\
&= -\frac{1}{2}(\mathrm{y}-\mu_1(x))^T\Sigma^{-1}(x)(\mathrm{y}-\mu_1(x)) \\
&\quad +\frac{1}{2}(\mathrm{y}-\mu_2(x))^T\Sigma^{-1}(x)(\mathrm{y}-\mu_2(x)) \\
&\quad +x\ln\left[\frac{\theta_1}{\theta_2}\right]+(1-x)\ln\left[\frac{1-\theta_1}{1-\theta_2}\right] \\
&= -\frac{1}{2}(\mathrm{y}^T\Sigma^{-1}(x)\mathrm{y}-2\mathrm{y}^T\Sigma^{-1}(x)\mu_1(x)+\mu_1^T(x)\Sigma^{-1}(x)\mu_1(x)) \\
&\quad +\frac{1}{2}(\mathrm{y}^T\Sigma^{-1}(x)\mathrm{y}-2\mathrm{y}^T\Sigma^{-1}(x)\mu_2(x)+\mu_2^T(x)\Sigma^{-1}(x)\mu_2(x)) \\
&\quad +x\ln\left[\frac{\theta_1}{\theta_2}\right]+(1-x)\ln\left[\frac{1-\theta_1}{1-\theta_2}\right] \\
&= \mathrm{y}^T\Sigma^{-1}(\mu_1(x)-\mu_2(x)) \\
&\quad -\frac{1}{2}\mu_1^T(x)\Sigma^{-1}(x)\mu_1(x)+\frac{1}{2}\mu_2^T(x)\Sigma^{-1}(x)\mu_2(x) \\
&\quad +x\ln\left[\frac{\theta_1}{\theta_2}\right]+(1-x)\ln\left[\frac{1-\theta_1}{1-\theta_2}\right] \\
&= \left\{\mathrm{y}-\frac{1}{2}(\mu_1(x)+\mu_2(x)\right\}\Sigma^{-1}(x)(\mu_1(x)-\mu_2(x)) \\
&\quad +x\ln\left[\frac{\theta_1}{\theta_2}\right]+(1-x)\ln\left[\frac{1-\theta_1}{1-\theta_2}\right]
\end{aligned} \tag{4.8}$$

Thus the classification procedure is to classify an observation $\mathbf{w}_0$ into Π_1 or Π_2 according as

$$V(x_0) \underset{<}{>} C \tag{4.9}$$

where the cut-off point C is determined so that the error of misclassification e_{12} (for example) is at a pre-specified level. From this we obtain two discriminant functions for $x_0 = 0$ and $x_0 = 1$. Chang and Afifi (1974) named this as the double discriminant function (DDF).

It is noted that when X and $\mathbf{Y}$ are independent (i.e., $\mu_i(0) = \mu_i(1),\ \ i = 1, 2$, and $\Sigma(0) = \Sigma(1)$) and X has the same distribution in Π_1 and Π_2 (i.e., $\theta_1 = \theta_2$), then we see that $V(0) = V(1)$. Thus all the information for classification comes from $\mathbf{Y}$ and the value of the dichotomous variable X is not involved in the decision. In this case, the conditional and the marginal distributions of $\mathbf{Y}$ are identically normal and the classical classification procedure based on simply the p continuous variables is the same as the DDF procedure.

4.3.2 Errors of misclassification

Since $V(x)$ is a linear combination of normal random variates, given $\mathbf{w}$ is from Π_1, the conditional distribution of $V(x)$ given $X = x$ is normal with mean

$$\frac{1}{2}\left\{\mu_1(x) - \mu_2(x)\right\}^T \Sigma^{-1}(x)(\mu_1(x) - \mu_2(x)) + x \ln\left[\frac{\theta_1}{\theta_2}\right] + (1-x)\ln\left[\frac{1-\theta_1}{1-\theta_2}\right]$$

and variance

$$\eta^2(x) = \left\{\mu_1(x) - \mu_2(x)\right\}^T \Sigma^{-1}(x)(\mu_1(x) - \mu_2(x)).$$

Also given $\mathbf{w}$ is from Π_2 the conditional distribution of $V(x)$ given $X = x$ is normal with mean $-\eta^2(x) + x\ln\left[\frac{\theta_1}{\theta_2}\right] + (1-x)\ln\left[\frac{1-\theta_1}{1-\theta_2}\right]$ and variance $\eta^2(x)$. The probability of misclassifying an observation from Π_1 into Π_2 is, therefore,

$$\begin{aligned} e_{12} &= Pr(V(x) < C | \mathbf{w} \in \Pi_1) \\ &= (1-\theta_1)P(V(0) < C | \mathbf{y} \sim \phi_p(\mu_1(0), \Sigma(0))) + \theta_1 P(V(1) < C | \mathbf{y} \sim \phi_p(\mu_1(1), \Sigma(1))) \\ &= (1-\theta_1)\Phi\left\{\frac{C - \frac{1}{2}\eta^2(0) - \ln\left[\frac{1-\theta_1}{1-\theta_2}\right]}{|\eta(0)|}\right\} + \theta_1\Phi\left\{\frac{C - \frac{1}{2}\eta^2(1) - \ln\left[\frac{\theta_1}{\theta_2}\right]}{|\eta(1)|}\right\}, \end{aligned} \tag{4.10}$$

and the probability of misclassifying an observation from Π_2 into Π_1 is

$$\begin{aligned} e_{21} &= Pr(V(x) > C|\mathrm{w} \in \Pi_2) \\ &= (1-\theta_2)P(V(0) > C|\mathrm{y} \sim \phi_p(\mu_2(0), \Sigma(0))) + \theta_2 P(V(1) > C|\mathrm{y} \sim \phi_p(\mu_2(1), \Sigma(1))) \\ &= (1-\theta_2)\Phi\left\{-\frac{C + \frac{1}{2}\eta^2(0) - \ln\left[\frac{1-\theta_1}{1-\theta_2}\right]}{|\eta(0)|}\right\} + \theta_2\Phi\left\{-\frac{C + \frac{1}{2}\eta^2(1) - \ln\left[\frac{\theta_1}{\theta_2}\right]}{|\eta(1)|}\right\} \end{aligned} \tag{4.11}$$

4.4 Classification when population parameters are unknown

Since the population parameters are unknown in most applications, the usual practice is to substitute estimates of the population parameters θ_i, $\mu_i(x)$, $\Sigma(x)$, $x = 0,\ 1$ into $V(x)$. Thus, let $\begin{pmatrix} x_{1i} \\ \mathrm{y}_{1i} \end{pmatrix}$, $i = 1, 2, \ldots, n_1$, be a random sample of size n_1 from population Π_1 which is the model specified above with θ_1 and $\mu_1(x)$. Similarly, let $\begin{pmatrix} x_{2i} \\ \mathrm{y}_{2i} \end{pmatrix}$, $i = 1, 2, \ldots, n_2$, be a random sample of size n_2 from population Π_2 which is the model specified above with θ_2 and $\mu_2(x)$.

4.4.1 Classical classification procedure

Chang and Afifi (1974) have used the maximum likelihood estimators obtained from the two samples mentioned above. These estimates of the parameters are given by

$$\hat{\theta}_i = \frac{n_i(1)}{n_i}, \quad (i = 1, 2),$$

where $n_i(1)$ is the number of observations in the ith sample that correspond to $x = 1$, and

$$\hat{\mu}_i(x) = \bar{\mathrm{y}}_i(x)$$

and

$$\hat{\Sigma}(x) = \mathrm{S}(x) = \frac{(n_1(x) - 1)\mathrm{S}_1(x) + (n_2(x) - 1)\mathrm{S}_2(x)}{n_1(x) + n_2(x) - 2}$$

for $x = 0$ and 1, and $i = 1$ and 2; here, $\mathbf{S}_i(x)$ is the sample variance-covariance matrix from the ith sample and is given by

$$\mathbf{S}_i(x) = \frac{1}{n_i(x) - 1} \sum_{l=1}^{n_i(x)} (\mathbf{y}_{il}(x) - \bar{\mathbf{y}}_i(x))(\mathbf{y}_{il}(x) - \bar{\mathbf{y}}_i(x))^T.$$

Let $\begin{pmatrix} x_0 \\ \mathbf{y}_0 \end{pmatrix}$ be an independent new observation that needs to be classified into either Π_1 or Π_2. Then, as shown by Chang and Afifi (1974) and Krzanowski (1975) the classical linear classification procedure is to classify $\begin{pmatrix} x_0 \\ \mathbf{y}_0 \end{pmatrix}$ into Π_1 or Π_2 according as

$$\begin{aligned} V(x_0) &= \left\{\mathbf{y}_0 - \frac{1}{2}(\bar{\mathbf{y}}_1(x_0) + \bar{\mathbf{y}}_2(x_0))\right\}^T \mathbf{S}^{-1}(x_0)(\bar{\mathbf{y}}_1(x_0) - \bar{\mathbf{y}}_2(x_0)) \\ &\quad + x_0 \ln\left[\frac{\hat{\theta}_1}{\hat{\theta}_2}\right] + (1 - x_0)\ln\left[\frac{1 - \hat{\theta}_1}{1 - \hat{\theta}_2}\right] \gtrless C^*, \end{aligned} \tag{4.12}$$

where the cut-off point C^* is determined so that the error of misclassification e_{12} (for example) is at a pre-specified level. The cut-off point may be obtained as the solution $\hat{C}^*$ of the equation (with e_{12} fixed as α) by substituting the appropriate estimators for the parameters $\eta^2(x)$ and θ_i in (4.10). The classical estimator of $\eta^2(x)$ is

$$\hat{\eta}^2(x) = \frac{n_1(x) + n_2(x) - p - 3}{n_1(x) + n_2(x) - 2}(\bar{\mathbf{y}}_1(x) - \bar{\mathbf{y}}_2(x))^T \mathbf{S}^{-1}(x)(\bar{\mathbf{y}}_1(x) - \bar{\mathbf{y}}_2(x)), \quad x = 0, 1,$$

i.e. Mahalanobis' sample squared distance. Note that this estimator is not unbiased but the unbiased estimator of $\eta^2(x)$ may take on inadmissible negative values as we mentioned in Chapter 2. Therefore the cut-off point $\hat{C}^*$ is the solution of

$$(1 - \hat{\theta}_1)\Phi\left\{\frac{\hat{C}^* - \frac{1}{2}\hat{\eta}^2(0) - \ln\left(\frac{1-\hat{\theta}_1}{1-\hat{\theta}_2}\right)}{|\hat{\eta}(0)|}\right\} + \hat{\theta}_1\Phi\left\{\frac{\hat{C}^* - \frac{1}{2}\hat{\eta}^2(1) - \ln\left(\frac{\hat{\theta}_1}{\hat{\theta}_2}\right)}{|\hat{\eta}(1)|}\right\} = \alpha. \tag{4.13}$$

In the case when both the errors of misclassification are allowed to float (that is, neither of the two errors are fixed), the classical linear classification procedure is to

classify the new independent observation $\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ into Π_1 or Π_2 or arbitrarily into one of them according as

$$\begin{aligned} V(x_0) &= \left\{y_0 - \frac{1}{2}(\bar{y}_1(x_0) + \bar{y}_2(x_0))\right\}^T S^{-1}(x_0)(\bar{y}_1(x_0) - \bar{y}_2(x_0)) \\ &\quad + x_0 \ln\left[\frac{\hat{\theta}_1}{\hat{\theta}_2}\right] + (1 - x_0)\ln\left[\frac{1-\hat{\theta}_1}{1-\hat{\theta}_2}\right] \gtrless = 0. \end{aligned} \tag{4.14}$$

4.4.2 Robust Linear Classification Procedure

Just as Balakrishnan and Tiku (1988b) proposed a robust univariate double linear classification procedure analogous to the classical univariate double linear procedure given by Chang and Afifi (1974) (see Section 3.2), we make use of the multivariate MML estimators derived in Chapter 2 to propose in this section a robust multivariate double discriminant function analogous to the classical one described in Section 4.4.1. We also discuss some asymptotic optimal properties of this robust linear classification procedure. This procedure generalizes the univariate procedure of Balakrishnan and Tiku (1988b) in a natural way to the multivariate situation and also extends the robust multivariate linear classification procedure proposed in Chapter 2 to handle situations where the data have multivariate continuous observations and an associated binary observation which may denote the gender or the social class. Let $(\hat{\mu}_1(x), \hat{\Sigma}_1(x))$ and $(\hat{\mu}_2(x), \hat{\Sigma}_2(x))$ be the MML estimators of $(\mu_1(x), \Sigma(x))$ and $(\mu_2(x), \Sigma(x))$ obtained from the samples $y_{1k}(x)$, $k = 1, 2, \ldots, n_1(x)$, and $y_{2k}(x)$, $k = 1, 2, \ldots, n_2(x)$, respectively for $x = 0, 1$. Let $\hat{\Sigma}(x)$ be the pooled MML estimator of $\Sigma(x)$ given by

$$\hat{\Sigma}(x) = \frac{(A_1(x) - 1)\hat{\Sigma}_1(x) + (A_2 - 1)\hat{\Sigma}_2(x)}{A_1(x) + A_2(x) - 2}. \tag{4.15}$$

Then the robust classification procedure for classifying a new independent observation $\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ is to classify it into Π_1 or Π_2 according as

$$V_R(x_0) = \left\{y_0 - \frac{1}{2}(\hat{\mu}_1(x_0) + \hat{\mu}_2(x_0))\right\}^T \hat{\Sigma}^{-1}(x_0)(\hat{\mu}_1(x_0) - \hat{\mu}_2(x_0))$$

$$+x_0 \ln\left[\frac{\hat{\theta}_1}{\hat{\theta}_2}\right] + (1-x_0)\ln\left[\frac{1-\hat{\theta}_1}{1-\hat{\theta}_2}\right] \gtrless C_R^*, \tag{4.16}$$

where the cut-off point C_R^* is determined so that the error of misclassification e_{12} is fixed as α. Therefore cut-off point C_R^* can be obtained as a solution of (4.10) after substituting MML estimators of $\mu(x)$ and $\Sigma(x)$ which is given by

$$(1-\hat{\theta}_1)\Phi\left\{\frac{\hat{C}_R^* - \frac{1}{2}\hat{\eta}_R^2(0) - \ln\left(\frac{1-\hat{\theta}_1}{1-\hat{\theta}_2}\right)}{|\hat{\eta}_R(0)|}\right\} + \hat{\theta}_1\Phi\left\{\frac{\hat{C}_R^* - \frac{1}{2}\hat{\eta}_R^2(1) - \ln\left(\frac{\hat{\theta}_1}{\hat{\theta}_2}\right)}{|\hat{\eta}_R(1)|}\right\} = \alpha, \tag{4.17}$$

where

$$\hat{\eta}_R^2(x) = \frac{A_1(x)+A_2(x)-p-3}{A_1(x)+A_2(x)-2}(\hat{\mu}_1(x)-\hat{\mu}_2(x))^T\Sigma^{-1}(x)(\hat{\mu}_1(x)-\hat{\mu}_2(x)), \quad x=0,1. \tag{4.18}$$

In the case when both the errors of misclassification are allowed to float, the robust linear classification procedure is to classify the new independent observation $\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ into Π_1 or Π_2, or arbitrarily into Π_1 or Π_2, according as

$$V_R(x_0) \gtrless 0, \tag{4.19}$$

where the statistic $V_R(x_0)$ is as given in (4.16).

Theorem 9 *The statistic $V_R(x_0)$ in (4.16) asymptotically has exactly the same distribution as the statistic $V(x_0)$ in (4.12), under normality.*

Proof: From the properties of multivariate MML estimators discussed in Chapter 2 we have for large n_1 and n_2 the distribution of the statistic $V_R(x_0)$ in (4.16) to be similar to the distribution of the statistic $V(x_0)$ in (4.8) under normality. Also, from the classical theory of ML estimators the distribution of the statistic $V(x_0)$ in (4.12)is similar to the distribution of the statistics $V(x_0)$ in (4.8) for large n_1 and n_2 under normality. Hence the theorem.

∎

Theorem 10 *Asymptotically $V(x_0)$ and $V_R(x_0)$ procedures have exactly the same $1 - e_{21}$ value, for a pre-assigned value of e_{12}, under normality.*

Proof: The result follows from Theorem 4.4.2 and the fact that $\hat{C}^*$ and $\hat{C}_R^*$ both converge to C given by (4.10) as n_1 and n_2 become large, since $\hat{\theta}_i$ converges to θ_i, $\bar{y}_i(x)$ and $\hat{\mu}_i(x)$ both converge to $\mu_i(x)$, $i = 1, 2$ $(x = 0, 1)$, and $\mathbf{S}^2(x)$ and $\hat{\Sigma}(x)$ both converge to Σ. ∎

4.5 Comparison of the Two Procedures

It is very difficult to evaluate the e_{12} and e_{21} values analytically for small sample sizes n_1 and n_2. However, to compare the $V(x)$ and $V_R(x)$ procedures, we carried out an extensive Monte Carlo investigation described below. First, n pairs of pseudo-random observations were simulated through the equations

$$y_{11} = \frac{1}{\sqrt{2}}\left\{\sqrt{1+\rho}\, z_1 + \sqrt{1-\rho}\, z_2\right\} \tag{4.20}$$

and

$$y_{12} = \frac{1}{\sqrt{2}}\left\{\sqrt{1+\rho}\, z_1 - \sqrt{1-\rho}\, z_2\right\}, \tag{4.21}$$

where z_1 and z_2 are pseudo-random observations from each of the following seven models:

(1) Normal $N(0,1)$

(2) $y_1, y_2, \ldots, y_{n-1}$ come from $N(0,1)$ and y_n from $N(1,100)$

(3) $y_1, y_2, \ldots, y_{n-2}$ come from $N(0,1)$ and y_{n-1} and y_n from $N(0,16)$

(4) Logistic: $\frac{\exp -y}{(1+\exp -y)^2}$, $\quad -\infty < y < \infty$,

(5) Double exponential: $\frac{1}{2}\exp -|y|$, $\quad -\infty < y < \infty$,

(6) Mixture model: $0.90N(0,1) + 0.10N(0,16)$

and

(7) Student's t_4 : $\frac{3}{4\sigma\sqrt{\pi}(1+y^2)^{5/2}}$, $\quad -\infty < y < \infty$

(4.22)

Table 4.1: Simulated values errors of misclassification
Simulated values of e_{12} and $1 - e_{21}$ for the classical and the robust linear classification procedures for the models in (4.22) when $\theta_1 = \theta_2 = 0.1, n_1 = n_2 = 50$,

$$\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mu_2 = \begin{pmatrix} \mu \\ \mu \end{pmatrix}, \delta = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \Sigma = \begin{pmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{pmatrix}, \text{ and } \Gamma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Model	e_{12} values						$1 - e_{21}$ values					
	$\mu = 1$		$\mu = 2$		$\mu = 3$		$\mu = 1$		$\mu = 2$		$\mu = 3$	
	V	V_R	V	V_R	V	V_R	V	V_R	V	V_R	V	V_R
(1)	.053	.053	.041	.049	.044	.041	.335	.337	.725	.730	.940	.952
(2)	.025	.054	.024	.048	.020	.040	.170	.332	.465	.710	.743	.916
(3)	.037	.045	.032	.043	.030	.043	.215	.309	.589	.685	.859	.909
(4)	.049	.056	.047	.051	.045	.051	.318	.343	.742	.771	.940	.953
(5)	.059	.059	.068	.066	.063	.065	.205	.263	.507	.609	.815	.887
(6)	.061	.063	.051	.060	.043	.062	.151	.260	.466	.627	.768	.867
(7)	.065	.067	.061	.061	.060	.069	.170	.250	.513	.598	.811	.865

The y_2 sample was obtained simply by adding the vector μ_2, which created a location shift between the two populations Π_1 and Π_2. Then, with e_{12} fixed as 0.05, the values of e_{12} and $1 - e_{21}$ were simulated (based on 1,000 Monte Carlo runs) for both the procedures under all seven models in (4.22) with

$$\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mu_2 = \begin{pmatrix} \mu \\ \mu \end{pmatrix}, \delta = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \Sigma = \begin{pmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{pmatrix}, \Gamma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

for $\mu = 1(1)3, \theta_1 = \theta_2 = 0.1$ and $n_1 = n_2 = 50$. From the Table 4.2 we can see that the robust procedure in (4.16) has its e_{12} value to be very stable and close to the presumed value of 0.05 while that of the classical procedure in (4.12) seems to be very sensitive to departures from bivariate normality. We also observe that the robust procedure has its $1 - e_{21}$ value to be exactly the same as that of the classical procedure under normality, and at the same time has its $1 - e_{21}$ value to be much larger than that of the classical classification procedure under all the non-normal models.

Now, we shall consider an example in an anthropometric set-up and illustrate the classical and the robust linear classification procedures described above.

Example:

In Tables 4.2 and 4.3, we have presented a simulated data set in an anthropometric set-up which gives measurements on upper facial length and head length along with the gender for $n_1 = 100$ and $n_2 = 80$ individuals from two different groups. From these two samples, we find:

$$\bar{y}_1(0) = \begin{bmatrix} 39.74904 \\ 41.97538 \end{bmatrix} \qquad \bar{y}_2(0) = \begin{bmatrix} 35.67731 \\ 37.77911 \end{bmatrix}$$

$$\hat{\mu}_1(0) = \begin{bmatrix} 39.88131 \\ 41.95518 \end{bmatrix} \qquad \hat{\mu}_2(0) = \begin{bmatrix} 35.61678 \\ 37.69569 \end{bmatrix}$$

$$\bar{y}_1(1) = \begin{bmatrix} 43.01504 \\ 43.62766 \end{bmatrix} \qquad \bar{y}_2(1) = \begin{bmatrix} 38.43471 \\ 40.08711 \end{bmatrix}$$

$$\hat{\mu}_1(1) = \begin{bmatrix} 42.83341 \\ 43.22960 \end{bmatrix} \qquad \hat{\mu}_2(1) = \begin{bmatrix} 38.45158 \\ 40.17460 \end{bmatrix}$$

$$\mathbf{S}(0) = \begin{bmatrix} 26.94348 & 14.52765 \\ 14.52765 & 31.39485 \end{bmatrix} \quad \hat{\Sigma}(0) = \begin{bmatrix} 23.34883 & 11.74393 \\ 11.74393 & 32.47124 \end{bmatrix}$$

$$\mathbf{S}(1) = \begin{bmatrix} 28.12694 & 12.48926 \\ 12.48926 & 26.80156 \end{bmatrix} \quad \hat{\Sigma}(1) = \begin{bmatrix} 25.52330 & 9.40057 \\ 9.40057 & 18.93859 \end{bmatrix}$$

$$\theta_1 = 0.47000 \quad \theta_2 = 0.41250 \quad C = -1.07127 \quad C_R = -1.11036$$

So, with e_{12} (the probability of wrongly classifying an individual from Group 1 into Group 2) fixed as 0.05, by applying the classical classification procedure in (4.12) and reclassifying each individual in Tables 4.2 and 4.3, we obtain estimates of e_{12} and $1 - e_{21}$ (the probability of correctly classifying an individual from Group 2 into itself) to be 0.042 and 0.522, respectively. Similarly, by applying the robust linear classification procedure in (4.16) and reclassifying each individual in Tables 4.2 and 4.3, we obtain estimates of e_{12} and $1 - e_{21}$ to be 0.040 and 0.560, respectively.

Table 4.2: Measurements on upper facial length and head length of $n_1 = 100$ individuals from Group 1 along with their gender

i	x_{1i}	y_{1i}^T	i	x_{1i}	y_{1i}^T	i	x_{1i}	y_{1i}^T
1	1	(54.5 38.4)	35	0	(45.9 50.6)	69	1	(39.7 41.7)
2	1	(54.7 46.7)	36	1	(45.1 42.2)	70	0	(39.3 40.5)
3	1	(53.8 42.8)	37	1	(42.9 39.8)	71	1	(42.6 45.1)
4	0	(39.4 46.7)	38	0	(44.0 43.5)	72	1	(39.3 30.1)
5	0	(39.3 47.8)	39	0	(43.5 48.1)	73	1	(41.1 45.7)
6	0	(34.5 41.5)	40	0	(37.6 47.7)	74	1	(36.6 42.0)
7	0	(50.4 50.1)	41	1	(43.7 40.0)	75	1	(28.8 40.5)
8	1	(41.1 38.9)	42	0	(45.1 39.1)	76	0	(49.0 44.7)
9	0	(33.8 31.9)	43	1	(36.4 43.8)	77	0	(38.8 42.2)
10	1	(41.3 45.3)	44	0	(32.7 33.8)	78	0	(33.9 35.8)
11	1	(47.7 59.3)	45	0	(42.7 40.8)	79	0	(37.2 41.3)
12	1	(39.8 35.4)	46	1	(42.0 43.2)	80	1	(39.0 40.8)
13	0	(22.6 32.1)	47	1	(46.6 53.1)	81	0	(40.7 40.5)
14	1	(42.1 46.2)	48	1	(42.1 40.8)	82	1	(44.2 44.7)
15	0	(34.6 47.4)	49	1	(46.3 43.3)	83	1	(44.2 46.2)
16	1	(49.4 48.4)	50	0	(40.0 47.6)	84	1	(42.2 40.0)
17	0	(37.3 38.8)	51	1	(46.8 41.7)	85	1	(40.4 38.2)
18	0	(38.8 37.7)	52	0	(44.6 35.6)	86	1	(36.2 41.6)
19	0	(38.6 39.5)	53	0	(39.7 39.6)	87	1	(38.5 41.1)
20	0	(37.4 44.5)	54	0	(45.3 37.6)	88	0	(37.4 39.5)
21	0	(45.0 35.3)	55	0	(45.4 46.8)	89	0	(30.4 48.4)
22	0	(36.3 42.0)	56	0	(47.3 50.8)	90	1	(38.2 38.0)
23	1	(43.0 42.0)	57	0	(35.3 37.9)	91	0	(42.8 36.9)
24	1	(35.2 46.0)	58	1	(39.0 47.5)	92	0	(40.5 46.9)
25	1	(42.8 44.2)	59	0	(40.9 46.5)	93	1	(35.9 43.3)
26	0	(33.1 36.1)	60	0	(41.1 51.2)	94	0	(41.4 38.0)
27	1	(43.1 46.1)	61	0	(37.7 48.4)	95	0	(46.7 51.5)
28	1	(49.2 52.0)	62	0	(40.4 43.2)	96	0	(43.9 49.0)
29	0	(41.7 35.3)	63	0	(46.0 43.4)	97	1	(50.3 47.4)
30	0	(39.8 43.8)	64	1	(48.6 43.2)	98	1	(52.7 43.8)
31	0	(33.5 34.9)	65	1	(40.4 40.4)	99	1	(44.4 56.1)
32	0	(37.2 36.4)	66	0	(45.0 39.5)	100	0	(35.1 37.5)
33	1	(38.7 43.2)	67	1	(46.3 52.6)			
34	0	(35.9 35.5)	68	1	(44.7 37.5)			

Table 4.3: Measurements on upper facial length and head length of $n_1 = 80$ individuals from Group 2 along with their gender

i	x_{2i}	$\mathbf{y}_{2i}^T$	i	x_{2i}	$\mathbf{y}_{2i}^T$	i	x_{2i}	$\mathbf{y}_{2i}^T$
1	0	(43.5 33.5)	28	0	(33.9 40.3)	55	1	(39.0 37.1)
2	0	(35.0 36.5)	29	1	(35.0 40.7)	56	1	(42.2 41.8)
3	0	(29.0 31.8)	30	1	(39.9 40.3)	57	1	(34.9 42.3)
4	0	(31.4 35.6)	31	1	(29.8 36.0)	58	0	(27.6 33.3)
5	1	(46.9 41.0)	32	0	(25.8 27.9)	59	1	(34.0 35.9)
6	0	(39.5 49.0)	33	0	(31.5 37.3)	60	1	(37.8 33.5)
7	1	(38.4 46.6)	34	0	(30.6 33.4)	61	1	(42.8 46.9)
8	0	(34.7 31.6)	35	1	(42.5 35.5)	62	1	(32.6 39.6)
9	1	(34.9 33.2)	36	1	(33.0 36.2)	63	1	(40.5 47.5)
10	0	(52.7 49.7)	37	0	(36.9 35.6)	64	0	(39.1 44.8)
11	0	(31.1 32.5)	38	0	(30.8 26.4)	65	0	(45.0 46.7)
12	1	(39.3 39.2)	39	1	(51.9 49.6)	66	0	(32.7 38.9)
13	0	(41.7 37.9)	40	1	(26.7 26.4)	67	0	(26.0 37.1)
14	0	(36.4 37.1)	41	1	(34.3 40.6)	68	0	(38.3 38.8)
15	1	(35.5 35.0)	42	0	(35.0 36.7)	69	1	(42.4 42.2)
16	0	(40.8 35.7)	43	0	(43.5 42.7)	70	0	(38.0 36.1)
17	0	(39.5 39.8)	44	1	(45.7 51.7)	71	0	(31.7 42.2)
18	1	(37.5 45.8)	45	0	(37.5 38.1)	72	0	(39.5 39.7)
19	1	(39.1 40.8)	46	0	(32.5 33.3)	73	0	(35.0 30.7)
20	0	(36.2 36.1)	47	1	(41.1 41.8)	74	0	(32.2 36.7)
21	1	(35.0 37.4)	48	1	(40.8 38.6)	75	0	(34.8 39.0)
22	1	(40.3 43.4)	49	0	(28.4 35.3)	76	1	(46.8 40.8)
23	0	(34.6 34.1)	50	0	(42.0 45.1)	77	1	(38.4 41.8)
24	0	(40.5 49.1)	51	0	(32.1 26.5)	78	1	(33.2 38.0)
25	0	(34.0 49.9)	52	0	(35.6 34.7)	79	1	(36.3 35.7)
26	0	(39.7 34.7)	53	0	(31.8 36.9)	80	0	(37.3 38.8)
27	0	(35.3 42.6)	54	0	(36.2 45.0)			

Chapter 5

Classical Classification Procedure Based On k dichotomous and one Continuous Variables

5.1 Introduction

The robustness of the linear classification procedure to nonnormality has been examined by several authors. Subrahmaniam and Ching'anda (1978) and Ching'anda and Subrahmaniam (1979) derived the distributional properties of the misclassification probabilities while sampling from a number of non-normal populations. These authors used a modification of the developments of John (1960a, 1960b) in the normal case. Amoh (1983) developed some similar results for the inverse Gaussian distribution. Balakrishnan and Kocehrlakota (1985) examined the case of sampling from a mixture of two normal populations.

Balakrishnan et al. (1986) derived the distributions and the expected values of the errors of misclassification in Chang and Afifi's (1974) classification procedure based on dichotmous and normal variables. Kocherlakota et al. (1987) presented a unified development with examples of the expansion applicable for the asymptotic distribution of the errors of misclassification in many non-normal situations. Through their work on the outlier- normal model, they illustrated that the errors

of misclassification get inflated when few outliers are present in the sample. Balakrishnan et al. (1988) studied the robustness of Chang and Afifi's classification procedure through the distribution functions and expected values of the errors of misclassification under truncated normal and mixture normal models.

In this chapter, we consider the classification problem based on $k-$variate dichotomous and an univariate normal variable. This situation will arise naturally in many practical problems. The following example illustrates on such problems: Two life insurance companies offer similar life insurance policies. It is a widely known fact that smokers and male policy holders pay somewhat higher insurance premiums. Each insurance company keeps a record per policy holder consisting of gender (1 for male, 0 for female) marital status (1 for smoking, 0 for non-smoking), etc., and also premium. Now, given a sample of records from each company, one may be interested in classifying a new record into one of the two companies. Yet another example in an anthropometric set-up is as follows: Some physical measurements along with information regarding the gender and age-group (say, adult or non-adult) are available from two distinct groups of individuals (groups may be different races or castes, etc.). Based on this data, one may be interested in classifying a new individual into one of these two groups.

In Section 2, we describe the basic model of the classification problem based on $k-$variate dichotomous and an univariate continuous variable. In Section 3, we derive the classification procedure under the normal distribution first for the case when all the parameters of the model are known and next for the case when they are all unknown. We concentrate on the case when both the errors of misclassification are allowed to float (that is, neither one is fixed), and derive the exact distribution functions, density functions and the expected values of the two errors of misclassification. In Section 4, we consider the outlier-normal model for the data on the continuous variable and derive the exact distribution functions, density functions and the expected values of the two errors of misclassification of the normal procedure when both errors are allowed to float. In Section 5, we carry out a similar work by assuming a general mixture-normal model for the continuous variable. Finally,

in Section 6 by making use of the results derived in Section 3-5 we assess the robustness features of the normal classical classification procedure (when both errors are allowed to float) under the outlier-normal and the mixture- normal models. For this robustness study, we consider the case when the data contains two dichotomous and one continuous variables.

5.2 The Model

Let us consider two random variables $\mathbf{X}$ and Y where $\mathbf{X}$ is a $k-$variate dichotomous variable and Y is a univariate continuous variable. As Krzanowski (1975) suggested, the k dichotomous variables may be expressed as multinomial $\mathbf{z}' = (z_1, z_2, \ldots, z_q)$, where $q = 2^k$. This follows from the fact that each vector $(x_1, x_2, \ldots, x_k)$ may be considered as a binary representation of base 10 integers. Thus, each distinct pattern of $\mathbf{x}$ defines a multinomial cell uniquely, with $\mathbf{x}' = (x_1, x_2, \ldots, x_k)$ falling in cell $c = \sum_{i=1}^{k} x_i 2^{i-1}$. From this it is clear that a $k-$ variate dichotomous and multinomial or polytomous variables have a one-to-one correspondence. However, in this chapter we will not treat $k-$ variate dichotmous as multinomial variables. Instead we write the probability mass function of $\mathbf{X}$ as

$$P\{\mathbf{X} = \mathbf{x}\} = p_{\mathbf{x}}$$

where each element of $\mathbf{x}$ takes 0 or 1. Following Olkin and Tate (1961), it is assumed that Y has a univariate normal distribution with mean $\mu_{\mathbf{x}}$ and variance $\sigma_{\mathbf{x}}^2$ for given vector $\mathbf{X}$. We assume that

$$\mu_{\mathbf{x}} = \mu + \delta^T \mathbf{x}, \qquad \sigma_{\mathbf{x}}^2 = \sigma^2 + \gamma^T \mathbf{x},$$

where δ and γ are both constant $k \times 1$ column vectors. Then the vector $\mathbf{W} = (\mathbf{X}, Y)^T$ has its probability density function to be

$$f(\mathbf{w}) = p_{\mathbf{x}} \left(\frac{1}{\sqrt{2\pi}\sigma_{\mathbf{x}}} \exp - \frac{(y - \mu_{\mathbf{x}})^2}{2\sigma_{\mathbf{x}}^2} \right), \qquad x_i = 0, 1; \quad \text{for } i = 1, 2, \ldots, k; \quad -\infty < y < \infty. \tag{5.1}$$

It can be seen directly that

$$E(\mathbf{W}) = \begin{pmatrix} \mathbf{P} \\ \mu + \delta^T \mathbf{P} \end{pmatrix} \tag{5.2}$$

where $\mathbf{P}$ is a $k \times 1$ vector with i^{th} element as $Pr[X_i = 1]$. i.e. marginal probability of $X_i = 1$. The marginal probability density function of Y can be determined from (5.1) to be

$$g(y) = \sum_{\mathbf{x}} p_{\mathbf{x}} \left(\frac{1}{\sqrt{2\pi}\sigma_{\mathbf{x}}} \exp\left\{ -\frac{(y-\mu_{\mathbf{x}})^2}{2\sigma_{\mathbf{x}}^2} \right\} \right) \tag{5.3}$$

where $\sum_{\mathbf{x}}$ denotes the summation over all 2^k possible values of $\mathbf{x}$. The problem we examine here is concerned with the classification of an observation $\mathbf{w}_0$ into one of the two populations Π_1 and Π_2 where Π_i has pdf of the form

$$f_i(\mathbf{w}) = p_{\mathbf{x}}^{(i)} \phi(y; \mu_{\mathbf{x}}^{(i)}, \sigma_{\mathbf{x}}^2), \tag{5.4}$$

where $p_{\mathbf{x}}^{(i)}$ is the probability of obtaining an observation in the vector $\mathbf{x}$ for population Π_i, $(i = 1, 2)$. Also, $\mu_{\mathbf{x}}^{(i)}$ is the mean for a given vector $\mathbf{X}$ and population Π_i $(i = 1, 2)$. It should be noted that we are assuming here that the conditional variances of Y under Π_1 and Π_2 are equal. However, $Var(\mathbf{W}|\Pi_1)$ and $Var(\mathbf{W}|\Pi_2)$ are not equal.

5.3 Results For The Normal Case

In this section, under the assumption that the conditional distribution of Y is univariate normal (as already explained in Section 2) we first derive the likelihood-ratio rule for classification when all the parameters are known and derive its errors of misclassification. Next, we describe the classification procedure when all the parameters in the model are unknown and by considering the case when both the errors are allowed to float we derive the exact distribution functions, density functions and expected values of the errors of misclassification. As mentioned already in Section 1, these results generalize the work by Balakrishnan et al. (1986).

5.3.1 Classification when all parameters are known

The log-likelihood ratio gives the following procedure based on the statistic $Z(\mathrm{w})$ for the observation w_0, where

$$\begin{aligned} Z(\mathrm{w}) &= log\left[\frac{f_1(\mathrm{w})}{f_2(\mathrm{w})}\right] \\ &= log\left[\frac{p_{\mathrm{x}}^{(1)}}{p_{\mathrm{x}}^{(2)}}\exp\left\{-\frac{y-\mu_{\mathrm{x}}^{(1)}}{2\sigma_{\mathrm{x}}^2}+\frac{y-\mu_{\mathrm{x}}^{(2)}}{2\sigma_{\mathrm{x}}^2}\right\}\right] \\ &= \left\{y-\frac{\mu_{\mathrm{x}}^{(1)}+\mu_{\mathrm{x}}^{(2)}}{2}\right\}\frac{\mu_{\mathrm{x}}^{(1)}-\mu_{\mathrm{x}}^{(2)}}{\sigma_{\mathrm{x}}^2}+\ln\left(\frac{p_{\mathrm{x}}^{(1)}}{p_{\mathrm{x}}^{(2)}}\right) \end{aligned} \tag{5.5}$$

Classify w_0 as belonging to Π_1 or Π_2 according to

$$Z(\mathrm{w}_0) \gtrless C, \tag{5.6}$$

where C is determined by fixing one of the two errors of misclassification.

5.3.2 Errors of misclassification

There are two types of errors of misclassification with any classification procedure. There errors are

$$e_{12} = Pr[\text{misclassifying } \mathrm{w}_0 \text{ into } \Pi_2 \text{ when in fact } \mathrm{w}_0 \text{ is from } \Pi_1]$$

and

$$e_{21} = Pr[\text{misclassifying } \mathrm{w}_0 \text{ into } \Pi_1 \text{ when in fact } \mathrm{w}_0 \text{ is from } \Pi_2].$$

We first of all have

$$\begin{aligned} e_{12} &= \Pr[Z(\mathrm{w}) < C|\mathrm{w} \in \Pi_1] \\ &= \sum_{\mathrm{x}} p_{\mathrm{x}}^{(1)} Pr\left[\left\{y-\frac{\mu_{\mathrm{x}}^{(1)}+\mu_{\mathrm{x}}^{(2)}}{2}\right\}\frac{\mu_{\mathrm{x}}^{(1)}-\mu_{\mathrm{x}}^{(2)}}{\sigma_{\mathrm{x}}^2}+\ln\left(\frac{p_{\mathrm{x}}^{(1)}}{p_{\mathrm{x}}^{(2)}}\right) < C|y \sim N(\mu_{\mathrm{x}}^{(1)},\sigma_{\mathrm{x}}^2)\right] \end{aligned}$$

Algebraic manipulation yields

$$\begin{aligned} e_{12} &= \sum_{\mathrm{x}} p_{\mathrm{x}}^{(1)} Pr\left[\frac{y-\mu_{\mathrm{x}}^{(1)}}{\sigma_{\mathrm{x}}} < \frac{C\sigma_{\mathrm{x}}}{\mu_{\mathrm{x}}^{(1)}-\mu_{\mathrm{x}}^{(2)}} - \frac{\mu_{\mathrm{x}}^{(1)}-\mu_{\mathrm{x}}^{(2)}}{2\sigma_{\mathrm{x}}} - \right. \\ &\quad \left. \frac{\sigma_{\mathrm{x}}}{\mu_{\mathrm{x}}^{(1)}-\mu_{\mathrm{x}}^{(2)}}\ln\left(\frac{p_{\mathrm{x}}^{(1)}}{p_{\mathrm{x}}^{(2)}}\right)|y \sim N(\mu_{\mathrm{x}}^{(1)},\sigma_{\mathrm{x}}^2)\right] \end{aligned} \tag{5.7}$$

Let us denote $\frac{\mu_{\mathbf{x}}^{(1)}-\mu_{\mathbf{x}}^{(2)}}{\sigma_{\mathbf{x}}}$ by $\delta_{\mathbf{x}}$, then

$$e_{12} = \sum_{\mathbf{x}} p_{\mathbf{x}}^{(1)} \Phi \left\{ \frac{C - \frac{1}{2}\delta_{\mathbf{x}}^2 - \ln\left(\frac{p_{\mathbf{x}}^{(1)}}{p_{\mathbf{x}}^{(2)}}\right)}{|\delta_{\mathbf{x}}|} \right\}, \tag{5.8}$$

where $\Phi(\ )$ denotes the distribution function of the standard normal distribution. Similarly,

$$\begin{aligned} e_{21} &= \Pr[Z(\mathbf{w}) > C | \mathbf{w} \in \Pi_2] \\ &= \sum_{\mathbf{x}} p_{\mathbf{x}}^{(2)} Pr \left[\left\{ y - \frac{\mu_{\mathbf{x}}^{(1)} + \mu_{\mathbf{x}}^{(2)}}{2} \right\} \frac{\mu_{\mathbf{x}}^{(1)} - \mu_{\mathbf{x}}^{(2)}}{\sigma_{\mathbf{x}}^2} + \ln\left(\frac{p_{\mathbf{x}}^{(1)}}{p_{\mathbf{x}}^{(2)}}\right) > C | y \sim N(\mu_{\mathbf{x}}^{(2)}, \sigma_{\mathbf{x}}^2) \right]. \end{aligned}$$

Algebraic manipulation yields

$$\begin{aligned} e_{12} &= \sum_{\mathbf{x}} p_{\mathbf{x}}^{(2)} Pr \left[\frac{y - \mu_{\mathbf{x}}^{(2)}}{\sigma_{\mathbf{x}}} > \frac{C\sigma_{\mathbf{x}}}{\mu_{\mathbf{x}}^{(1)} - \mu_{\mathbf{x}}^{(2)}} + \frac{\mu_{\mathbf{x}}^{(1)} - \mu_{\mathbf{x}}^{(2)}}{2\sigma_{\mathbf{x}}} - \right. \\ & \left. \frac{\sigma_{\mathbf{x}}}{\mu_{\mathbf{x}}^{(1)} - \mu_{\mathbf{x}}^{(2)}} \ln\left(\frac{p_{\mathbf{x}}^{(1)}}{p_{\mathbf{x}}^{(2)}}\right) | y \sim N(\mu_{\mathbf{x}}^{(2)}, \sigma_{\mathbf{x}}^2) \right] \end{aligned} \tag{5.9}$$

which yields

$$e_{12} = \sum_{\mathbf{x}} p_{\mathbf{x}}^{(2)} \Phi \left\{ -\frac{C + \frac{1}{2}\delta_{\mathbf{x}}^2 - \ln\left(\frac{p_{\mathbf{x}}^{(1)}}{p_{\mathbf{x}}^{(2)}}\right)}{|\delta_{\mathbf{x}}|} \right\}. \tag{5.10}$$

5.3.3 Classification when the parameters are unknown

Consider independent random samples of sizes n_1 and n_2 from each of the populations. Denote by $y_{\mathbf{x},j}^{(i)}$ the jth observation from the ith population of which the value of $\mathbf{X}$ is $\mathbf{x}$. Let $n_i = \sum_{\mathbf{x}} n_{\mathbf{x}}^{(i)}$, where $n_{\mathbf{x}}^{(i)}$ is the number of observations in the ith sample with $\mathbf{X} = \mathbf{x}$, and

$$\bar{y}_{\mathbf{x}}^{(i)} = \frac{1}{n_{\mathbf{x}}^{(i)}} \sum_{j=1}^{n_{\mathbf{x}}^{(i)}} y_{\mathbf{x},j}^{(i)}$$

and

$$s_{\mathbf{x}}^2 = \frac{1}{n_{\mathbf{x}}^{(1)} + n_{\mathbf{x}}^{(2)} - 2} \sum_{i=1}^{2} \sum_{j=1}^{n_{\mathbf{x}}^{(i)}} (y_{\mathbf{x},j}^{(i)} - \bar{y}_{\mathbf{x}}^{(i)})^2.$$

It can be shown that the MLE's of the parameters are, for $i = 1, 2$,

$$\hat{p}_{\mathbf{x}}^{(i)} = \frac{n_{\mathbf{x}}^{(i)}}{n_i}, \qquad \hat{\mu}_{\mathbf{x}}^{(i)} = \bar{y}_{\mathbf{x}}^{(i)}, \qquad \hat{\sigma}_{\mathbf{x}}^2 = s_{\mathbf{x}}^2 \ \text{(corrected for bias)}.$$

The likelihood classification rule is in this case to classify an observation w as belonging to Π_1 or Π_2 according as

$$Z(\mathbf{w}) = \left\{ y - \frac{\bar{y}_{\mathbf{x}}^{(1)} + \bar{y}_{\mathbf{x}}^{(2)}}{2} \right\} \frac{\bar{y}_{\mathbf{x}}^{(1)} - \bar{y}_{\mathbf{x}}^{(2)}}{s_{\mathbf{x}}^2} + \ln \left(\frac{n_{\mathbf{x}}^{(1)} n_2}{n_{\mathbf{x}}^{(2)} n_1} \right) \gtrless C,$$

where C is determined by fixing one of the two errors of misclassifications. For this one has to make use of the asymptotic distributional properties of the statistic $Z(\mathbf{w})$.

5.3.4 Classification when both the errors are allowed to float

In certain instances, however, it may not be of interest to restrict e_{12} at a pre-fixed level; that is, one may be interested in allowing both e_{12} and e_{21} to float. This will be the case when committing one type of error is not considered to be any more serious than committing the other type of error.In this case, the classification procedure is to classify the independent observation $\mathbf{w}_0$ into Π_1, Π_2, or arbitrarily into Π_1 or Π_2, according as

$$Z(\mathbf{w}_0) \gtrless = 0.$$

For simplicity, in order to discuss the errors of misclassification we shall assume that $p_{\mathbf{x}}^{(1)} = p_{\mathbf{x}}^{(2)}$. In this case, the last term in (5.5) is zero. Therefore the classification statistic reduces to

$$Z^*(\mathbf{w}) = \left\{ y - \frac{\bar{y}_{\mathbf{x}}^{(1)} + \bar{y}_{\mathbf{x}}^{(2)}}{2} \right\} \left(\bar{y}_{\mathbf{x}}^{(1)} - \bar{y}_{\mathbf{x}}^{(2)} \right).$$

By defining two random variables $U_{\mathbf{x}}$ and $V_{\mathbf{x}}$ such that

$$U_{\mathbf{x}} = \bar{Y}_{\mathbf{x}}^{(1)} + \bar{Y}_{\mathbf{x}}^{(2)},$$

and

$$V_{\mathbf{x}} = \bar{Y}_{\mathbf{x}}^{(2)} - \bar{Y}_{\mathbf{x}}^{(1)},$$

and letting $u_{\mathbf{x}}$ and $v_{\mathbf{x}}$ be their respective realizations, we have

$$Z^*(\mathbf{w}) = -\left\{y - \frac{u_{\mathbf{x}}}{2}\right\} v_{\mathbf{x}}. \tag{5.11}$$

Therefore

$$\begin{aligned} Z^*(\mathbf{w}) \ > 0 \ &\Rightarrow \begin{cases} y > \frac{u_{\mathbf{x}}}{2} & \text{for} \quad v_{\mathbf{x}} < 0 \\ y < \frac{u_{\mathbf{x}}}{2} & \text{for} \quad v_{\mathbf{x}} > 0 \end{cases} \\ \text{and} \\ Z^*(\mathbf{w}) \ < 0 \ &\Rightarrow \begin{cases} y < \frac{u_{\mathbf{x}}}{2} & \text{for} \quad v_{\mathbf{x}} < 0 \\ y > \frac{u_{\mathbf{x}}}{2} & \text{for} \quad v_{\mathbf{x}} > 0 \end{cases} \end{aligned} \tag{5.12}$$

The simplified classification rule is to classify $\mathbf{w} = \begin{bmatrix} x \\ y \end{bmatrix}$ as

1. Belonging to Π_1, if

$$Z^*(\mathbf{w}) > 0 \Rightarrow \begin{cases} y > \frac{u_{\mathbf{x}}}{2} & \text{for} \quad v_{\mathbf{x}} < 0 \\ y < \frac{u_{\mathbf{x}}}{2} & \text{for} \quad v_{\mathbf{x}} > 0 \end{cases}$$

2. Belonging to Π_2, if

$$Z^*(\mathbf{w}) < 0 \Rightarrow \begin{cases} y < \frac{u_{\mathbf{x}}}{2} & \text{for} \quad v_{\mathbf{x}} < 0 \\ y > \frac{u_{\mathbf{x}}}{2} & \text{for} \quad v_{\mathbf{x}} > 0 \end{cases}.$$

In this special case, we derive the exact distribution functions, density functions and expected values of the errors of misclassification in the following.

5.3.5 Exact distribution functions of the two errors of misclassification

The joint distribution of $(U_{\mathbf{x}}, V_{\mathbf{x}})$ is $BVN[\mu_{u\mathbf{x}}, \mu_{v\mathbf{x}}; \sigma^2_{u\mathbf{x}}, \sigma^2_{v\mathbf{x}}, \rho_{\mathbf{x}}]$ where

$$\mu_{u\mathbf{x}} \ = \ \mu_{\mathbf{x}}^{(1)} + \mu_{\mathbf{x}}^{(2)},$$

$$\begin{aligned}
\mu_{vx} &= \mu_x^{(2)} - \mu_x^{(1)}, \\
\sigma_{ux}^2 = \sigma_{vx}^2 &= \sigma_x^2 \left(\frac{1}{n_x^{(1)}} + \frac{1}{n_x^{(2)}} \right) \\
\rho_x &= \left(\frac{n_x^{(1)} - n_x^{(2)}}{n_x^{(1)} + n_x^{(2)}} \right).
\end{aligned}$$

By using U_x and V_x we can write the error of misclassification as

$$e_{12} = \begin{cases} \Phi\left[\frac{\frac{U_x}{2} - \mu_x^{(1)}}{\sigma_x}\right] & \text{if} \quad V_x < 0 \\ \\ \Phi\left[\frac{\mu_x^{(1)} - \frac{U_x}{2}}{\sigma_x}\right] & \text{if} \quad V_x > 0 \end{cases} \tag{5.13}$$

Let

$$G_1(z) = \Pr[e_{12} \leq z],$$

then by substituting the expression for e_{12} from (5.13) we obtain

$$\begin{aligned}
G_1(z) &= \sum_x p_x^{(1)} \left\{ \Pr\left[U_x \leq k_x^{(1)}, V_x < 0\right] + \right. \\
&\quad \left. \Pr\left[U_x > k_x^{(2)}, V_x > 0\right] \right\},
\end{aligned} \tag{5.14}$$

where

$$\begin{aligned}
k_x^{(1)} &= 2\{\mu_x^{(1)} + \sigma_x \Phi^{-1}(z)\}, \\
k_x^{(2)} &= 2\{\mu_x^{(1)} - \sigma_x \Phi^{-1}(z)\}.
\end{aligned}$$

By denoting $H(z_1, z_2; \rho)$ as the distribution function of the standard BVN distribution with the coefficient of correlation ρ, we can write the cdf of e_{12}

$$G_1(z) = \sum_x p_x^{(1)} \left[H\left(\frac{k_x^{(1)} - \mu_{ux}}{\sigma_{ux}}, -\frac{\mu_{vx}}{\sigma_{vx}}; \rho_x \right) + H\left(\frac{\mu_{ux} - k_x^{(2)}}{\sigma_{ux}}, \frac{\mu_{vx}}{\sigma_{vx}}; \rho_x \right) \right]. \tag{5.15}$$

with

$$\delta_x = \frac{\mu_x^{(1)} - \mu_x^{(2)}}{\sigma_x}$$

we have

$$\begin{aligned}
\frac{k_x^{(1)} - \mu_{ux}}{\sigma_{ux}} &= \frac{2\mu_x^{(1)} + 2\sigma_x \Phi^{-1}(z) - \mu_x^{(1)} - \mu_x^{(2)}}{\sigma_x m_x} \\
&= \frac{\delta_x + 2\Phi^{-1}(z)}{m_x},
\end{aligned} \tag{5.16}$$

$$\begin{aligned}\frac{\mu_{ux} - k_x^{(2)}}{\sigma_{ux}} &= \frac{\mu_x^{(1)} + \mu_x^{(2)} - 2\mu_x^{(1)} + 2\sigma_x \Phi^{-1}(z)}{m_x \sigma_x} \\ &= \frac{-\delta_x + 2\Phi^{-1}(z)}{m_x}, \end{aligned} \tag{5.17}$$

and

$$\frac{\mu_{vx}}{\sigma_{vx}} = -\frac{\delta_x}{m_x}, \tag{5.18}$$

where

$$m_x = \left[\frac{1}{n_x^{(1)}} + \frac{1}{n_x^{(2)}}\right]^{1/2}.$$

Therefore

$$G_1(z) = \sum_x p_x^{(1)} \left[H\left(\frac{\delta_x + 2\Phi^{-1}(z)}{m_x}, \frac{\delta_x}{m_x}; \rho_x\right) + H\left(\frac{-\delta_x + 2\Phi^{-1}(z)}{m_x}, -\frac{\delta_x}{m_x}; \rho_x\right)\right] \tag{5.19}$$

The density function of e_{12} can be obtained by directly differentiating the distribution function. For this differentiation we use **Result 2** in Appendix A and the fact that

$$\frac{d\Phi^{-1}(z)}{dz} = \frac{1}{\phi(\Phi^{-1}(z))}. \tag{5.20}$$

Thus

$$\begin{aligned} g_1(z) &= 2\sum_x p_x^{(1)} \frac{1}{m_x \phi(\Phi^{-1}(z))} \left[\Phi\left\{C_x^{(1)}\right\} \phi\left(\frac{\delta_x + 2\Phi^{-1}(z)}{m_x}\right)\right. \\ &\quad \left. + \Phi\left\{C_x^{(2)}\right\} \phi\left(\frac{-\delta_x + 2\Phi^{-1}(z)}{m_x}\right)\right] \end{aligned} \tag{5.21}$$

where

$$C_x^{(1)} = \left[\frac{(1-\rho_x)\delta_x - 2\rho_x \Phi^{-1}(z)}{m_x (1-\rho_x^2)^{1/2}}\right]$$

and

$$C_x^{(2)} = -\left[\frac{(1-\rho_x)\delta_x + 2\rho_x \Phi^{-1}(z)}{m_x (1-\rho_x^2)^{1/2}}\right].$$

We may similarly derive the distribution function and density function of e_{21} given by

$$e_{21} = \begin{cases} \Phi\left[\frac{\frac{U_x}{2} - \mu_x^{(1)}}{\sigma_x}\right] & \text{if} \quad V_x > 0 \\ \Phi\left[\frac{\mu_x^{(1)} - \frac{U_x}{2}}{\sigma_x}\right] & \text{if} \quad V_x < 0 \end{cases} \tag{5.22}$$

Let

$$G_2(z) = \Pr[e_{21} \leq z],$$

then by substituting the expression for e_{21} from (5.22) we obtain

$$\begin{aligned} G_2(z) &= \sum_{\mathbf{x}} p_{\mathbf{x}}^{(2)} \left\{ \Pr\left[U_{\mathbf{x}} \leq l_{\mathbf{x}}^{(1)}, V_{\mathbf{x}} > 0\right] + \right. \\ &\quad \left. \Pr\left[U_{\mathbf{x}} > l_{\mathbf{x}}^{(2)}, V_{\mathbf{x}} < 0\right] \right\}, \end{aligned} \tag{5.23}$$

where

$$\begin{aligned} l_{\mathbf{x}}^{(1)} &= 2\{\mu_{\mathbf{x}}^{(2)} + \sigma_{\mathbf{x}}\Phi^{-1}(z)\}, \\ l_{\mathbf{x}}^{(2)} &= 2\{\mu_{\mathbf{x}}^{(2)} - \sigma_{\mathbf{x}}\Phi^{-1}(z)\}. \end{aligned}$$

The cdf of e_{21} then becomes

$$G_2(z) = \sum_{\mathbf{x}} p_{\mathbf{x}}^{(2)} \left[H\left(\frac{l_{\mathbf{x}}^{(1)} - \mu_{u\mathbf{x}}}{\sigma_{u\mathbf{x}}}, \frac{\mu_{v\mathbf{x}}}{\sigma_{v\mathbf{x}}}; -\rho_{\mathbf{x}}\right) + H\left(\frac{\mu_{u\mathbf{x}} - l_{\mathbf{x}}^{(2)}}{\sigma_{u\mathbf{x}}}, -\frac{\mu_{v\mathbf{x}}}{\sigma_{v\mathbf{x}}}; -\rho_{\mathbf{x}}\right)\right]. \tag{5.24}$$

Now we have

$$\begin{aligned} \frac{l_{\mathbf{x}}^{(1)} - \mu_{u\mathbf{x}}}{\sigma_{u\mathbf{x}}} &= \frac{2\mu_{\mathbf{x}}^{(2)} + 2\sigma_{\mathbf{x}}\Phi^{-1}(z) - \mu_{\mathbf{x}}^{(1)} - \mu_{\mathbf{x}}^{(2)}}{m_{\mathbf{x}}\sigma_{\mathbf{x}}} \\ &= \frac{-\delta_{\mathbf{x}} + 2\Phi^{-1}(z)}{m_{\mathbf{x}}}, \end{aligned} \tag{5.25}$$

and

$$\begin{aligned} \frac{\mu_{u\mathbf{x}} - l_{\mathbf{x}}^{(2)}}{\sigma_{u\mathbf{x}}} &= \frac{\mu_{\mathbf{x}}^{(1)} + \mu_{\mathbf{x}}^{(2)} + 2\sigma_{\mathbf{x}}\Phi^{-1}(z) - 2\mu_{\mathbf{x}}^{(1)}}{m_{\mathbf{x}}\sigma_{\mathbf{x}}} \\ &= \frac{\delta_{\mathbf{x}} + 2\Phi^{-1}(z)}{m_{\mathbf{x}}}. \end{aligned} \tag{5.26}$$

Therefore

$$G_2(z) = \sum_{\mathbf{x}} p_{\mathbf{x}}^{(2)} \left[H\left(\frac{-\delta_{\mathbf{x}} + 2\Phi^{-1}(z)}{m_{\mathbf{x}}}, \frac{-\delta_{\mathbf{x}}}{m_{\mathbf{x}}}, -\rho_{\mathbf{x}}\right) + H\left(\frac{\delta_{\mathbf{x}} + 2\Phi^{-1}(z)}{m_{\mathbf{x}}}, \frac{\delta_{\mathbf{x}}}{m_{\mathbf{x}}}, -\rho_{\mathbf{x}}\right)\right] \tag{5.27}$$

The density function of e_{21} can be obtained by directly differentiating the distribution function. Thus

$$\begin{aligned} g_2(z) &= 2\sum_{\mathbf{x}} p_{\mathbf{x}}^{(2)} \frac{1}{m_{\mathbf{x}}\phi\left(\Phi^{-1}(z)\right)} \left[\Phi\left\{C_{\mathbf{x}}^{*(1)}\right\}\phi\left(\frac{-\delta_{\mathbf{x}} + 2\Phi^{-1}(z)}{m_{\mathbf{x}}}\right)\right. \\ &\quad \left. + \Phi\left\{C_{\mathbf{x}}^{*(2)}\right\}\phi\left(\frac{\delta_{\mathbf{x}} + 2\Phi^{-1}(z)}{m_{\mathbf{x}}}\right)\right] \end{aligned} \tag{5.28}$$

where

$$C_{\mathbf{x}}^{*(1)} = \left[\frac{-(1+\rho_{\mathbf{x}})\delta_{\mathbf{x}} + 2\rho_{\mathbf{x}}\Phi^{-1}(z)}{m_{\mathbf{x}}(1-\rho_{\mathbf{x}}^2)^{1/2}}\right]$$

and

$$C_{\mathbf{x}}^{*(2)} = \left[\frac{(1+\rho_{\mathbf{x}})\delta_{\mathbf{x}} + 2\rho_{\mathbf{x}}\Phi^{-1}(z)}{m_{\mathbf{x}}(1-\rho_{\mathbf{x}}^2)^{1/2}}\right].$$

Alternatively one can obtain the cdf and pdf of e_{21} by interchanging sample 1 and sample 2, i.e. reversing the sign of $\delta_{\mathbf{x}}$ and $\rho_{\mathbf{x}}$, in the expressions for cdf and pdf of e_{12} derived in Eqs. (5.19) and (5.21), respectively.

5.3.6 Expected values of the error rates

Since

$$e_{12} = \begin{cases} Pr\left[2Y > \bar{y}_{\mathbf{x}}^{(1)} + \bar{y}_{\mathbf{x}}^{(2)} | \bar{y}_{\mathbf{x}}^{(1)}, \bar{y}_{\mathbf{x}}^{(2)}\right] & \text{if} \quad \bar{y}_{\mathbf{x}}^{(1)} < \bar{y}_{\mathbf{x}}^{(2)} \\ Pr\left[2Y < \bar{y}_{\mathbf{x}}^{(1)} + \bar{y}_{\mathbf{x}}^{(2)} | \bar{y}_{\mathbf{x}}^{(1)}, \bar{y}_{\mathbf{x}}^{(2)}\right] & \text{if} \quad \bar{y}_{\mathbf{x}}^{(1)} > \bar{y}_{\mathbf{x}}^{(2)} \end{cases} \tag{5.29}$$

the expected value of e_{12} (also called the unconditional probability of misclassification) may be computed as

$$\begin{aligned} E(e_{12}) &= \sum_{\mathbf{x}} p_{\mathbf{x}}^{(1)} \left\{ Pr\left[2Y > \bar{y}_{\mathbf{x}}^{(1)} + \bar{y}_{\mathbf{x}}^{(2)}, \bar{y}_{\mathbf{x}}^{(1)} < \bar{y}_{\mathbf{x}}^{(2)} | Y \in \Pi_1\right] + \right. \\ &\quad \left. Pr\left[2Y < \bar{y}_{\mathbf{x}}^{(1)} + \bar{y}_{\mathbf{x}}^{(2)}, \bar{y}_{\mathbf{x}}^{(1)} > \bar{y}_{\mathbf{x}}^{(2)} | Y \in \Pi_1\right]\right\} \end{aligned} \tag{5.30}$$

or

$$\begin{aligned} E(e_{12}) &= \sum_{\mathbf{x}} p_{\mathbf{x}}^{(1)} \{ Pr\left[U_{\mathbf{x}} < 2Y, V_{\mathbf{x}} > 0 | Y \sim N(\mu_{\mathbf{x}}^{(1)}\sigma_{\mathbf{x}}^2)\right] + \\ &\quad Pr\left[U_{\mathbf{x}} > 2Y, V_{\mathbf{x}} < 0 | Y \sim N(\mu_{\mathbf{x}}^{(1)}\sigma_{\mathbf{x}}^2)\right] \} \end{aligned} \tag{5.31}$$

We can simplify (5.31) by using **Result 3** in Appendix A and obtain the following expression

$$E(e_{12}) = \sum_{\mathbf{x}} p_{\mathbf{x}}^{(1)} \left[H\left(\frac{2\mu_{\mathbf{x}}^{(1)} - \mu_{u\mathbf{x}}}{(4\sigma_{\mathbf{x}}^2 + \sigma_{u\mathbf{x}}^2)^{1/2}}, \frac{\mu_{v\mathbf{x}}}{\sigma_{v\mathbf{x}}}; -\frac{\sigma_{u\mathbf{x}}\rho_{\mathbf{x}}}{(4\sigma_{\mathbf{x}}^2 + \sigma_{u\mathbf{x}}^2)^{1/2}} \right) + H\left(-\frac{2\mu_{\mathbf{x}}^{(1)} - \mu_{u\mathbf{x}}}{(4\sigma_{\mathbf{x}}^2 + \sigma_{u\mathbf{x}}^2)^{1/2}}, -\frac{\mu_{v\mathbf{x}}}{\sigma_{v\mathbf{x}}}; -\frac{\sigma_{u\mathbf{x}}\rho_{\mathbf{x}}}{(4\sigma_{\mathbf{x}}^2 + \sigma_{u\mathbf{x}}^2)^{1/2}} \right) \right]. \tag{5.32}$$

But

$$\frac{2\mu_{\mathbf{x}}^{(1)} - \mu_{u\mathbf{x}}}{(4\sigma_{\mathbf{x}}^2 + \sigma_{u\mathbf{x}}^2)^{1/2}} = \frac{2\mu_{\mathbf{x}}^{(1)} - \mu_{\mathbf{x}}^{(1)} - \mu_{\mathbf{x}}^{(2)}}{(4\sigma_{\mathbf{x}}^2 + m_{\mathbf{x}}^2\sigma_{\mathbf{x}}^2)^{1/2}} = \frac{\delta_{\mathbf{x}}}{(4 + m_{\mathbf{x}}^2)^{1/2}}. \tag{5.33}$$

Therefore, we get

$$E(e_{12}) = \sum_{\mathbf{x}} p_{\mathbf{x}}^{(1)} \left[H\left(\frac{\delta_{\mathbf{x}}}{(4 + m_{\mathbf{x}}^2)^{1/2}}, -\frac{\delta_{\mathbf{x}}}{m_{\mathbf{x}}}; -\frac{m_{\mathbf{x}}\rho_{\mathbf{x}}}{(4 + m_{\mathbf{x}}^2)^{1/2}} \right) + H\left(-\frac{\delta_{\mathbf{x}}}{(4 + m_{\mathbf{x}}^2)^{1/2}}, \frac{\delta_{\mathbf{x}}}{m_{\mathbf{x}}}; -\frac{m_{\mathbf{x}}\rho_{\mathbf{x}}}{(4 + m_{\mathbf{x}}^2)^{1/2}} \right) \right]. \tag{5.34}$$

Similarly, we have

$$E(e_{21}) = \sum_{\mathbf{x}} p_{\mathbf{x}}^{(2)} \{ Pr\left[U_{\mathbf{x}} < 2Y, V_{\mathbf{x}} < 0 | Y \sim N(\mu_{\mathbf{x}}^{(2)}, \sigma_{\mathbf{x}}^2)\right] + Pr\left[U_{\mathbf{x}} > 2Y, V_{\mathbf{x}} > 0 | Y \sim N(\mu_{\mathbf{x}}^{(2)}, \sigma_{\mathbf{x}}^2)\right] \} \tag{5.35}$$

We can simplify (5.35) by using **Result 3** in Appendix A and obtain the following expression

$$E(e_{21}) = \sum_{\mathbf{x}} p_{\mathbf{x}}^{(2)} \left[H\left(\frac{2\mu_{\mathbf{x}}^{(2)} - \mu_{u\mathbf{x}}}{(4\sigma_{\mathbf{x}}^2 + \sigma_{u\mathbf{x}}^2)^{1/2}}, -\frac{\mu_{v\mathbf{x}}}{\sigma_{v\mathbf{x}}}; \frac{\sigma_{u\mathbf{x}}\rho_{\mathbf{x}}}{(4\sigma_{\mathbf{x}}^2 + \sigma_{u\mathbf{x}}^2)^{1/2}} \right) + H\left(-\frac{2\mu_{\mathbf{x}}^{(1)} - \mu_{u\mathbf{x}}}{(4\sigma_{\mathbf{x}}^2 + \sigma_{u\mathbf{x}}^2)^{1/2}}, \frac{\mu_{v\mathbf{x}}}{\sigma_{v\mathbf{x}}}; \frac{\sigma_{u\mathbf{x}}\rho_{\mathbf{x}}}{(4\sigma_{\mathbf{x}}^2 + \sigma_{u\mathbf{x}}^2)^{1/2}} \right) \right]. \tag{5.36}$$

But

$$\frac{2\mu_{\mathbf{x}}^{(2)} - \mu_{u\mathbf{x}}}{(4\sigma_{\mathbf{x}}^2 + \sigma_{u\mathbf{x}}^2)^{1/2}} = \frac{2\mu_{\mathbf{x}}^{(2)} - \mu_{\mathbf{x}}^{(1)} - \mu_{\mathbf{x}}^{(2)}}{(4\sigma_{\mathbf{x}}^2 + m_{\mathbf{x}}^2\sigma_{\mathbf{x}}^2)^{1/2}} \tag{5.37}$$

$$= -\frac{\delta_{\mathbf{x}}}{(4 + m_{\mathbf{x}}^2)^{1/2}}. \tag{5.38}$$

Therefore, we get

$$E(e_{21}) = \sum_{\mathbf{x}} p_{\mathbf{x}}^{(1)} \left[H\left(-\frac{\delta_{\mathbf{x}}}{(4+m_{\mathbf{x}}^2)^{1/2}}, \frac{\delta_{\mathbf{x}}}{m_{\mathbf{x}}}; \frac{m_{\mathbf{x}}\rho_{\mathbf{x}}}{(4+m_{\mathbf{x}}^2)^{1/2}} \right) + H\left(\frac{\delta_{\mathbf{x}}}{(4+m_{\mathbf{x}}^2)^{1/2}}, -\frac{\delta_{\mathbf{x}}}{m_{\mathbf{x}}}; \frac{m_{\mathbf{x}}\rho_{\mathbf{x}}}{(4+m_{\mathbf{x}}^2)^{1/2}} \right) \right]. \tag{5.39}$$

Alternatively one can obtain the expected value of e_{21} by interchanging sample 1 and sample 2, i.e. reversing the sign of $\delta_{\mathbf{x}}$ and $\rho_{\mathbf{x}}$, in the expression for the expected value of e_{12} derived in Eq. (5.34)

5.4 Results For The Outlier-normal case

Barnett and Lewis (1978) define an outlier as an observation (or subset of observations) in a set of data which appears to be inconsistent with the remainder of that set of data. As Barnett and Lewis suggested the phrase 'appears to be inconsistent' is crucial; it is a matter of subjective judgement on the part of the observer whether or not he/she picks out some observation (or set of observations) for scrutiny.

Several models for outliers have been proposed; see Barnett and Lewis (1978) who give a full discussion of these models. For various outlier testing procedures one can also refer to Tiku, Tan and Balakrishnan (1986).

The most popular model perhaps is the one proposed by Dixon (1950) as follows.

For a single general outlier in the sample, this model stipulates that in a random sample of size n, $n-1$ observations $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ come from a normal $N(\mu, \sigma^2)$, where μ and σ^2 are unknown, and the ith observation x_i comes from a normal $N(\mu + a\sigma, (b\sigma)^2)$, where i, a, b are unknown. In this section we use the above mentioned outlier model as the underlying distribution for the samples in hand.

In this section, by considering the general outlier-normal model introduced above for the data on the continuous variable in the classification problem (see Section 2) we derive the exact distribution functions, density functions and expected

values of the errors of misclassification of the procedure based on the normal distribution (described in Section 3) in the case when both the errors are allowed to float. This work extends the results of Kocherlakota et al. (1987) and will be used in Section 6 to assess the effect some outliers in the data have on the classification procedure based on the normal distribution.

5.4.1 Exact distribution functions of the two errors of misclassification

Let us consider sampling from normal populations under Π_i, $i = 1, 2$, with s_i outliers in each case. Then write $s_i = \sum_{\mathbf{x}} s_{\mathbf{x}}^{(i)}$, where $s_{\mathbf{x}}^{(i)}$ the number of observations in Π_i with $\mathbf{X} = \mathbf{x}$. That is, for $i = 1,\ 2$,

$$y_{\mathbf{x},j}^{(i)} \sim N(\mu_{\mathbf{x}}^{(i)}, \sigma_{\mathbf{x}}^2), \qquad j = 1, 2, \ldots, n_{\mathbf{x}}^{(i)} - s_{\mathbf{x}}^{(i)}$$

and

$$y_{\mathbf{x},j}^{(i)} \sim N(\mu_{\mathbf{x}}^{(i)} + a_i\sigma_{\mathbf{x}}, b_i^2\sigma_{\mathbf{x}}^2), \qquad j = n_{\mathbf{x}}^{(i)} - s_{\mathbf{x}}^{(i)} + 1, \ldots, n_{\mathbf{x}}^{(i)},$$

where a_i is the location contamination and b_i is the scale contamination. Then it can be seen that the sample mean

$$\bar{y}_{\mathbf{x}}^{(i)} \sim N\left(\mu_{\mathbf{x}}^{(i)} + q_{\mathbf{x}}^{(i)} a_i\sigma, \frac{\sigma_{\mathbf{x}}^2}{n_{\mathbf{x}}^{(i)}}\left[1 + q_{\mathbf{x}}^{(i)}(b_i^2 - 1)\right]\right),$$

independently, for $i = 1, 2$. Here $q_{\mathbf{x}}^{(i)} = \frac{s_{\mathbf{x}}^{(i)}}{n_{\mathbf{x}}^{(i)}}$, the proportion of outliers in the ith sample. In this case $U_{\mathbf{x}} = \bar{Y}_{\mathbf{x}}^{(1)} + \bar{Y}_{\mathbf{x}}^{(2)}$ and $V_{\mathbf{x}} = \bar{Y}_{\mathbf{x}}^{(2)} - \bar{Y}_{\mathbf{x}}^{(1)}$ are jointly $BVN[\mu_{u\mathbf{x}}, \mu_{v\mathbf{x}}; \sigma_{u\mathbf{x}}, \sigma_{v\mathbf{x}}, \rho_{\mathbf{x}}]$, where

$$\begin{aligned}
\mu_{u\mathbf{x}} &= \mu_{\mathbf{x}}^{(1)} + \mu_{\mathbf{x}}^{(2)} + \sigma_{\mathbf{x}}(q_{\mathbf{x}}^{(1)}a_1 + q_{\mathbf{x}}^{(2)}a_2), \\
\mu_{v\mathbf{x}} &= \mu_{\mathbf{x}}^{(2)} - \mu_{\mathbf{x}}^{(1)} + \sigma_{\mathbf{x}}(q_{\mathbf{x}}^{(2)}a_2 - q_{\mathbf{x}}^{(1)}a_1) \\
\sigma_{u\mathbf{x}}^2 = \sigma_{v\mathbf{x}}^2 &= \sigma_{\mathbf{x}}^2\left[\frac{1 + q_{\mathbf{x}}^{(1)}(b_1^2 - 1)}{n_{\mathbf{x}}^{(1)}} + \frac{1 + q_{\mathbf{x}}^{(2)}(b_2^2 - 1)}{n_{\mathbf{x}}^{(2)}}\right], \\
&\text{and} \\
\rho_{\mathbf{x}}\sigma_{u\mathbf{x}}^2 &= \sigma_{\mathbf{x}}^2\left[\frac{1 + q_{\mathbf{x}}^{(2)}(b_2^2 - 1)}{n_{\mathbf{x}}^{(2)}} - \frac{1 + q_{\mathbf{x}}^{(1)}(b_1^2 - 1)}{n_{\mathbf{x}}^{(1)}}\right].
\end{aligned}$$

For our convenience let us denote

$$\begin{aligned}
a_{\mathbf{x}}^{*} &= (q_{\mathbf{x}}^{(2)}a_2 + q_{\mathbf{x}}^{(1)}a_1), \\
b_{\mathbf{x}}^{*} &= (q_{\mathbf{x}}^{(2)}a_2 - q_{\mathbf{x}}^{(1)}a_1), \\
c_{\mathbf{x}}^{2} &= \left[\frac{1+q_{\mathbf{x}}^{(2)}(b_2^2-1)}{n_{\mathbf{x}}^{(2)}} + \frac{1+q_{\mathbf{x}}^{(1)}(b_1^2-1)}{n_{\mathbf{x}}^{(1)}}\right], \\
d_{\mathbf{x}}^{2} &= \left[\frac{1+q_{\mathbf{x}}^{(2)}(b_2^2-1)}{n_{\mathbf{x}}^{(2)}} - \frac{1+q_{\mathbf{x}}^{(1)}(b_1^2-1)}{n_{\mathbf{x}}^{(1)}}\right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mu_{u\mathbf{x}} &= \mu_{\mathbf{x}}^{(1)} + \mu_{\mathbf{x}}^{(2)} + a_{\mathbf{x}}^{*}\sigma_{\mathbf{x}}, \\
\mu_{u\mathbf{x}} &= \mu_{\mathbf{x}}^{(2)} - \mu_{\mathbf{x}}^{(1)} + b_{\mathbf{x}}^{*}\sigma_{\mathbf{x}}, \\
\sigma_{u\mathbf{x}} = \sigma_{v\mathbf{x}} &= c_{\mathbf{x}}\sigma_{\mathbf{x}}, \\
&\text{and} \\
\rho_{\mathbf{x}} &= \frac{d_{\mathbf{x}}^2}{c_{\mathbf{x}}^2}.
\end{aligned}$$

Hence by using (5.14) the distribution function of e_{12} becomes

$$G_1(z) = \sum_{\mathbf{x}} p_{\mathbf{x}}^{(1)}\left[H\left(\frac{k_{\mathbf{x}}^{(1)}-\mu_{u\mathbf{x}}}{\sigma_{u\mathbf{x}}}, -\frac{\mu_{v\mathbf{x}}}{\sigma_{v\mathbf{x}}}; \rho_{\mathbf{x}}\right) + H\left(\frac{\mu_{u\mathbf{x}}-k_{\mathbf{x}}^{(2)}}{\sigma_{u\mathbf{x}}}, \frac{\mu_{v\mathbf{x}}}{\sigma_{v\mathbf{x}}}; \rho_{\mathbf{x}}\right)\right] \tag{5.40}$$

But

$$\begin{aligned}
\frac{k_{\mathbf{x}}^{(1)}-\mu_{u\mathbf{x}}}{\sigma_{u\mathbf{x}}} &= \frac{2(\mu_{\mathbf{x}}^{(1)} + 2\sigma_{\mathbf{x}}\Phi^{-1}(z)) - \mu_{\mathbf{x}}^{(1)} - \mu_{\mathbf{x}}^{(2)} - a_{\mathbf{x}}^{*}\sigma_{\mathbf{x}}}{c_{\mathbf{x}}\sigma_{\mathbf{x}}} \\
&= \frac{\delta_{\mathbf{x}} + 2\Phi^{-1}(z) - a_{\mathbf{x}}^{*}}{c_{\mathbf{x}}},
\end{aligned} \tag{5.41}$$

$$\begin{aligned}
\frac{\mu_{u\mathbf{x}}-k_{\mathbf{x}}^{(2)}}{\sigma_{u\mathbf{x}}} &= \frac{\mu_{\mathbf{x}}^{(1)} + \mu_{\mathbf{x}}^{(2)} + a_{\mathbf{x}}^{*}\sigma_{\mathbf{x}} - 2\mu_{\mathbf{x}}^{(1)} + 2\sigma_{\mathbf{x}}\Phi^{-1}(z)}{c_{\mathbf{x}}\sigma_{\mathbf{x}}} \\
&= \frac{-\delta_{\mathbf{x}} + 2\Phi^{-1}(z) + a_{\mathbf{x}}^{*}}{c_{\mathbf{x}}}
\end{aligned} \tag{5.42}$$

and

$$\frac{\mu_{v\mathbf{x}}}{\sigma_{v\mathbf{x}}} = \frac{-\delta_{\mathbf{x}} + b_{\mathbf{x}}^{*}}{c_{\mathbf{x}}}. \tag{5.43}$$

Therefore, we get

$$\begin{aligned} G_1(z) &= \sum_{\mathbf{x}} p_{\mathbf{x}}^{(1)} \left[H\left(\frac{\delta_{\mathbf{x}} + 2\Phi^{-1}(z) - a_{\mathbf{x}}^{*}}{c_{\mathbf{x}}}, \frac{\delta_{\mathbf{x}} - b_{\mathbf{x}}^{*}}{c_{\mathbf{x}}}; \rho_{\mathbf{x}} \right) \right. \\ &\quad \left. + H\left(\frac{-\delta_{\mathbf{x}} + 2\Phi^{-1}(z) + a_{\mathbf{x}}^{*}}{c_{\mathbf{x}}}, \frac{-\delta_{\mathbf{x}} + b_{\mathbf{x}}^{*}}{c_{\mathbf{x}}}; \rho_{\mathbf{x}} \right) \right] \end{aligned} \tag{5.44}$$

The density function of e_{12} is obtained by differentiating the distribution function and using **Result 2** in Appendix A to be

$$\begin{aligned} g_1(z) &= 2\sum_{\mathbf{x}} p_{\mathbf{x}}^{(1)} \frac{1}{c_{\mathbf{x}}\phi(\Phi^{-1}(z))} \left[\Phi\left\{ D_{\mathbf{x}}^{(1)} \right\} \phi\left(\frac{\delta_{\mathbf{x}} + 2\Phi^{-1}(z) - a_{\mathbf{x}}^{*}}{c_{\mathbf{x}}} \right) + \right. \\ &\quad \left. \Phi\left\{ D_{\mathbf{x}}^{(2)} \right\} \phi\left(\frac{-\delta_{\mathbf{x}} + 2\Phi^{-1}(z) + a_{\mathbf{x}}^{*}}{c_{\mathbf{x}}} \right) \right], \end{aligned} \tag{5.45}$$

where

$$\begin{aligned} D_{\mathbf{x}}^{(1)} &= \left\{ \frac{\delta_{\mathbf{x}} - b_{\mathbf{x}}^{*}}{c_{\mathbf{x}}} - \rho_{\mathbf{x}} \frac{\delta_{\mathbf{x}} + 2\Phi^{-1}(z) - a_{\mathbf{x}}^{*}}{c_{\mathbf{x}}} \right\} / (1 - \rho_{\mathbf{x}}^2)^{1/2} \\ &= \frac{(1 - \rho_{\mathbf{x}})\delta_{\mathbf{x}} - 2\rho_{\mathbf{x}}\Phi^{-1}(z) + a_{\mathbf{x}}^{*}\rho_{\mathbf{x}} - b_{\mathbf{x}}^{*}}{(1 - \rho_{\mathbf{x}}^2)^{1/2} c_{\mathbf{x}}} \end{aligned}$$

and

$$\begin{aligned} D_{\mathbf{x}}^{(2)} &= \left\{ \frac{-\delta_{\mathbf{x}} + b_{\mathbf{x}}^{*}}{c_{\mathbf{x}}} - \rho_{\mathbf{x}} \frac{-\delta_{\mathbf{x}} + 2\Phi^{-1}(z) + a_{\mathbf{x}}^{*}}{c_{\mathbf{x}}} \right\} / (1 - \rho_{\mathbf{x}}^2)^{1/2} \\ &= -\frac{(1 - \rho_{\mathbf{x}})\delta_{\mathbf{x}} + 2\rho_{\mathbf{x}}\Phi^{-1}(z) + a_{\mathbf{x}}^{*}\rho_{\mathbf{x}} - b_{\mathbf{x}}^{*}}{(1 - \rho_{\mathbf{x}}^2)^{1/2} c_{\mathbf{x}}}. \end{aligned}$$

By using (5.23) the distribution function of e_{21} becomes

$$G_2(z) = \sum_{\mathbf{x}} p_{\mathbf{x}}^{(2)} \left[H\left(\frac{l_{\mathbf{x}}^{(1)} - \mu_{u\mathbf{x}}}{\sigma_{u\mathbf{x}}}, \frac{\mu_{v\mathbf{x}}}{\sigma_{v\mathbf{x}}}; -\rho_{\mathbf{x}} \right) + H\left(\frac{\mu_{u\mathbf{x}} - l_{\mathbf{x}}^{(2)}}{\sigma_{u\mathbf{x}}}, -\frac{\mu_{v\mathbf{x}}}{\sigma_{v\mathbf{x}}}; -\rho_{\mathbf{x}} \right) \right] \tag{5.46}$$

By realizing

$$\begin{aligned} \frac{l_{\mathbf{x}}^{(1)} - \mu_{u\mathbf{x}}}{\sigma_{u\mathbf{x}}} &= \frac{2(\mu_{\mathbf{x}}^{(2)} + 2\sigma_{\mathbf{x}}\Phi^{-1}(z)) - \mu_{\mathbf{x}}^{(1)} - \mu_{\mathbf{x}}^{(2)} - a_{\mathbf{x}}^{*}\sigma_{\mathbf{x}}}{c_{\mathbf{x}}\sigma_{\mathbf{x}}} \\ &= \frac{-\delta_{\mathbf{x}} + 2\Phi^{-1}(z) - a_{\mathbf{x}}^{*}}{c_{\mathbf{x}}}, \end{aligned} \tag{5.47}$$

$$\frac{\mu_{ux} - l_x^{(2)}}{\sigma_{ux}} = \frac{\mu_x^{(1)} + \mu_x^{(2)} + a_x^- \sigma_x - 2\mu_x^{(2)} + 2\sigma_x \Phi^{-1}(z)}{c_x \sigma_x}$$
$$= \frac{\delta_x + 2\Phi^{-1}(z) + a_x^-}{c_x} \tag{5.48}$$

and

$$\frac{\mu_{vx}}{\sigma_{vx}} = \frac{-\delta_x + b_x^-}{c_x}, \tag{5.49}$$

we obtain

$$G_2(z) = \sum_x p_x^{(2)} \left[H\left(\frac{-\delta_x + 2\Phi^{-1}(z) - a_x^-}{c_x}, \frac{-\delta_x + b_x^-}{c_x}; -\rho_x \right) + H\left(\frac{\delta_x + 2\Phi^{-1}(z) + a_x^-}{c_x}, \frac{\delta_x - b_x^-}{c_x}; -\rho_x \right) \right]. \tag{5.50}$$

The density function of e_{21} is obtained by differentiating the distribution function and using **Result 2** in Appendix A to be

$$g_2(z) = 2\sum_x p_x^{(1)} \frac{1}{c_x \phi(\Phi^{-1}(z))} \left[\Phi\left\{ D_x^{-(1)} \right\} \phi\left(\frac{-\delta_x + 2\Phi^{-1}(z) - a_x^-}{c_x} \right) + \Phi\left\{ D_x^{-(2)} \right\} \phi\left(\frac{\delta_x + 2\Phi^{-1}(z) + a_x^-}{c_x} \right) \right]. \tag{5.51}$$

where

$$D_x^{-(1)} = \left\{ \frac{-\delta_x + b_x^-}{c_x} + \rho_x \frac{-\delta_x + 2\Phi^{-1}(z) - a_x^-}{c_x} \right\} / (1 - \rho_x^2)^{1/2}$$
$$= \frac{-(1+\rho_x)\delta_x + 2\rho_x \Phi^{-1}(z) - a_x^- \rho_x + b_x^-}{(1-\rho_x^2)^{1/2} c_x}$$

and

$$D_x^{-(2)} = \left\{ \frac{\delta_x - b_x^-}{c_x} + \rho_x \frac{\delta_x + 2\Phi^{-1}(z) + a_x^-}{c_x} \right\} / (1 - \rho_x^2)^{1/2}$$
$$= \frac{(1+\rho_x)\delta_x + 2\rho_x \Phi^{-1}(z) + a_x^- \rho_x - b_x^-}{(1-\rho_x^2)^{1/2} c_x}$$

Alternatively one can obtain the cdf and pdf of e_{21} by interchanging sample 1 and sample 2, i.e. by reversing the sign of δ_x, ρ_x and b_x^- of the expressions for cdf and pdf of the e_{12} derived in Eqs. (5.44) and (5.45), respectively.

5.4.2 Expected values of error rates

We obtain the following expression for the expected value of e_{12} by substituting the joint distribution of (U_x, V_x) in (5.31) and using **Result 3** in Appendix A:

$$E(e_{12}) = \sum_x p_x^{(1)} \left[H\left(\frac{2\mu_x^{(1)} - \mu_{ux}}{(c_x^2\sigma_x^2 + 4\sigma_x^2)^{1/2}}, \frac{\mu_{vx}}{c_x\sigma_x}; -\frac{\rho_x c_x \sigma_x}{(c_x^2\sigma_x^2 + 4\sigma_x^2)^{1/2}} \right) + H\left(-\frac{2\mu_x^{(1)} - \mu_{ux}}{(c_x^2\sigma_x^2 + 4\sigma_x^2)^{1/2}}, -\frac{\mu_{vx}}{c_x\sigma_x}; -\frac{\rho_x \tau_x}{(c_x^2\sigma_x^2 + 4\sigma_x^2)^{1/2}} \right) \right] \tag{5.52}$$

which further reduces to

$$E(e_{12}) = \sum_x p_x^{(1)} \left[H\left(\frac{\delta_x - a_x^*}{(c_x^2 + 4)^{1/2}}, \frac{-\delta_x + b_x^*}{c_x}; -\frac{\rho_x c_x}{(c_x^2 + 4)^{1/2}} \right) + H\left(-\frac{\delta_x - a_x^*}{(c_x^2 + 4)^{1/2}}, -\frac{-\delta_x + b_x^*}{c_x}; -\frac{\rho_x c_x}{(c_x^2 + 4)^{1/2}} \right) \right]. \tag{5.53}$$

Similarly, we have

$$E(e_{21}) = \sum_x p_x^{(2)} \left[H\left(\frac{2\mu_x^{(2)} - \mu_{ux}}{(c_x^2\sigma_x^2 + 4\sigma_x^2)^{1/2}}, -\frac{\mu_{vx}}{c_x\sigma_x}; \frac{\rho_x c_x \sigma_x}{(c_x^2\sigma_x^2 + 4\sigma_x^2)^{1/2}} \right) + H\left(-\frac{2\mu_x^{(2)} - \mu_{ux}}{(c_x^2\sigma_x^2 + 4\sigma_x^2)^{1/2}}, \frac{\mu_{vx}}{c_x\sigma_x}; \frac{\rho_x \tau_x}{(c_x^2\sigma_x^2 + 4\sigma_x^2)^{1/2}} \right) \right] \tag{5.54}$$

which further reduces to

$$E(e_{21}) = \sum_x p_x^{(2)} \left[H\left(\frac{-\delta_x - a_x^*}{(c_x^2 + 4)^{1/2}}, \frac{\delta_x - b_x^*}{c_x}; \frac{\rho_x c_x}{(c_x^2 + 4)^{1/2}} \right) + H\left(\frac{\delta_x - a_x^*}{(c_x^2 + 4)^{1/2}}, -\frac{\delta_x - b_x^*}{c_x}; \frac{\rho_x c_x}{(c_x^2 + 4)^{1/2}} \right) \right]. \tag{5.55}$$

Alternatively one can obtain the expected value of e_{21} by interchanging sample 1 and sample 2, i.e. by reversing the sign of δ_x, ρ_x and b_x^* of the expression for expected value of e_{12} derived in Eq. (5.53).

5.5 Results For The Mixture-normal Model

Mixtures of distributions, in particular normal, have been used extensively as models in a wide variety of important practical situations where data can be viewed

as arising from two or more populations mixed in varying proportions. Recent problems which have been addressed by normal mixture models include the identification of outliers (Aitkin and Tunnicliffe Wilson, 1980) and the investigation of the robustness of certain statistics to departures from normality, for example, the sample correlation coefficient as studied by Srivastava and Lee (1984). In addition to this latter role of assessing the performances of estimators in nonnormal situations, normal mixtures have been used of course in the development of robust estimators. For example, under the contaminated normal family as suggested by Tukey (1960), the density of an observation was taken to be a mixture of two univariate normal densities with the same means but where the second component had a greater variance than the first. This family was introduced to model a population which follows a normal distribution except on those few occasions where a grossly atypical observation is recorded. Huber (1964) subsequently considered more general forms of contamination of the normal distribution in the development of his robust M-estimators of a location parameter. Balakrishnan and Kocherlakota (1985) and Balakrishnan et al. (1988) have used a mixture-normal model in assessing the robustness features of the univariate classical linear classification procedure (see Chapter 2) and the univariate classical linear classification procedure based on normal and a dichotomous variable (see Chapter 3), respectively.

In this section, by considering a two-component mixture-normal model for the data on the continuous variable in the classification problem (see Section 2) we derive some approximate expressions for the distribution functions, density functions and expected values of the errors of misclassification of the procedure based on the normal distribution (described in Section 3) in the case when both the errors are allowed to float. This work extends the results of Balakrishnan et al. (1988) and will be used later in Section 6 to assess the robustness features of the classification procedure based on the normal distribution.

5.5.1 Distribution functions of the two errors of misclassification

Let $Y|X = x$ under Π_i be defined by the density function

$$f_i(y|X = x) = p\phi(y; \mu_x^{(i)}, \sigma_x^2) + (1-p)\phi(y; \mu_x^{(i)} + a\sigma_x, \sigma_x^2)$$

for $i = 1,\ 2$. In this case it is known that the exact distribution of (U_x, V_x) is

$$\begin{aligned} g^*(u_x, v_x) &= \sum_r \sum_s \binom{n_x^{(1)}}{r} \binom{n_x^{(2)}}{s} p^{r+s}(1-p)^{n_x-(r+s)} \\ &\quad h^*\left[u_x, v_x, \mu_{ux}, \mu_{vx}, \sigma_{ux}^2, \sigma_{vx}^2; \rho_x\right] \end{aligned} \tag{5.56}$$

where h^* is the bivariate normal density function with the parameters as indicated and

$$\begin{aligned} \mu_{ux} &= \mu_x^{(1)} + \mu_x^{(2)} + 2a\sigma_x - a\sigma_x\left(\frac{r}{n_x^{(1)}} + \frac{s}{n_x^{(2)}}\right), \\ \mu_{vx} &= \mu_x^{(2)} - \mu_x^{(1)} + a\sigma_x\left(\frac{r}{n_x^{(1)}} - \frac{s}{n_x^{(2)}}\right), \\ \sigma_{ux}^2 = \sigma_{vx}^2 &= \sigma_x^2\left(\frac{1}{n_x^{(1)}} + \frac{1}{n_x^{(2)}}\right), \\ n_x &= n_x^{(1)} + n_x^{(2)}, \\ &\text{and} \\ \rho_x &= \frac{n_x^{(1)} - n_x^{(2)}}{n_x}. \end{aligned}$$

By following the same line of reasoning as in the last section we obtain

$$e_{12} = \begin{cases} 1 - F_1\left(\frac{u_x}{2}\right), & \text{for} \quad V_x > 0 \\ F_1\left(\frac{u_x}{2}\right), & \text{for} \quad V_x \le 0 \end{cases} \tag{5.57}$$

where $F_1(\ )$ is the conditional distribution function of population 1. By writing $\Phi_1(z)$ for $p\Phi(z) + (1-p)\Phi(z-a)$ we obtain

$$e_{12} = \begin{cases} 1 - \Phi_1\left(\frac{\frac{u_x}{2} - \mu_x^{(1)}}{\sigma_x}\right), & \text{for} \quad V_x > 0 \\ \Phi_1\left(\frac{\frac{u_x}{2} - \mu_x^{(1)}}{\sigma_x}\right), & \text{for} \quad V_x \le 0 \end{cases} \tag{5.58}$$

Therefore the asymptotic distribution function of e_{12} is

$$\begin{aligned} G_1(z) &= \sum_{\mathbf{x}} p_{\mathbf{x}}^{(1)} \left\{ Pr\left\{ U_{\mathbf{x}} \geq k_{\mathbf{x}}^{(1)}, V_{\mathbf{x}} > 0 \right\} + \right. \\ &\quad \left. Pr\left\{ U_{\mathbf{x}} \leq k_{\mathbf{x}}^{(2)}, V_{\mathbf{x}} < 0 \right\} \right\} \end{aligned} \tag{5.59}$$

where

$$\begin{aligned} k_{\mathbf{x}}^{(1)} &= 2\{\mu_{\mathbf{x}}^{(1)} + \Phi_1^{-1}(1-z)\} \\ k_{\mathbf{x}}^{(2)} &= 2\{\mu_{\mathbf{x}}^{(1)} + \Phi_1^{-1}(z)\} \end{aligned}$$

Using the joint p.d.f. of $(U_{\mathbf{x}}, V_{\mathbf{x}})$ given in (5.56) we obtain the cumulative distribution function of e_{12} to be

$$\begin{aligned} G_1(z) &= \sum_{\mathbf{x}} p_{\mathbf{x}}^{(1)} \sum_r \sum_s \binom{n_{\mathbf{x}}^{(1)}}{r} \binom{n_{\mathbf{x}}^{(2)}}{s} p^{r+s}(1-p)^{n_{\mathbf{x}}-(r+s)} \\ &\quad \left\{ H\left(\frac{\mu_{u\mathbf{x}} - k_{\mathbf{x}}^{(1)}}{\sigma_{u\mathbf{x}}}, \frac{\mu_{v\mathbf{x}}}{\sigma_{v\mathbf{x}}}; \rho_{\mathbf{x}} \right) + \right. \\ &\quad \left. H\left(\frac{k_{\mathbf{x}}^{(2)} - \mu_{u\mathbf{x}}}{\sigma_{u\mathbf{x}}}, -\frac{\mu_{v\mathbf{x}}}{\sigma_{v\mathbf{x}}}; \rho_{\mathbf{x}} \right) \right\} \end{aligned} \tag{5.60}$$

Since

$$\begin{aligned} \frac{\mu_{u\mathbf{x}} - k_{\mathbf{x}}^{(1)}}{\sigma_{u\mathbf{x}}} &= \frac{\mu_{\mathbf{x}}^{(1)} + \mu_{\mathbf{x}}^{(2)} + 2a\sigma_{\mathbf{x}} - a\sigma_{\mathbf{x}}\left(\frac{r}{n_{\mathbf{x}}^{(1)}} + \frac{s}{n_{\mathbf{x}}^{(2)}}\right) - 2\{\mu_{\mathbf{x}}^{(1)} + \Phi_1^{-1}(1-z)\}}{\sigma_{\mathbf{x}} m_{\mathbf{x}}} \\ &= \frac{-\frac{\mu_{\mathbf{x}}^{(1)} - \mu_{\mathbf{x}}^{(2)}}{\sigma_{\mathbf{x}}} + 2a - a\left(\frac{r}{\nu_{\mathbf{x}}^{(1)}} + \frac{s}{\nu_{\mathbf{x}}^{(2)}}\right) - 2\Phi_1^{-1}(1-z)}{m_{\mathbf{x}}} \\ &= \frac{-\delta_{\mathbf{x}} + 2a - a\left(\frac{r}{n_{\mathbf{x}}^{(1)}} + \frac{s}{n_{\mathbf{x}}^{(2)}}\right) - 2\Phi_1^{-1}(1-z)}{m_{\mathbf{x}}} \end{aligned}$$

and

$$\begin{aligned} \frac{k_{\mathbf{x}}^{(2)} - \mu_{u\mathbf{x}}}{\sigma_{u\mathbf{x}}} &= \frac{2\{\mu_{\mathbf{x}}^{(1)} + \Phi_1^{-1}(z)\} - \left\{\mu_{\mathbf{x}}^{(1)} + \mu_{\mathbf{x}}^{(2)} + 2a\sigma_{\mathbf{x}} - a\sigma_{\mathbf{x}}\left(\frac{r}{n_{\mathbf{x}}^{(1)}} + \frac{s}{n_{\mathbf{x}}^{(2)}}\right)\right\}}{\sigma_{\mathbf{x}} m_{\mathbf{x}}} \\ &= \frac{\frac{\mu_{\mathbf{x}}^{(1)} - \mu_{\mathbf{x}}^{(2)}}{\sigma_{\mathbf{x}}} - 2a + a\left(\frac{r}{n_{\mathbf{x}}^{(1)}} + \frac{s}{n_{\mathbf{x}}^{(2)}}\right) + 2\Phi_1^{-1}(z)}{m_{\mathbf{x}}} \end{aligned}$$

$$= \frac{\delta_x - 2a + a\left(\frac{r}{n_x^{(1)}} + \frac{s}{n_x^{(2)}}\right) + 2\Phi_1^{-1}(z)}{m_x}$$

Also

$$\frac{\mu_{vx}}{\sigma_{vx}} = \frac{\mu_x^{(2)} - \mu_x^{(1)} + a\sigma_x\left(\frac{r}{n_x^{(1)}} - \frac{s}{n_x^{(2)}}\right)}{\sigma_x m_x}$$
$$= \frac{-\delta_x + a\left(\frac{r}{n_x^{(1)}} - \frac{s}{n_x^{(2)}}\right)}{m_x}$$

Therefore we can write the probability distribution function of e_{12} in the following form:

$$G_1(z) = \sum_x p_x^{(1)} \sum_r \sum_s \binom{n_x^{(1)}}{r} \binom{n_x^{(2)}}{s} p^{r+s}(1-p)^{n_x-(r+s)}$$
$$\left\{ H\left[\frac{-\delta_x + 2a - a\left(\frac{r}{n_x^{(1)}} + \frac{s}{n_x^{(2)}}\right) - 2\Phi_1^{-1}(1-z)}{m_x}, \frac{-\delta_x + a\left(\frac{r}{n_x^{(1)}} - \frac{s}{n_x^{(2)}}\right)}{m_x}, \rho_x\right] + \right.$$
$$\left. H\left[\frac{\delta_x - 2a + a\left(\frac{r}{n_x^{(1)}} + \frac{s}{n_x^{(2)}}\right) + 2\Phi_1^{-1}(z)}{m_x}, -\frac{-\delta_x + a\left(\frac{r}{n_x^{(1)}} - \frac{s}{n_x^{(2)}}\right)}{m_x}, \rho_x\right]\right\} \quad (5.61)$$

The density function of e_{12} can be obtained by differentiating the above distribution function and using **Result 2** in Appendix A and the fact that

$$\frac{d\Phi_1^{-1}(z)}{dz} = \frac{1}{\phi_1(\Phi_1^{-1}(z))}$$
$$\frac{d\Phi_1^{-1}(1-z)}{dz} = -\frac{1}{\phi_1(\Phi_1^{-1}(1-z))}$$

Thus

$$g_1(z) = \sum_x \sum_r \sum_s \binom{n_x^{(1)}}{r} \binom{n_x^{(2)}}{s} p^{r+s}(1-p)^{n_x-(r+s)} p_x^{(1)}$$
$$\left[-\frac{2}{m_x \phi_1(\Phi_1^{-1}(1-z))} \Phi\left\{E_x^{(1)}\right\} \phi\left(\frac{-\delta_x + 2a - a\left(\frac{r}{n_x^{(1)}} + \frac{s}{n_x^{(2)}}\right) - 2\Phi_1^{-1}(1-z)}{m_x}\right) + \right.$$

$$\frac{2}{m_x\phi_1(\Phi_1^{-1}(z))}\,\Phi\left\{E_x^{(2)}\right\}\phi\left(\frac{-\delta_x-2a+a\left(\frac{r}{n_x^{(1)}}+\frac{s}{n_x^{(2)}}\right)+2\Phi_1^{-1}(z)}{m_x}\right)\Bigg] \tag{5.62}$$

where

$$E_x^{(1)} = \frac{-\delta_x+a\left(\frac{r}{n_x^{(1)}}-\frac{s}{n_x^{(2)}}\right)-\rho_x\left[-\delta_x+2a-a\left(\frac{r}{n_x^{(1)}}+\frac{s}{n_x^{(2)}}\right)-2\Phi_1^{-1}(1-z)\right]}{(1-\rho_x)^{1/2}m_x} \tag{5.63}$$

$$= \frac{-(1-\rho_x)\delta_x-2\rho_x a+\frac{ar(1+\rho_x)}{n_x^{(1)}}-\frac{as(1-\rho_x)}{n_x^{(2)}}+2\rho_x\Phi_1^{-1}(1-z)}{(1-\rho_x)^{1/2}m_x} \tag{5.64}$$

and

$$E_x^{(2)} = \frac{\delta_x-a\left(\frac{r}{n_x^{(1)}}-\frac{s}{n_x^{(2)}}\right)-\rho_x\left[\delta_x-2a+a\left(\frac{r}{n_x^{(1)}}+\frac{s}{n_x^{(2)}}\right)+2\Phi_1^{-1}(z)\right]}{(1-\rho_x)^{1/2}m_x} \tag{5.65}$$

$$= \frac{(1-\rho_x)\delta_x+2\rho_x a-\frac{ar(1+\rho_x)}{n_x^{(1)}}-\frac{as(1-\rho_x)}{n_x^{(2)}}-2\rho_x\Phi_1^{-1}(z)}{(1-\rho_x)^{1/2}m_x} \tag{5.66}$$

Similarly, we can derive the distribution function and density function of e_{21}.

5.5.2 Expected values of error rates

Following the same line of argument as used in Section 4, the expected value of e_{12} may be computed as

$$\begin{aligned} E(e_{12}) &= \sum_x p_x^{(1)}\{\, Pr\left[U_x<2Y, V_x>0|Y\sim pN(\mu_x^{(1)},\sigma_x^2)+(1-p)N(\mu_x^{(1)}+a\sigma_x,\sigma_x^2)\right] + \\ &\quad Pr\left[U_x>2Y, V_x<0|Y\sim pN(\mu_x^{(1)},\sigma_x^2)+(1-p)N(\mu_x^{(1)}+a\sigma_x,\sigma_x^2)\right]\,\} \end{aligned} \tag{5.67}$$

We can simplify (5.67) by applying **Result 4** in Appendix A to obtain

$$\begin{aligned} E(e_{12}) &= \sum_x\sum_r\sum_s\sum_i^2 p_i\binom{n_x^{(1)}}{r}\binom{n_x^{(2)}}{r}p^{r+s}(1-p)^{n_x-(r+s)}p_x^{(1)} \\ &\quad \Bigg[H\left(\frac{2\mu_i-\mu_{ux}}{(4\sigma_x^2+\sigma_{ux}^2)^{1/2}},\frac{\mu_{vx}}{\sigma_{vx}},-\frac{\rho_x\sigma_{ux}}{(4\sigma_x^2+\sigma_{ux}^2)^{1/2}}\right)+ \\ &\quad H\left(-\frac{2\mu_i-\mu_{ux}}{(4\sigma_x^2+\sigma_{ux}^2)^{1/2}},-\frac{\mu_{vx}}{\sigma_{vx}},-\frac{\rho_x\sigma_{ux}}{(4\sigma^2+\sigma_{ux}^2)^{1/2}}\right)\Bigg], \end{aligned} \tag{5.68}$$

where

$$p_1 = 1-p_2 \;=\; p,$$

$$\mu_1 = \mu_X^{(1)},$$

and

$$\mu_2 = \mu_X^{(1)} + a\sigma_X.$$

In a similar way, we can derive the expected value of e_{21}.

5.6 Robustness study for two dichotomous and one continuous variable case

In this section, we concentrate on the special case $k = 2$, that is, when the data contains two dichotomous and one continuous variables. By making use of the results developed in the last two sections for the outlier-normal and the mixture-normal models, we examine the robustness of the procedure developed in Section 3 based on the normal distribution. For this purpose All the calculations were carried out in single precision using Fortran on VAX 8530 computer. To calculate the cumulative distribution function of bivariate normal distribution we used the formula (Owen (1980))

$$H[x, y; \rho] = \frac{1}{2}\{\Phi(x) + \Phi(y) - \delta(x, y)\} - T(x, a_1) - T(y, a_2),$$

where

$$T(h, a) = \frac{1}{2\pi}\int_0^a \frac{\exp -\frac{1}{2}h^2(1 + x^2)}{1 + x^2}dx,$$

$$\delta(x, y) = \begin{cases} 0 & \text{if } xy > 0 \text{ or } xy = 0 \text{ and } x + y \geq 0 \\ 1 & \text{otherwise} \end{cases},$$

$$a_1 = \frac{\frac{y}{x} - \rho}{(1 - \rho^2)^{1/2}},$$

and

$$a_2 = \frac{\frac{x}{y} - \rho}{(1 - \rho^2)^{1/2}}.$$

We utilised the IMSL/STAT *anordf* and IMSL/MATH *qdagis* for the calculation of cdf of univariate normal and the T function, respectively.

5.6.1 Parameter Selection

The choice of the sample sizes, n_1 and n_2, were 50 each. We considered the following five different set of values of the parameters:

Model I	$\mu_1 = 0$	$\mu_2 = 1$	$\sigma^2 = 1$	$\underline{\delta}' = (0.25, 0.25)$	$\underline{\gamma}' = (1, 1)$
Model II	$\mu_1 = 0$	$\mu_2 = 2$	$\sigma^2 = 1$	$\underline{\delta}' = (0.25, 0.25)$	$\underline{\gamma}' = (1, 1)$
Model III	$\mu_1 = 0$	$\mu_2 = 1$	$\sigma^2 = 4$	$\underline{\delta}' = (0.25, 0.25)$	$\underline{\gamma}' = (1, 1)$
Model IV	$\mu_1 = 0$	$\mu_2 = 1$	$\sigma^2 = 1$	$\underline{\delta}' = (0.25, 0.25)$	$\underline{\gamma}' = (4, 1)$
Model V	$\mu_1 = 0$	$\mu_2 = 1$	$\sigma^2 = 1$	$\underline{\delta}' = (0.25, 0.25)$	$\underline{\gamma}' = (1, 4)$

In each case we used four different distributions of P, namely P_1, P_2, P_3, and P_4 where

$$P_1 = \begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 0.25 \end{bmatrix} P_2 = \begin{bmatrix} 0.2 & 0.3 \\ 0.3 & 0.2 \end{bmatrix} P_3 = \begin{bmatrix} 0.4 & 0.1 \\ 0.1 & 0.4 \end{bmatrix} P_4 = \begin{bmatrix} 0.1 & 0.4 \\ 0.1 & 0.4 \end{bmatrix}$$

and we studied the distribution functions of the errors corresponding to these four different P 's in the same graph. In the outlier-normal case we used $a_1 = a_2 = 1$ and $b_1 = b_2 = 5$ and total number of outliers in each sample as 0, 2 and 5. In the mixture-normal case we used a to be 2, 4 and p to be 0.9, 0.8 respectively.

5.6.2 Comments on the effects of non-normality

For the various choices of parameters described above, the distribution functions and the expected values of the two errors of misclassification were computed under the normal and non-normal cases and compared. For the purpose of illustration, we have presented a few selected plots of the distribution functions at the end of this chapter.

Based on these comparisons, we bring out the following points with regard to the effects of non-normality on the linear procedure based on the normal distribution:

1. In all cases, a shift in the mean results in the distribution function $G_1(z)$ rising more steeply. This means that the error e_{12} remains small with a higher

probability. This is understandable as a shift in the mean increases the distance between the two populations. This also results in smaller values for the expected values of the error rates.

2. As the distribution functions remained relatively stable over the different choices of P and the expected values of the errors also reflected this pattern, we noted that P does not have a marked effect on the error rates.

3. The distribution function $G_1(z)$ in the outlier-normal case is observed to be higher than $G_1(z)$ for the normal case whenever $z \leq E(e_{12})$ but tends to be flatter when $z > E(e_{12})$. This makes the probability of the error rates being large much higher under the outlier-normal model than in the normal case.

4. The effect of outliers being present in both samples is much more pronounced than in the case when outliers are present only in one sample. For example, the distribution functions for the case when $s_1 \neq 0$ are much flatter than when $s_1 = 0$. The result of this is that when $s_1 \neq 0$ the classification procedure has a higher probability of the errors being large.

5. The effect on the distribution functions under the mixture- normal model was similar to that of the outlier-normal model. In general, the distribution function of e_{12} under the mixture- normal model was shifted to the right as compared to the distribution of e_{12} under the normal distribution thus indicating that the errors of misclassification e_{12} tends to be stochastically larger in the mixture case than under the normal. This pattern was observed through the expected values of the error rates as well.

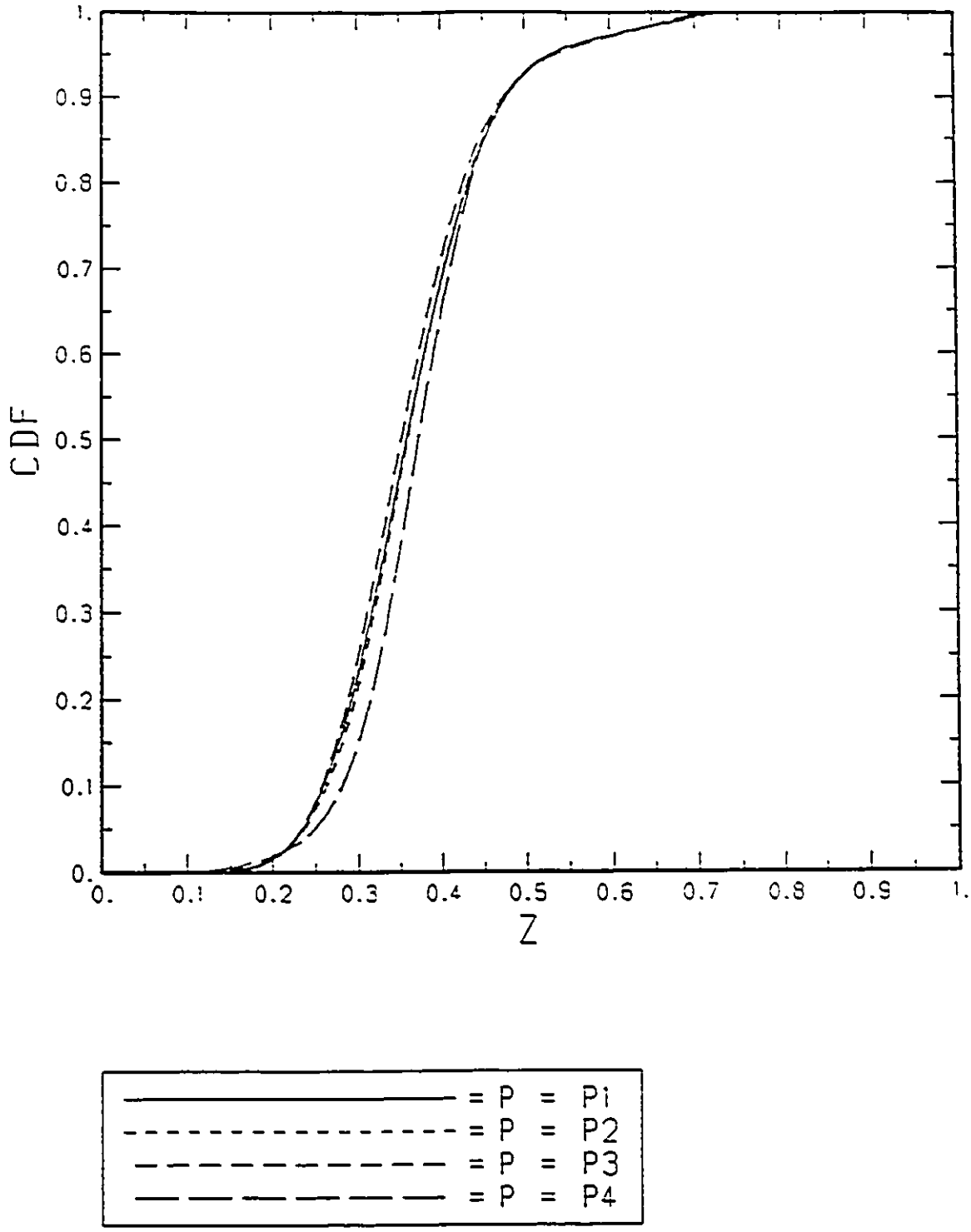
Table 5.1: Expected values of e_{12} under outlier distribution for five models with $n_1 = n_2 = 50$

P_1	P_2	P_3	P_4
$a_1 = a_2 = 1$	$b_1 = b_2 = 1$	$s_x^{(1)} = 0$	$s_x^{(2)} = 0$
0.36814	0.36839	0.36312	0.37897
0.23477	0.23666	0.22966	0.25342
0.43612	0.43501	0.43097	0.43510
0.39337	0.39442	0.38461	0.40047
0.39337	0.39442	0.38461	0.41891
$a_1 = a_2 = 1$	$b_1 = b_2 = 5$	$s_x^{(1)} = 2$	$s_x^{(2)} = 2$
0.37559	0.37487	0.36469	0.38397
0.23107	0.23231	0.22716	0.24883
0.44269	0.44111	0.43534	0.43967
0.39940	0.40020	0.38551	0.40480
0.39940	0.40020	0.38551	0.42570
$a_1 = a_2 = 1$	$b_1 = b_2 = 5$	$s_x^{(1)} = 5$	$s_x^{(2)} = 5$
0.38362	0.38240	0.36710	0.38938
0.23283	0.23191	0.22615	0.24714
0.44611	0.44437	0.43767	0.44189
0.40521	0.40591	0.38650	0.40877
0.40521	0.40591	0.38650	0.43012

Table 5.2: Expected values of e_{12} under mixture distribution for five models with $n_1 = n_2 = 50$.

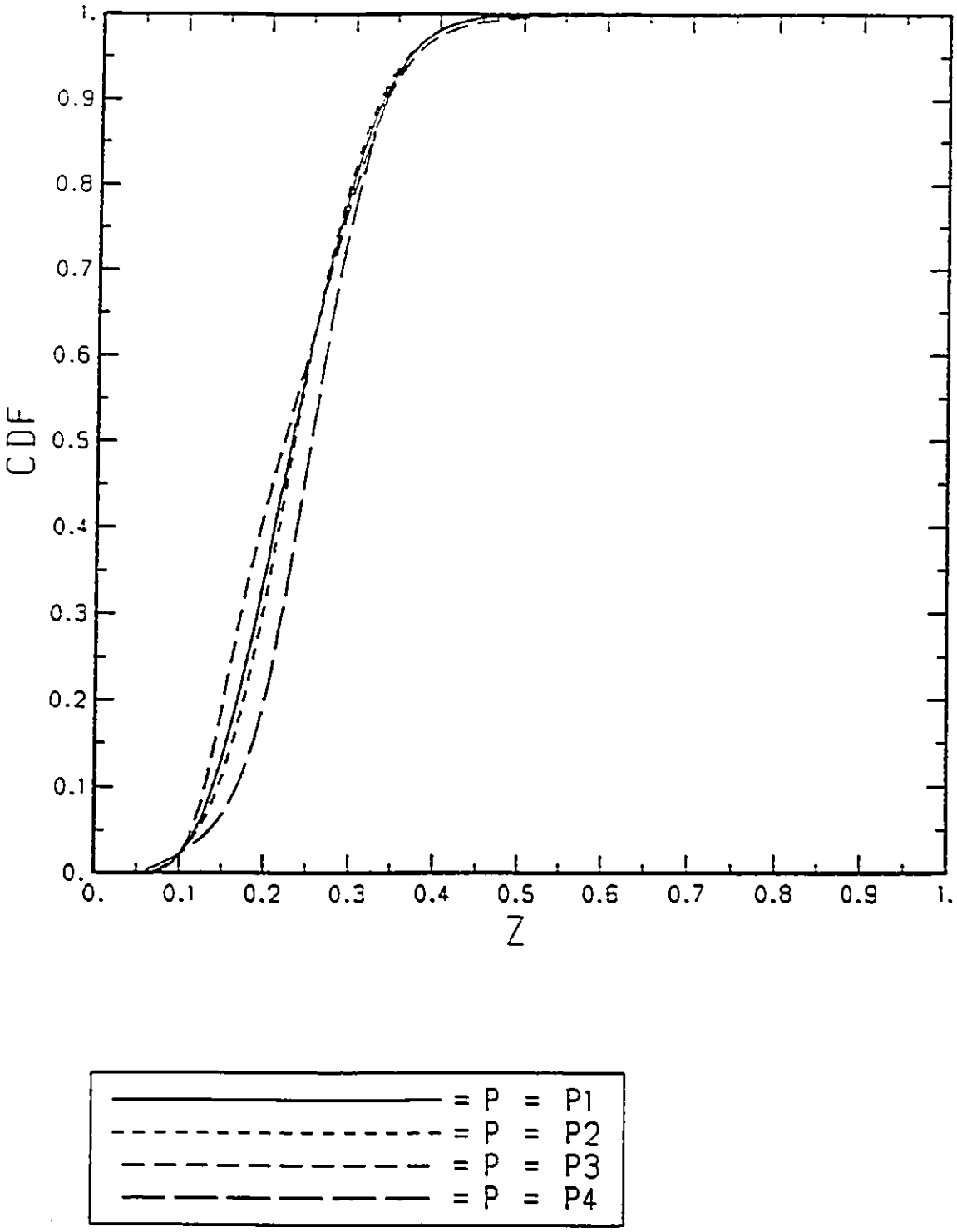
P_1	P_2	P_3	P_4
$p = 0.9$		$a = 2.0$	
0.37094	0.37047	0.36422	0.37914
0.24857	0.25009	0.24518	0.26436
0.43523	0.43373	0.42830	0.43240
0.39469	0.39516	0.38464	0.39956
0.39469	0.39516	0.38464	0.41801
$p = 0.9$		$a = 4.0$	
0.35928	0.35748	0.34541	0.36302
0.23451	0.23421	0.23277	0.24492
0.42557	0.42314	0.41433	0.41901
0.38347	0.38313	0.36686	0.38448
0.38347	0.38313	0.36686	0.40605
$p = 0.8$		$a = 2.0$	
0.38193	0.38123	0.37437	0.38837
0.26958	0.27084	0.26684	0.28352
0.44001	0.43848	0.43286	0.43665
0.40323	0.40351	0.39283	0.40683
0.40323	0.40351	0.39283	0.42409
$p = 0.8$		$a = 4.0$	
0.37584	0.37368	0.36038	0.37587
0.27557	0.27390	0.27322	0.27952
0.42995	0.42749	0.41850	0.42256
0.39571	0.39493	0.37845	0.39395
0.39571	0.39493	0.37845	0.41270

Figure 5.1: Distribution function of error of misclassification e_{12}: plot 1



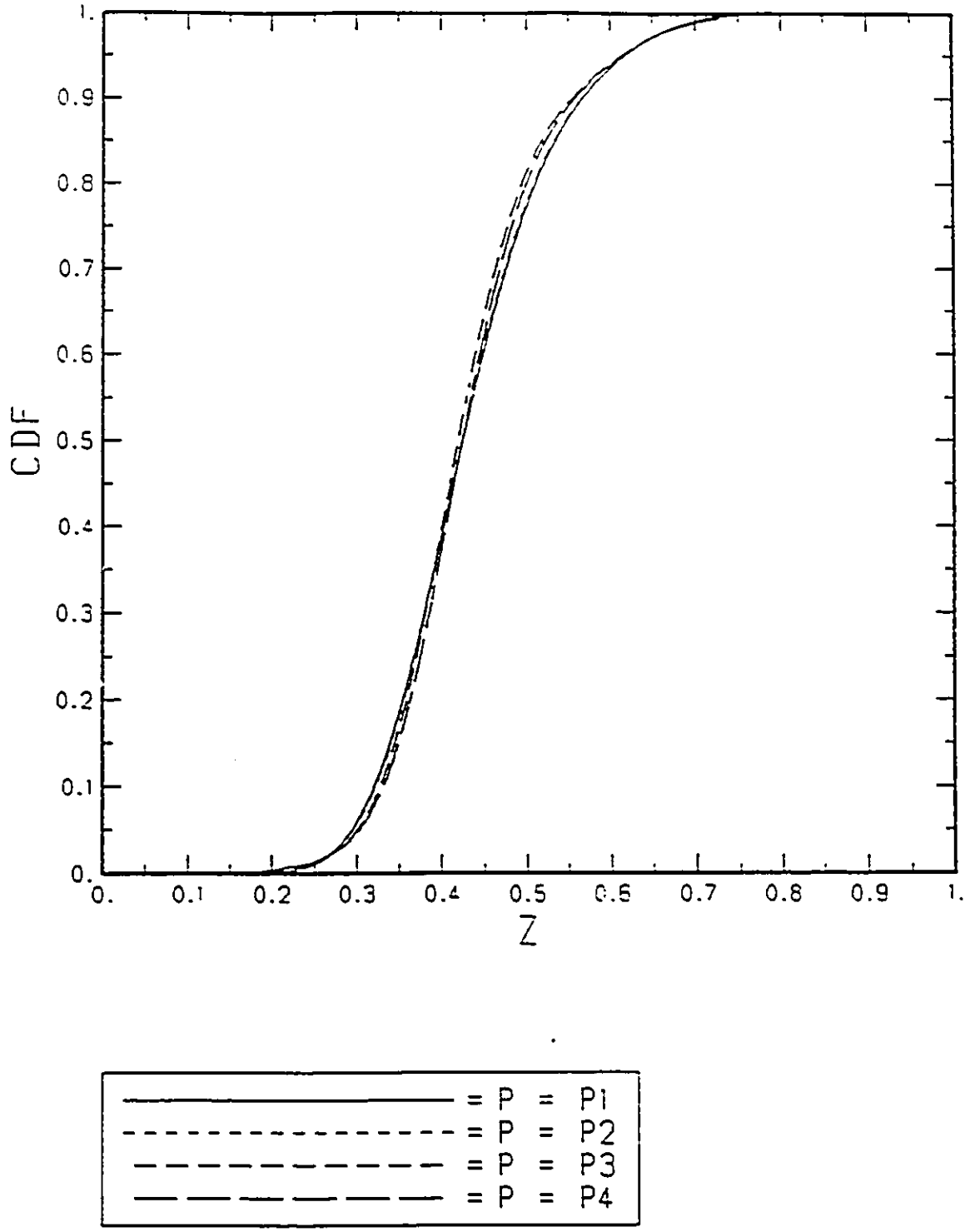
Model I

Figure 5.2: Distribution function of error of misclassification e_{12}: plot 2



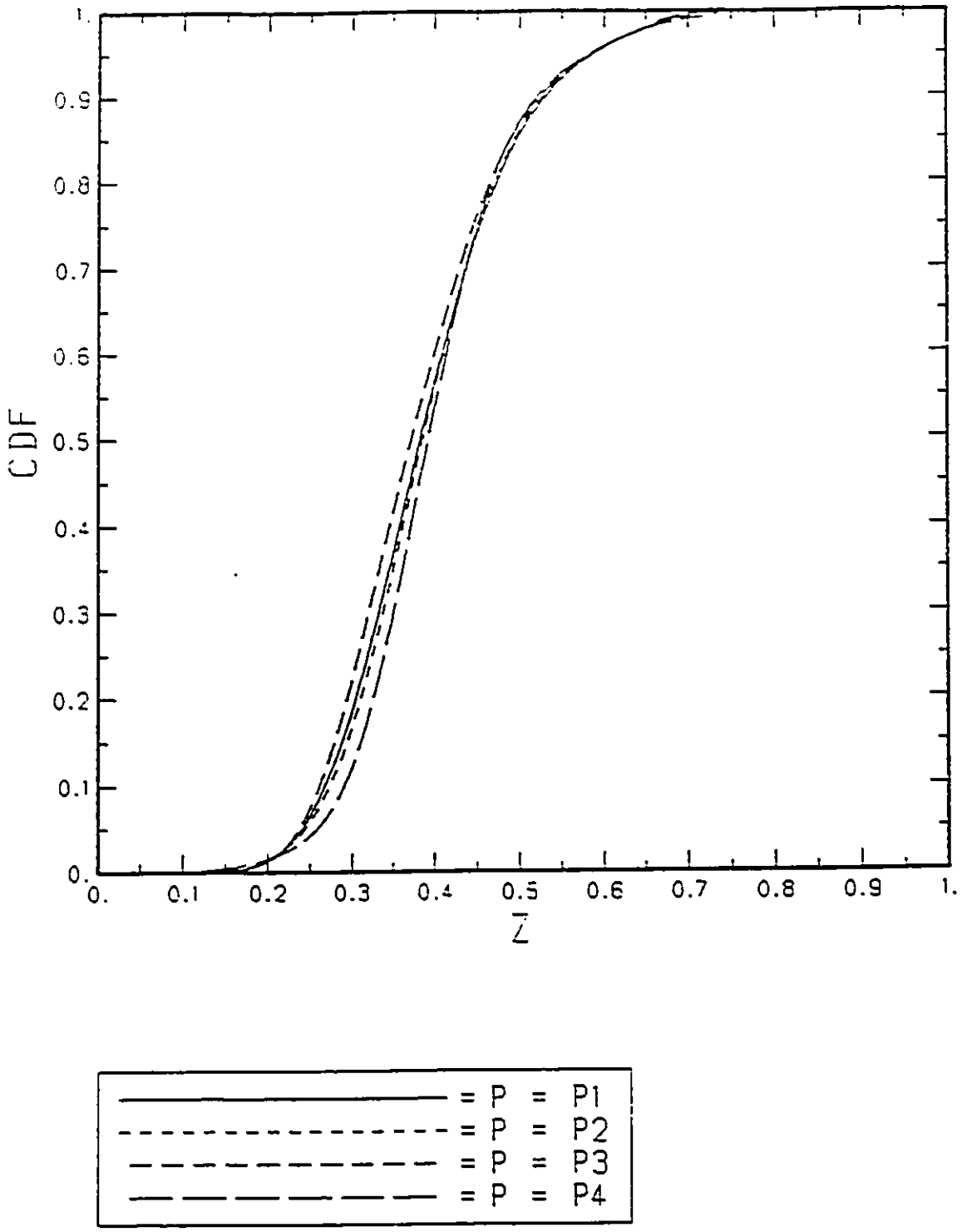
Model II

Figure 5.3: Distribution function of error of misclassification e_{12}: plot 3



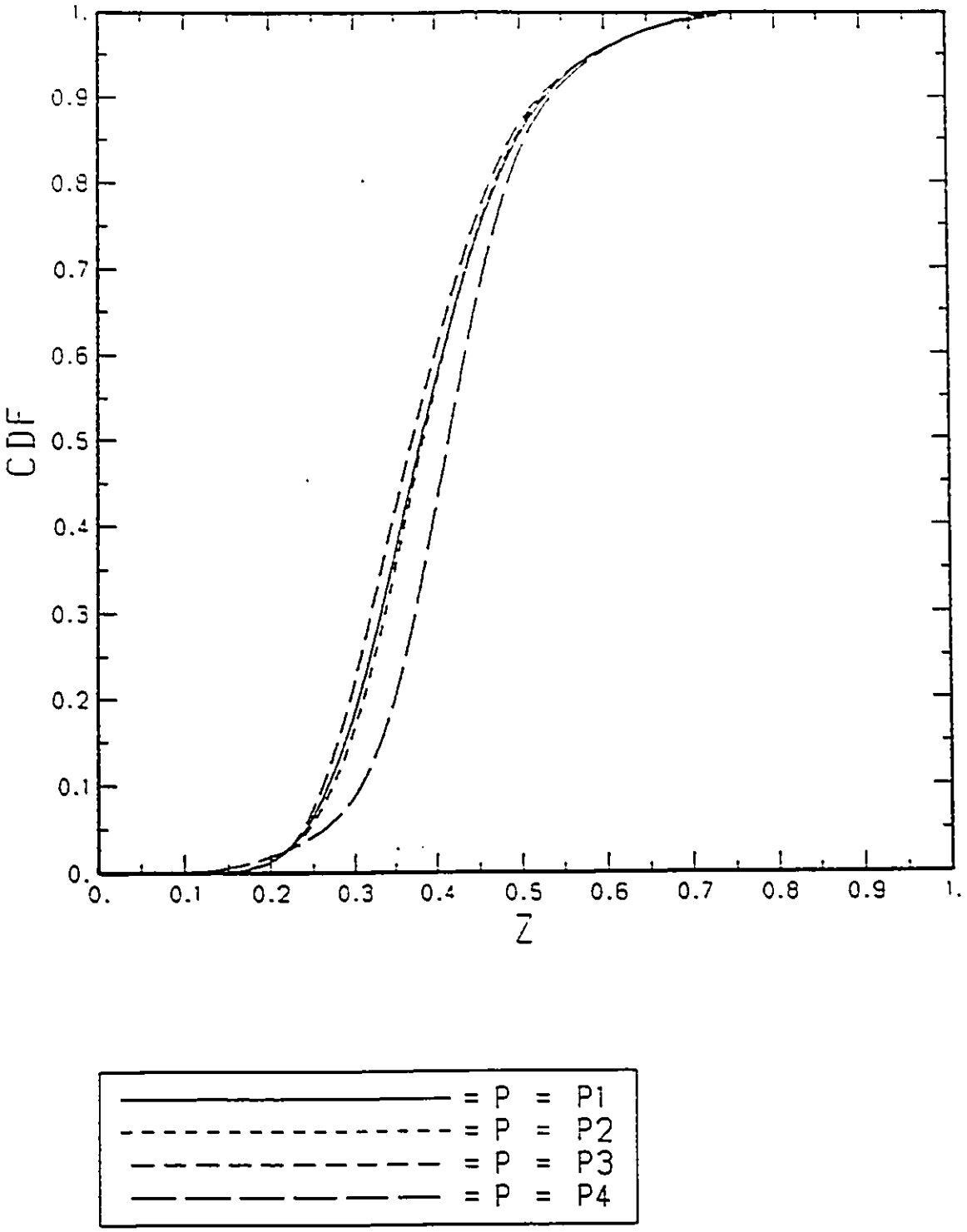
Model III

Figure 5.4: Distribution function of error of misclassification e_{12}: plot 4



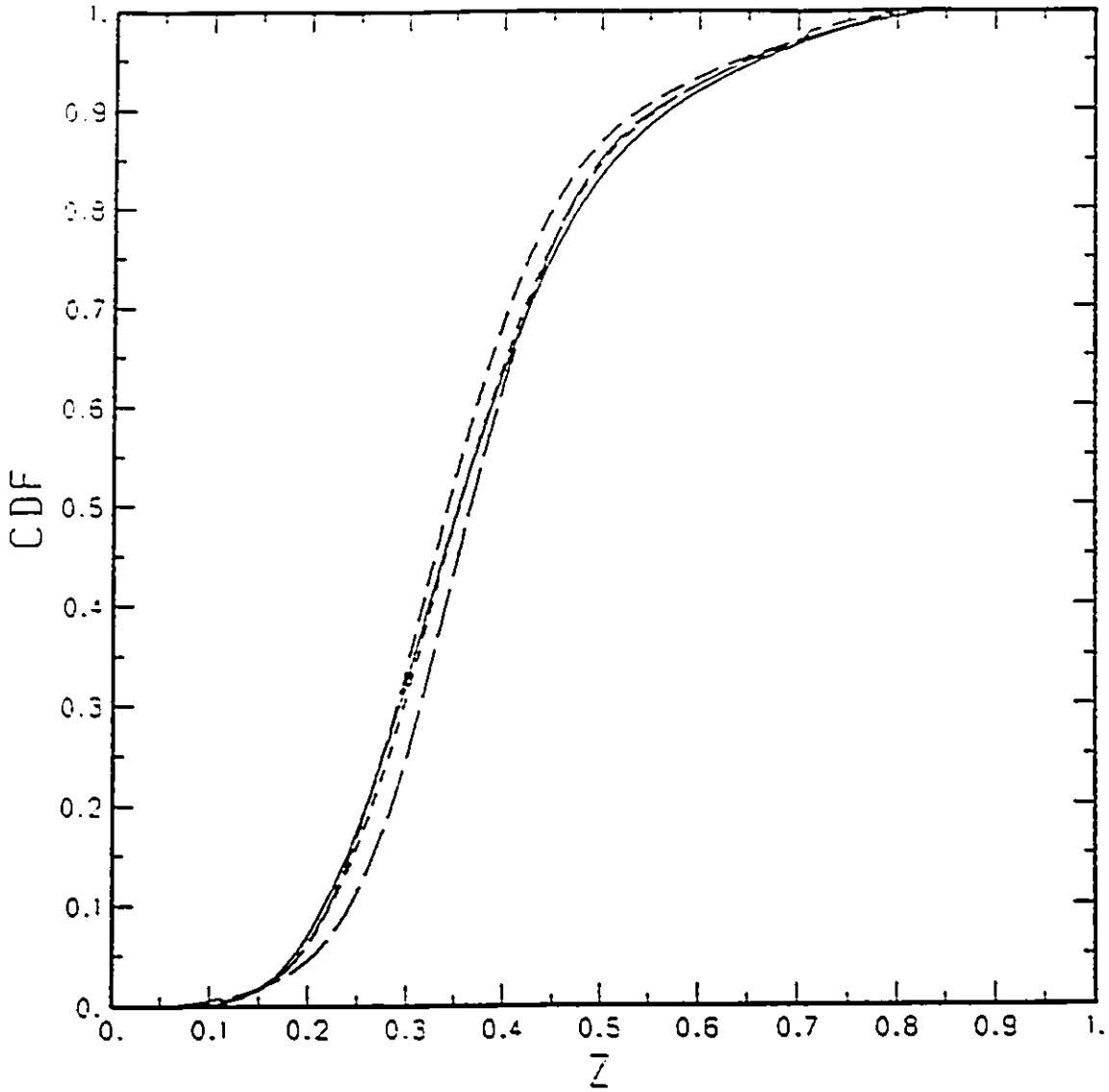
Model IV

Figure 5.5: Distribution function of error of misclassification e_{12}: plot 5



Model V

Figure 5.6: Distribution function of error of misclassification e_{12}: plot 6



= P = P1
= P = P2
= P = P3
= P = P4

Model I
A1=A2=1 B1=B2=5
S1=2 S2=2

Figure 5.7: Distribution function of error of misclassification e_{12}: plot 7

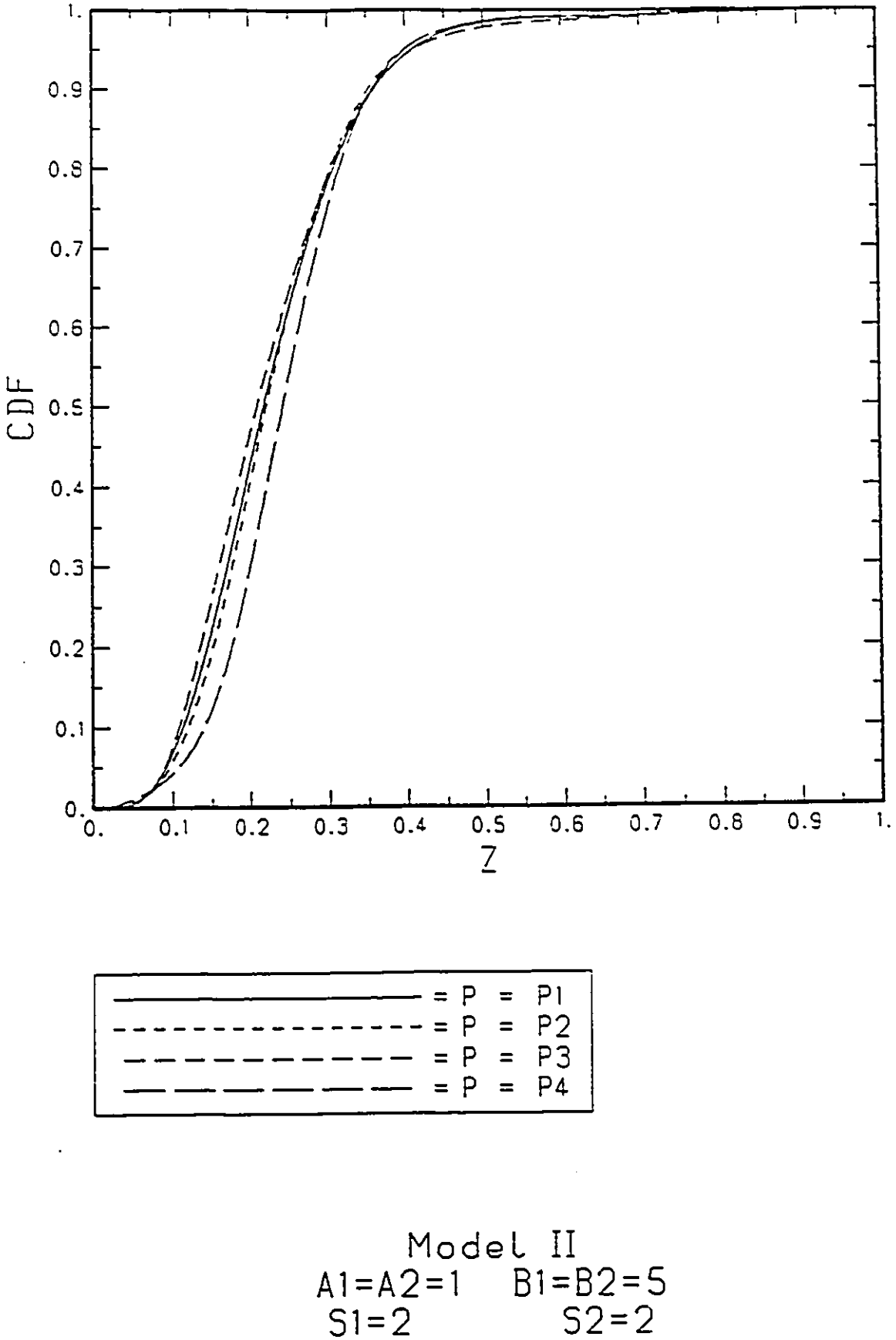
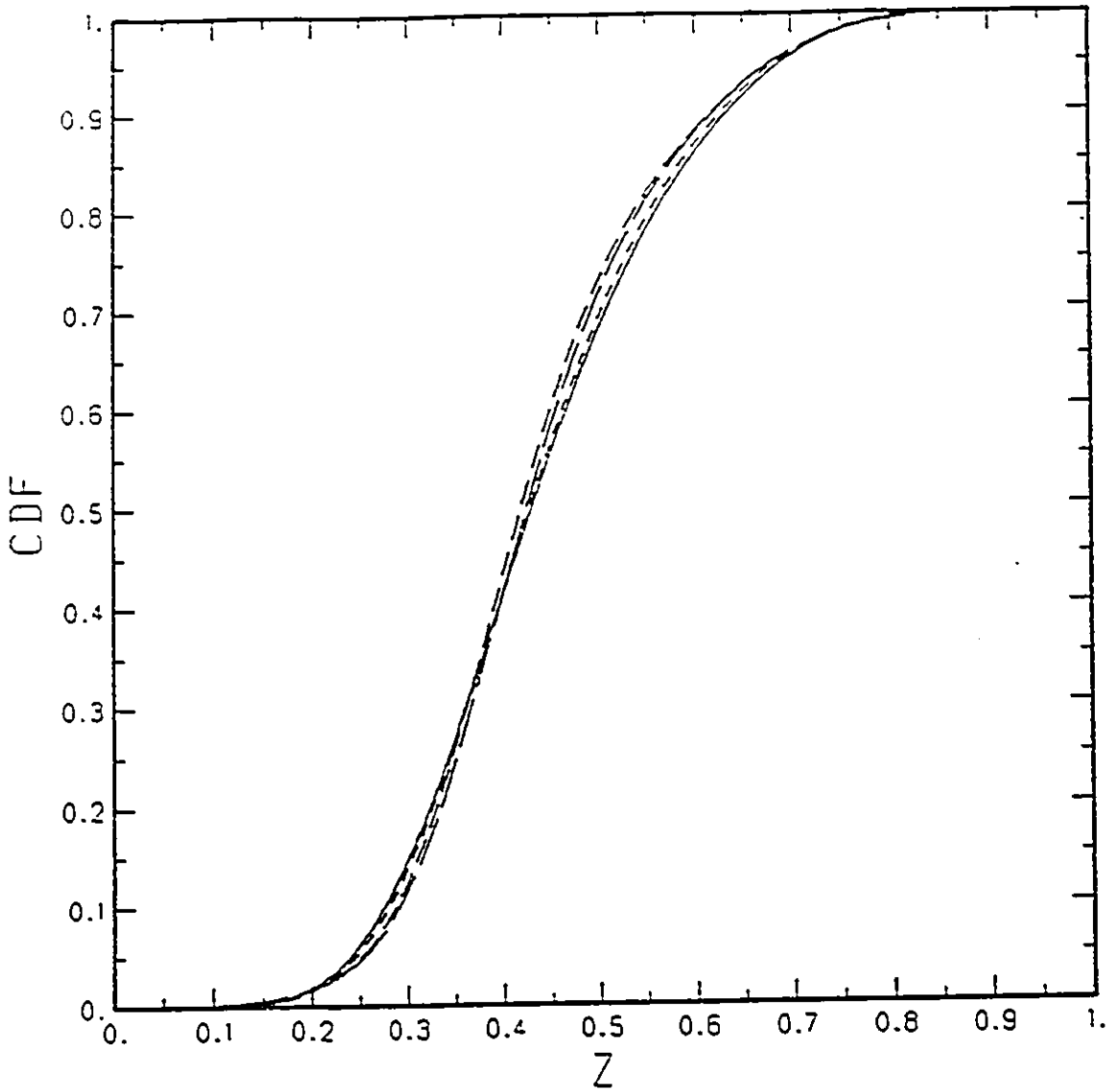


Figure 5.8: Distribution function of error of misclassification e_{12}: plot S



Model III
A1=A2=1 B1=B2=5
S1=2 S2=2

Figure 5.9: Distribution function of error of misclassification e_{12}: plot 9

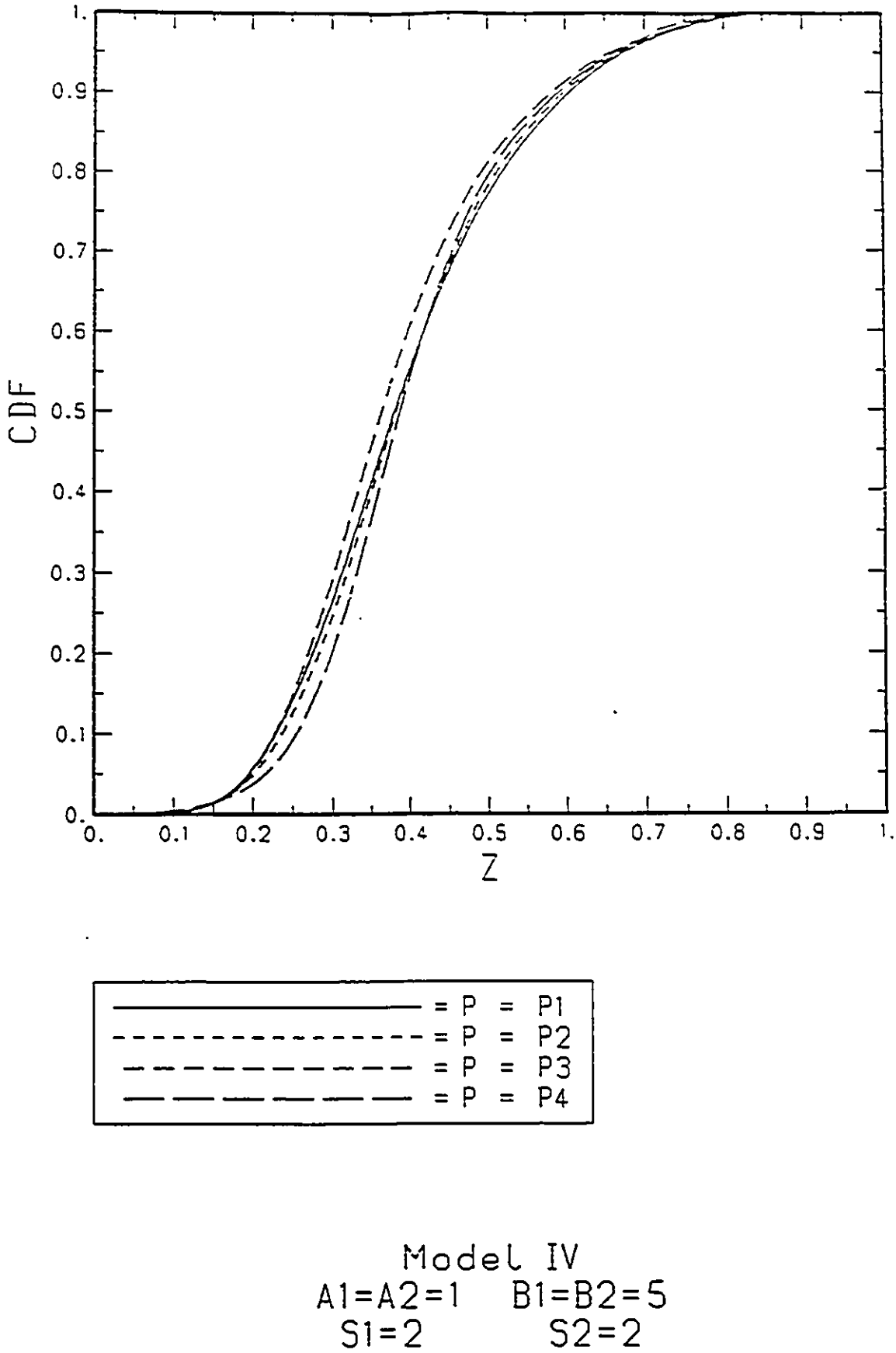
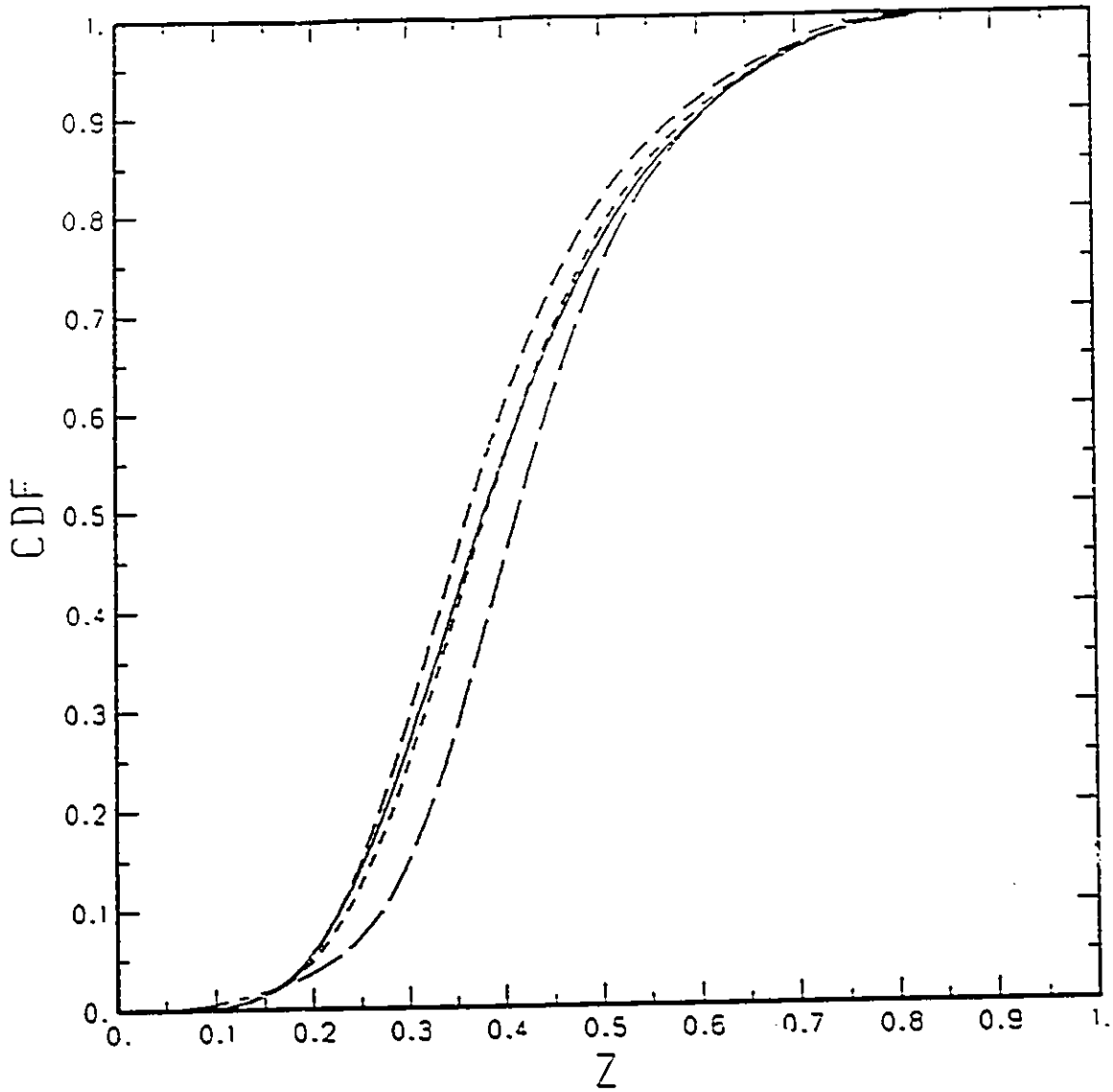


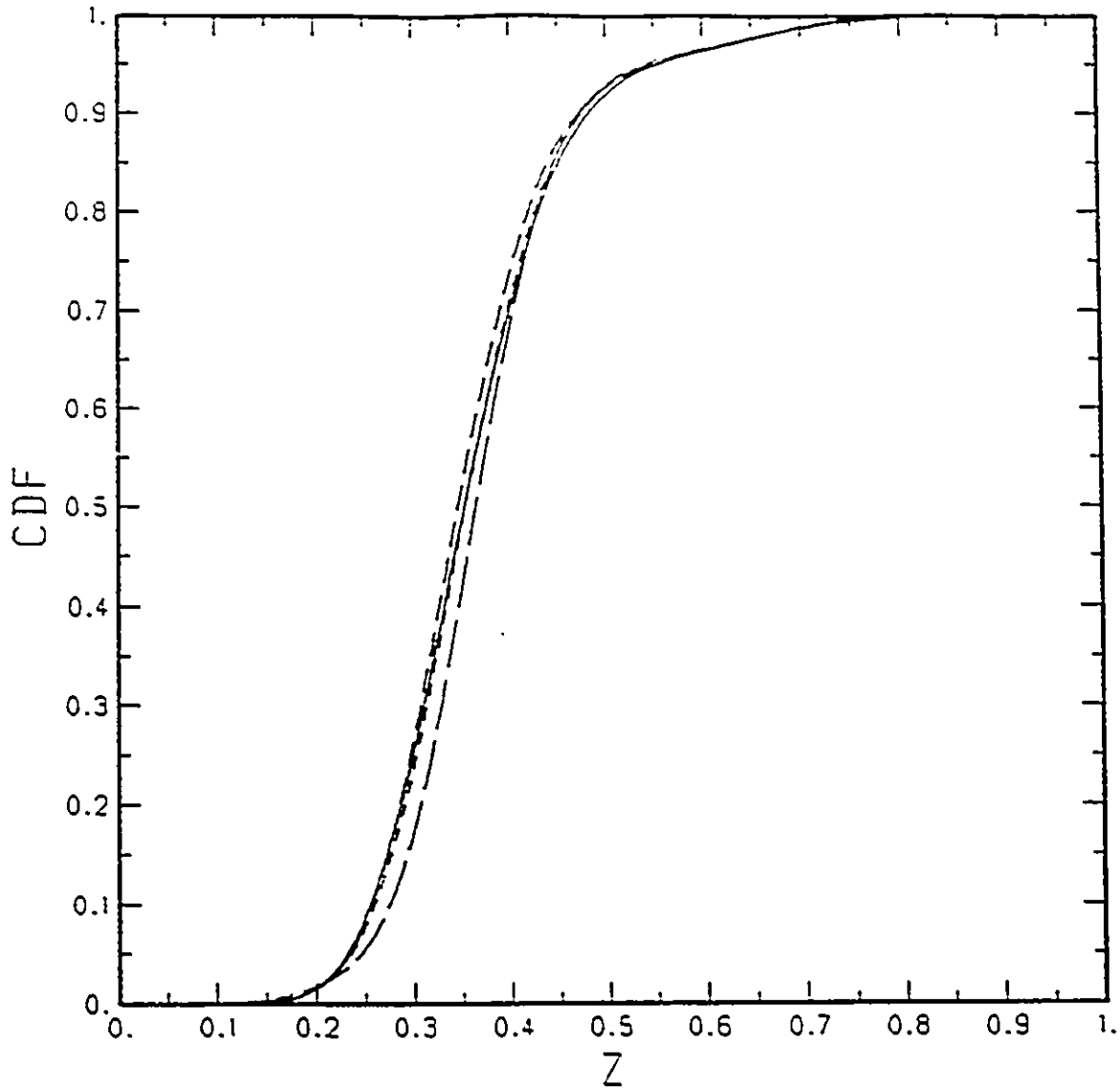
Figure 5.10: Distribution function of error of misclassification e_{12}: plot 10



= P = P1
= P = P2
= P = P3
= P = P4

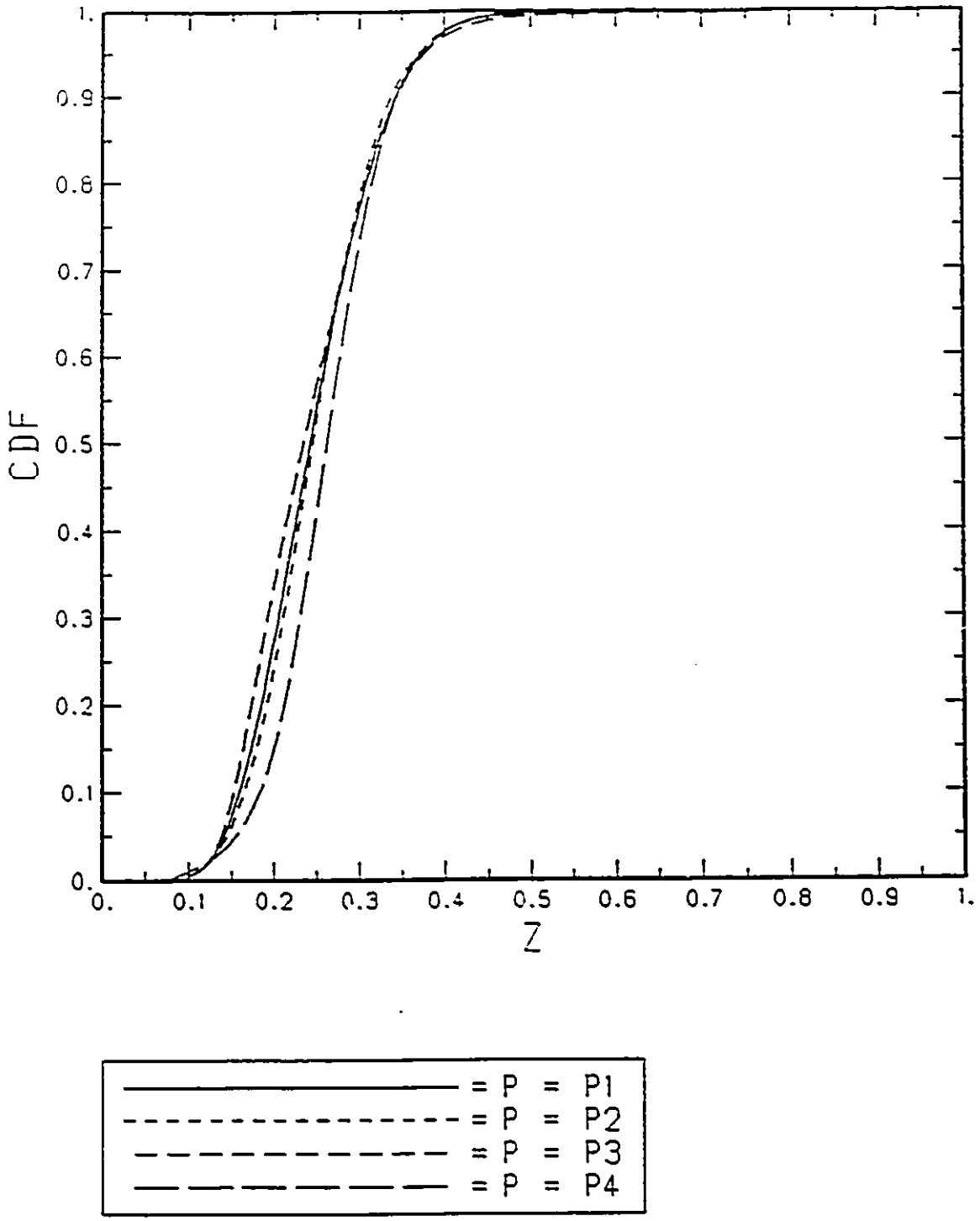
Model V
A1=A2=1 B1=B2=5
S1=2 S2=2

Figure 5.11: Distribution function of error of misclassification e_{12}: plot 11



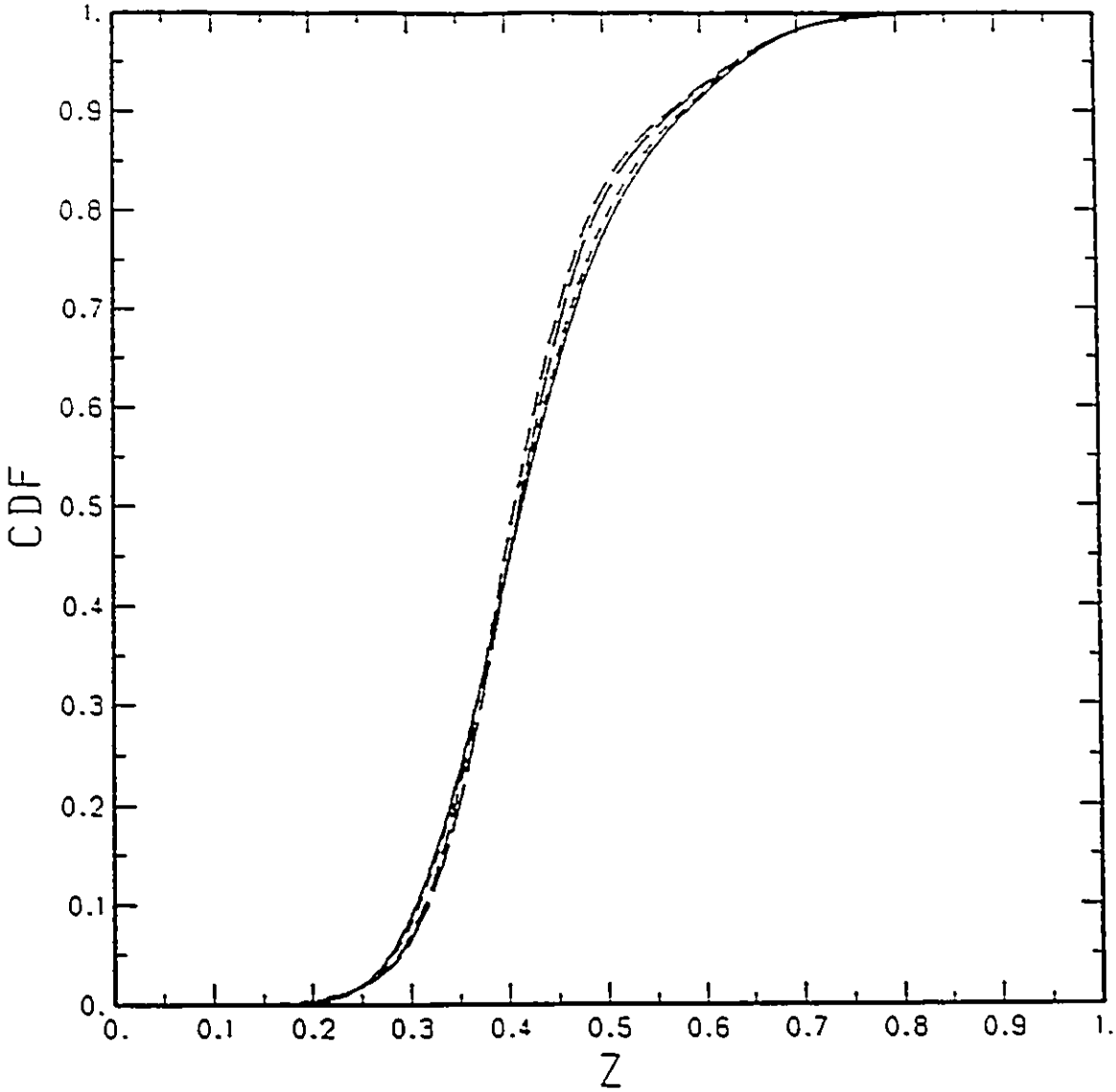
Model I
Mixture of Normals
A = 2.0 p = 0.9

Figure 5.12: Distribution function of error of misclassification c_{12}: plot 12



Model II
Mixture of Normals
A = 2.0 p = 0.9

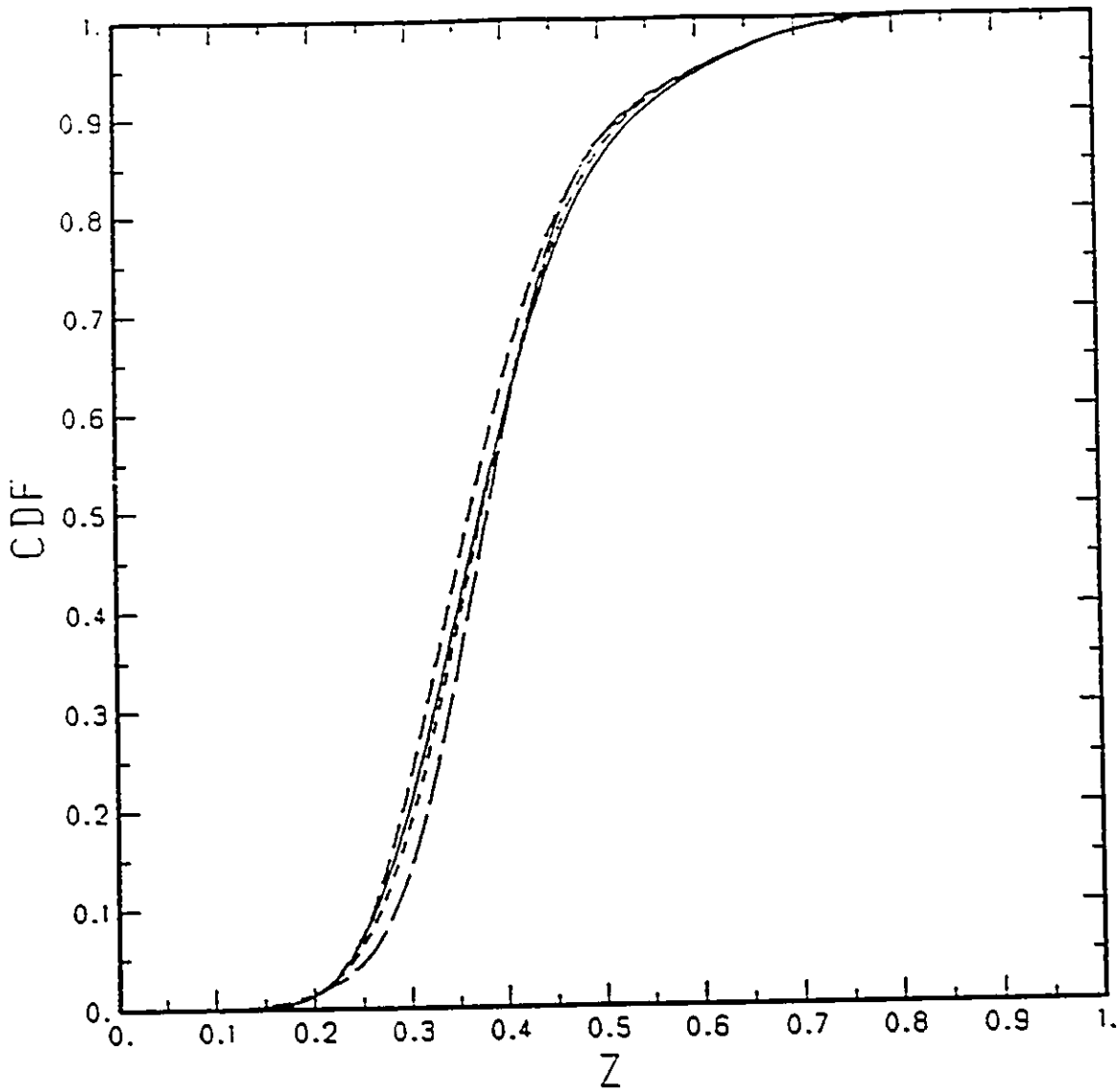
Figure 5.13: Distribution function of error of misclassification e_{12}: plot 13



= P = P1
= P = P2
= P = P3
= P = P4

Model III
Mixture of Normals
A = 2.0 p = 0.9

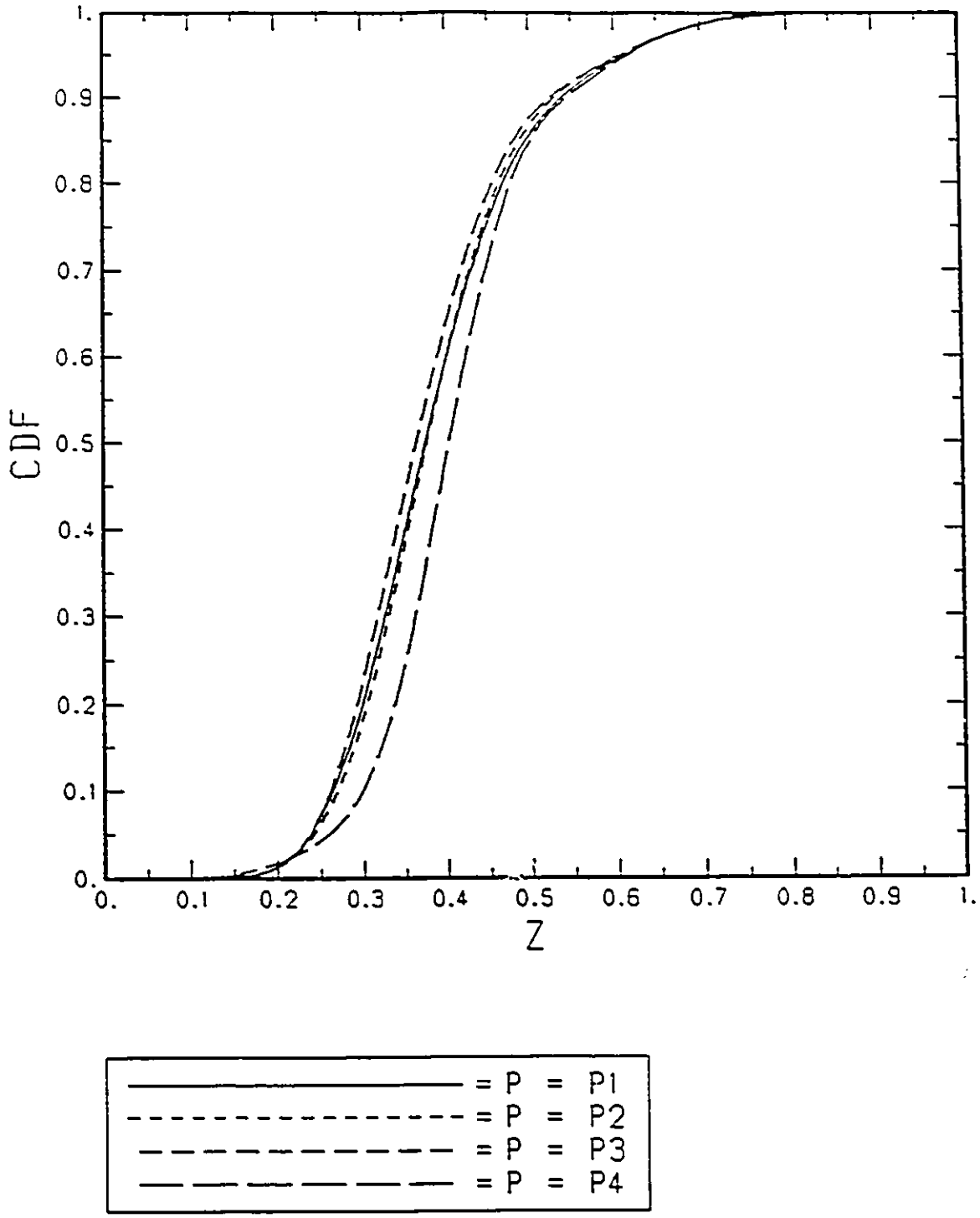
Figure 5.14: Distribution function of error of misclassification e_{12}: plot 14



= P = P1
= P = P2
= P = P3
= P = P4

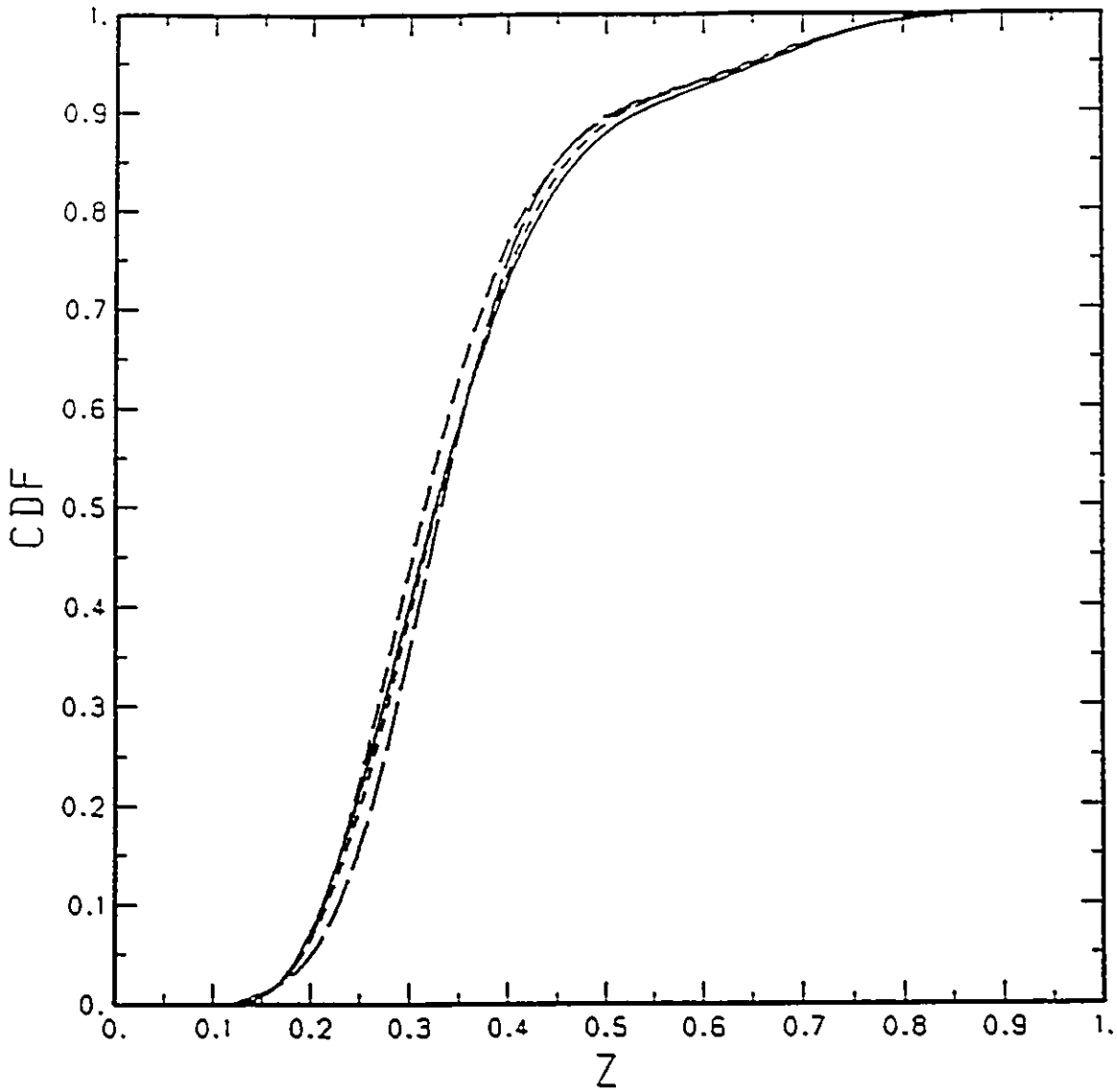
Model IV
Mixture of Normals
A = 2.0 p = 0.9

Figure 5.15: Distribution function of error of misclassification e_{12}: plot 15



Model V
Mixture of Normals
A = 2.0 p = 0.9

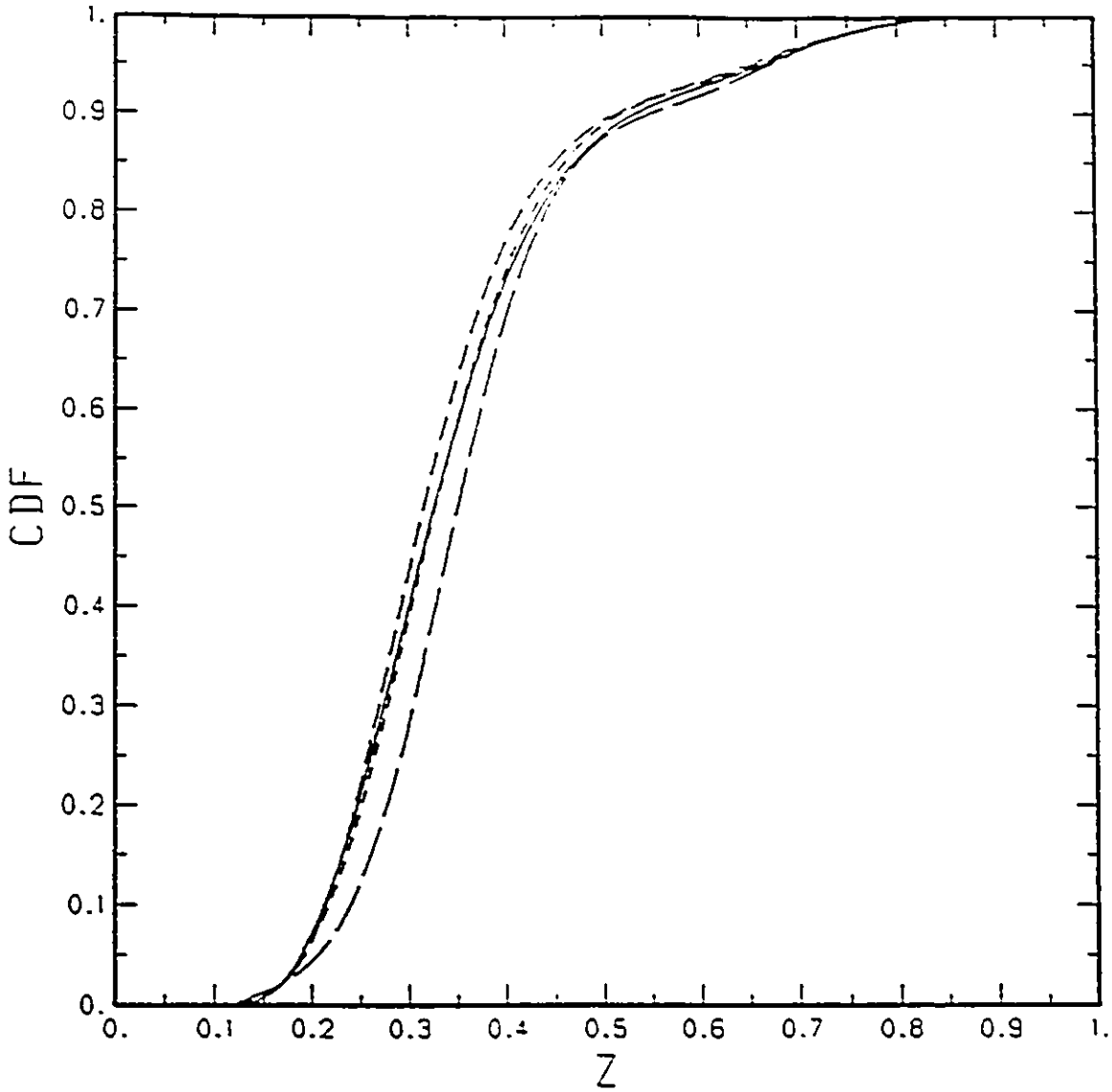
Figure 5.16: Distribution function of error of misclassification e_{12}: plot 16



= P = P1
= P = P2
= P = P3
= P = P4

Model IV
Mixture of Normals
A = 4.0 p = 0.9

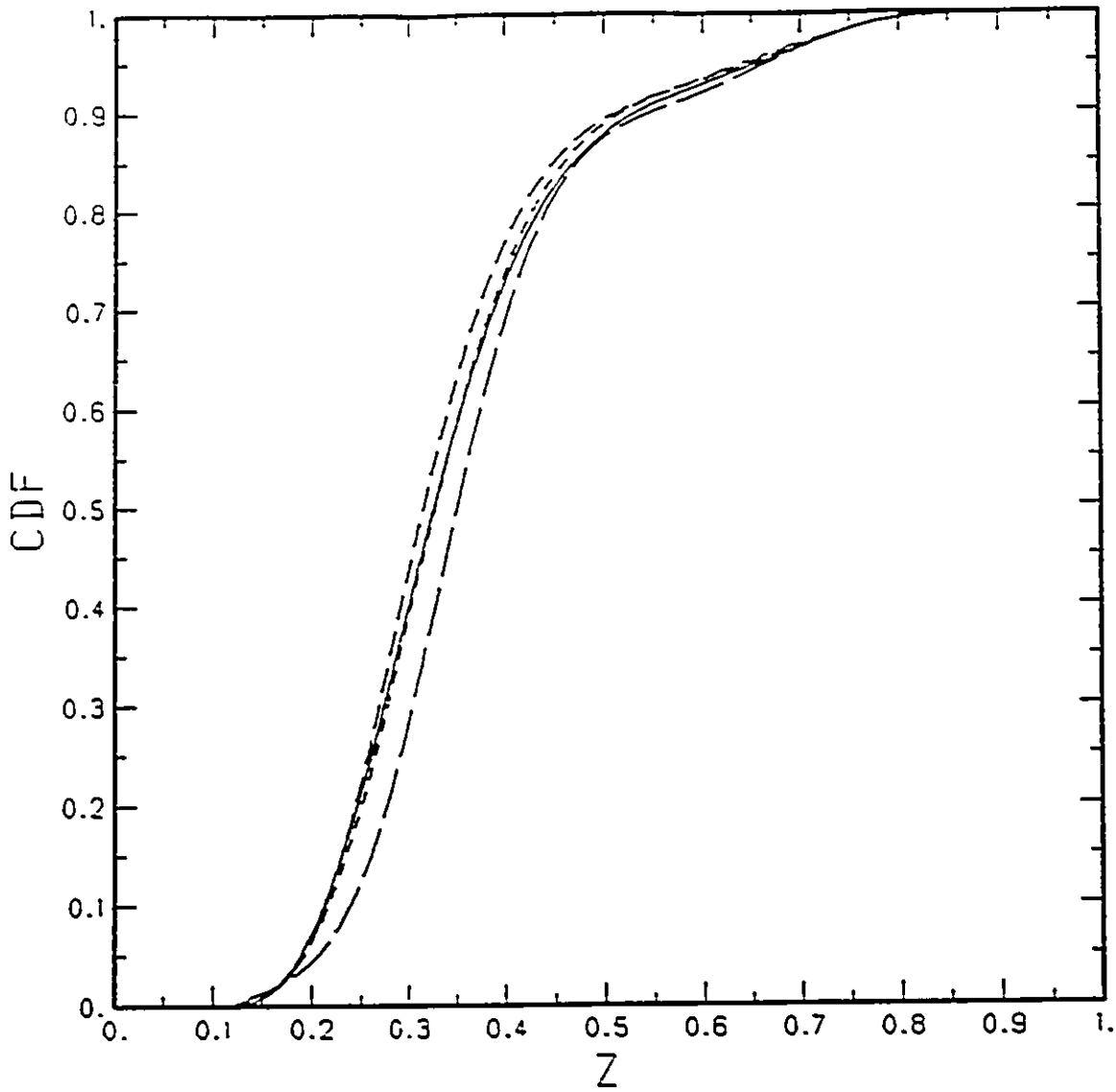
Figure 5.17: istribution function of error of misclassification e_{12}: plot 17



= P = P1
= P = P2
= P = P3
= P = P4

Model V
Mixture of Normals
A = 4.0 p = 0.9

Figure 5.18: Distribution function of error of misclassification e_{12}: plot 18



Model V
Mixture of Normals
A = 4.0 p = 0.9

Chapter 6

Univariate Classical and Robust Linear Classification Procedures Based On k Dichotomous and One Continuous Variables

6.1 Introduction

In the last chapter we considered the classification problem for data comprising of a k—-variate dichotomous variable and an associated continuous variable. We derived the linear classification procedure by assuming a normal distribution for the conditional distribution of the continuous variable and discussed some distributional properties of the two errors of misclassification for the case when both these errors are allowed to float (that is, neither is fixed at a specified level). By deriving similar results under the outlier-normal and mixture- normal models for the continuous variable and comparing them with the results for the normal case, we displayed that the classical linear classification procedure is sensitive to departures from normality. Hence, there is a need for proposing a robust procedure for this classification problem.

In Section 2, we present the classical linear classification procedure for this

problem when all the parameters in the model are known. In Section 3, we consider the case when all the parameters are unknown and present the classical linear classification procedure and discuss the asymptotic determination of the cut-off point of this procedure when one of the errors of misclassification is at a pre-fixed level. Next, in Section 4 we propose a robust linear classification procedure based on the univariate MML estimators (see Section 2) analogous to the classical procedure and explain the asymptotic determination of the cut-off point of this procedure when one of the two errors is at a pre-specified level. We also discuss some asymptotic optimal properties of this procedure. In Section 5, we derive the asymptotic formulas for the two errors of both the procedures and explain why the procedure based on the MML estimators will be robust to departures from normality while the classical procedure will be non-robust in nature. In Section 6, we make a comparison of the classical and the robust procedures for small samples using Monte Carlo simulations by considering normal and a wide variety of non-normal models. Through this comparative study, we show that the procedure based on the MML estimators proposed in this chapter is quite robust to departures from normality and also highly efficient as compared to the classical linear procedure. This work generalizes the results of Balakrishnan and Tiku (1988b) and Chang and Afifi (1974) which have been discussed briefly in Chapter 3. Finally, in Section 7 we consider an example with two dichotomous and one continuous variables and illustrate the two classification procedures developed and discussed in the earlier sections.

6.2 Classification when the population parameters are known

Let us consider two random variables $\mathbf{X}$ and Y where $\mathbf{X}$ is a $k-$ variate dichotomous variable with the probability mass function

$$P\{\mathbf{X} = \mathbf{x}\} = p_{\mathbf{x}},$$

where each element of x takes 0 or 1, and Y, conditional on $\mathbf{X} = \mathbf{x}$, has a normal distribution $N(\mu_{\mathbf{x}}, \sigma^2_{\mathbf{x}})$. We assume that

$$\mu_{\mathbf{x}} = \mu + \delta^T \mathbf{x}, \qquad \sigma^2_{\mathbf{x}} = \sigma^2 + \gamma^T \mathbf{x}.$$

Then the vector $\mathbf{W} = (\mathbf{X}, Y)^T$ has probability density function

$$f(\mathbf{w}) = p_{\mathbf{x}} \left(\frac{1}{\sqrt{2\pi}\sigma_{\mathbf{x}}} \exp - \frac{(y - \mu_{\mathbf{x}})^2}{2\sigma^2_{\mathbf{x}}} \right). \qquad x_i = 0, 1; \quad \text{for } i = 1, 2; \quad -\infty < y < \infty, \tag{6.1}$$

and the marginal probability density function of Y is

$$f(y) = \sum_{\mathbf{x}} f(\mathbf{w}). \tag{6.2}$$

The problem we discuss here is concerned with the classification of an observation $\mathbf{w}_0$ into one of the two populations Π_1 and Π_2, where Π_i has the pdf of the form

$$f_i(\mathbf{w}) = p^{(i)}_{\mathbf{x}} \phi(y; \mu^{(i)}_{\mathbf{x}}, \sigma^2_{\mathbf{x}}). \tag{6.3}$$

It should be noted that we are assuming here that the conditional variances of Y under Π_1 and Π_2 are equal. However, $Var(\mathbf{W}|\Pi_1)$ and $Var(\mathbf{W}|\Pi_2)$ are not equal.

The log-likelihood ratio gives the following procedure for the observation $\mathbf{w}_0$. Classify it as belonging to Π_1 or Π_2 according as

$$Z(\mathbf{x}_0) = \left[y_0 - \frac{\mu^{(1)}_{\mathbf{x}_0} + \mu^{(2)}_{\mathbf{x}_0}}{2} \right] \frac{\mu^{(1)}_{\mathbf{x}_0} - \mu^{(2)}_{\mathbf{x}_0}}{\sigma^2_{\mathbf{x}_0}} + \ln \left[\frac{p^{(1)}_{\mathbf{x}_0}}{p^{(2)}_{\mathbf{x}_0}} \right] \quad \gtrless C \tag{6.4}$$

and the errors of misclassification e_{12} and e_{21} are given by

$$e_{12} = \sum_{\mathbf{x}} p^{(1)}_{\mathbf{x}} \Phi \left\{ \frac{C - \frac{1}{2}\delta^2_{\mathbf{x}} - \log \left[\frac{p^{(1)}_{\mathbf{x}}}{p^{(2)}_{\mathbf{x}}} \right]}{|\delta_{\mathbf{x}}|} \right\} \tag{6.5}$$

and

$$e_{21} = \sum_{\mathbf{x}} p^{(2)}_{\mathbf{x}} \Phi \left\{ - \frac{C + \frac{1}{2}\delta^2_{\mathbf{x}} - \log \left[\frac{p^{(1)}_{\mathbf{x}}}{p^{(2)}_{\mathbf{x}}} \right]}{|\delta_{\mathbf{x}}|} \right\} \tag{6.6}$$

The cut-off point C of the classification procedure in (6.4) may be chosen by fixing one of the two errors given above at a specified level.

6.3 Classical classification procedure when the parameters are unknown

Consider independent random samples of sizes n_1, n_2 from each of the populations. Denote by $y_{\mathbf{x},j}^{(i)}$ the jth observation from the ith population of which the value of $\mathbf{X}$ is $\mathbf{x}$.

Let $n_i = \sum_{\mathbf{x}} n_{\mathbf{x}}^{(i)}$, where $n_{\mathbf{x}}^{(i)}$ is the number of observations in the ith sample with $\mathbf{X} = \mathbf{x}$, and

$$\begin{aligned} \bar{y}_{\mathbf{x}}^{(i)} &= \frac{1}{n_{\mathbf{x}}^{(i)}} \sum_{j=1}^{n_{\mathbf{x}}^{(i)}} y_{\mathbf{x},j}^{(i)} \\ s_{\mathbf{x}}^2 &= \frac{1}{n_{\mathbf{x}}^{(1)} + n_{\mathbf{x}}^{(2)} - 2} \sum_{i=1}^{2} \sum_{j=1}^{n_{\mathbf{x}}^{(i)}} (y_{\mathbf{x},j}^{(i)} - \bar{y}_{\mathbf{x}}^{(i)})^2. \end{aligned}$$

It can be shown that the MLE's of the parameters are, for $i = 1, 2$,

$$\hat{p}_{\mathbf{x}}^{(i)} = \frac{n_{\mathbf{x}}^{(i)}}{n_i}, \quad \hat{\mu}_{\mathbf{x}}^{(i)} = \bar{y}_{\mathbf{x}}^{(i)}, \quad \hat{\sigma}_{\mathbf{x}}^2 = s_{\mathbf{x}}^2 \ \text{(corrected for bias)}.$$

By substituting the ML estimators for parameters in (6.4) we obtain the following rule which is a direct extension of the procedure given by Chang and Afifi (1974) (also see Chapter 3). This procedure classifies an observation $\mathbf{w}_0$ as belonging to Π_1 or Π_2 according as

$$\hat{Z}(\mathbf{x}_0) = \left[y_0 - \frac{\bar{y}_{\mathbf{x}_0}^{(1)} + \bar{y}_{\mathbf{x}_0}^{(2)}}{2} \right] \frac{\bar{y}_{\mathbf{x}_0}^{(1)} - \bar{y}_{\mathbf{x}_0}^{(2)}}{s_{\mathbf{x}_0}^2} + \ln \left[\frac{n_{\mathbf{x}_0}^{(1)} n_2}{n_{\mathbf{x}_0}^{(2)} n_1} \right] \gtrless C, \tag{6.7}$$

where C is determined by fixing one of the two errors of misclassification at a specified level. An estimate $\hat{C}$ of C can be obtained by substituting estimators of the parameters in (6.5) and fixing e_{12} to be α, i.e. the estimate of the cut-off point C may be obtained as the solution $\hat{C}$ of the equation (with e_{12} fixed as α)

$$\sum_{\mathbf{x}} \hat{p}_{\mathbf{x}}^{(1)} \Phi \left\{ \frac{\hat{C} - \frac{1}{2}\hat{\eta}^2(\mathbf{x}) - \ln \left[\frac{n_{\mathbf{x}}^{(1)} n_2}{n_{\mathbf{x}}^{(2)} n_1} \right]}{|\hat{\eta}(\mathbf{x})|} \right\} = \alpha, \tag{6.8}$$

where

$$\hat{\eta}^2(\mathrm{x}) = \left(\frac{\bar{y}_{\mathrm{x}}^{(1)} - \bar{y}_{\mathrm{x}}^{(2)}}{s_{\mathrm{x}}}\right)^2 \left(\frac{n_{\mathrm{x}}^{(1)} + n_{\mathrm{x}}^{(2)} - 4}{n_{\mathrm{x}}^{(1)} + n_{\mathrm{x}}^{(2)} - 2}\right).$$

is the Mahalanobis' sample squared distance. We may note that $\hat{\eta}^2(x)$ but guarantees an admissible non-negative estimate for $\eta^2(x)$.

6.4 Robust classification procedure when the parameters are unknown

In this section, we make use of the univariate MML estimators described in Chapter 2 to propose a robust linear classification procedure analogous to the classical linear classification procedure presented in the last section. We discuss the asymptotic determination of the cut-off point of this procedure when one of the errors of misclassification is at a pre-fixed level and also some asymptotic optimality properties of this procedure.

By using the univariate MML estimators in place of the parameters in the classification rule in (6.4), we obtain a classification procedure which classifies an observation w_0 into Π_1 or Π_2 according as

$$\hat{Z}_R(\mathrm{x}_0) = \left[y - \frac{\hat{\mu}_{\mathrm{x}_0}^{(1)} + \hat{\mu}_{\mathrm{x}_0}^{(2)}}{2}\right] \frac{\hat{\mu}_{\mathrm{x}_0}^{(1)} - \hat{\mu}_{\mathrm{x}_0}^{(2)}}{\hat{\sigma}_{\mathrm{x}_0}^2} + \ln\left[\frac{n_{\mathrm{x}_0}^{(1)} n_2}{n_{\mathrm{x}_0}^{(2)} n_1}\right] \gtrless C_R, \tag{6.9}$$

where $\hat{\mu}_{\mathrm{x}}^{(i)}$ is the MML estimator of the ith population mean, i=1,2; $\hat{\sigma}_{\mathrm{x}}^2$ is the pooled MML estimator of σ_{x}^2. An estimate $\hat{C}_R$ of C_R can be obtained by substituting the MML estimators for the parameters in (6.5). i.e. the estimate of the cut-off point C_R may be obtained as the solution $\hat{C}_R$ of the equation (with e_{12} fixed as α)

$$\sum_{\mathrm{x}} \hat{p}_{\mathrm{x}}^{(1)} \Phi \left\{ \frac{\hat{C}_R - \frac{1}{2}\hat{\eta}_R^2(\mathrm{x}) - \ln\left[\frac{n_{\mathrm{x}}^{(1)} n_2}{n_{\mathrm{x}}^{(2)} n_1}\right]}{|\hat{\eta}_R(\mathrm{x})|} \right\} = \alpha, \tag{6.10}$$

where

$$\hat{\eta}_R^2(\mathrm{x}) = \left(\frac{\hat{\mu}_{\mathrm{x}}^{(1)} - \hat{\mu}_{\mathrm{x}}^{(2)}}{\hat{\sigma}_{\mathrm{x}}}\right)^2 \left(\frac{A_{\mathrm{x}}^{(1)} + A_{\mathrm{x}}^{(2)} - 4}{A_{\mathrm{x}}^{(1)} + A_{\mathrm{x}}^{(2)} - 2}\right),$$

$$A_{\mathbf{x}}^{(i)} = n_{\mathbf{x}}^{(i)} - 2r_{\mathbf{x}}^{(i)},$$

$r_{\mathbf{x}}^{(i)}$ being the number of observations censored for robustness purposes.

We present in the following two theorems some asymptotic optimal properties of the robust procedure based on the MML estimators that is presented in this section.

Theorem 11 *The statistic $\hat{Z}(\mathbf{x}_0)$ in (6.7) and $\hat{Z}_R(\mathbf{x}_0)$ in (6.9) asymptotically has exactly the same distribution as the statistic $Z(\mathbf{x}_0)$ in (6.7))*

Proof: This theorem follows from the fact that when n_1 and n_2 tend to infinity both ML estimators and MML estimators tend to their population parameters. i.e. $\bar{y}_{\mathbf{x}}^{(i)}$, $\hat{\mu}_{\mathbf{x}}^{(i)}$ both tends to $\mu_{\mathbf{x}}^{(i)}$ and $s_{\mathbf{x}}$, $\hat{\sigma}_{\mathbf{x}}$ both tends to $\sigma_{\mathbf{x}}$. Also $\hat{p}_{\mathbf{x}}^{(i)}$ tends to $p_{\mathbf{x}}^{(i)}$. Hence the proof. ∎

Theorem 12 *Asymptotically $\hat{Z}(\mathbf{x}_0)$ and $\hat{Z}_R(\mathbf{x}_0)$ procedures have exactly the same e_{21} of e_{12} values, under normality.*

Proof: The result follows from Theorem (11) and the fact that $\hat{C}$ and $\hat{C}_R$ both converge to C given by (6.5) as n_1 and n_2 become large, since $\frac{n_{\mathbf{x}}^{(i)}}{n_i}$ converges to $p_{\mathbf{x}}^{(i)}$, $\bar{y}_{\mathbf{x}}^{(i)}$ and $\hat{\mu}_{\mathbf{x}}^{(i)}$ both converge to $\mu_{\mathbf{x}}^{(i)}$, $i = 1, 2$ $(x = 0, 1)$, and $s_{\mathbf{x}}^2$ and $\hat{\sigma}_{\mathbf{x}}$ both converge to $\sigma_{\mathbf{x}}^2$.

6.5 Asymptotic Values of e_{12} and e_{21}

As n_i become large, $\frac{n_{\mathbf{x}}^{(i)}}{n_i}$, $y_{\mathbf{x}}^{(i)}$ and $s_{\mathbf{x}}$ converge to their expected values $p_{\mathbf{x}}^{(i)}$, $\mu_{\mathbf{x}}^{(i)}$ and $\sigma_{\mathbf{x}}$, respectively. In that case, $\hat{Z}(\mathbf{x})$ is a linear function of $y_{\mathbf{x}_0}$. The asymptotic distribution of $\hat{Z}(\mathbf{x})$ given $\mathbf{X} = \mathbf{x}$ is, therefore, similar to the distribution of $y_{\mathbf{x}_0}$, with a different mean and variance. Moreover, $\hat{C}$ converges to C as n_1 and n_2 become large. If we define

$$Z_1 = \frac{\hat{Z}(\mathbf{x}) - E(\hat{Z}(\mathbf{x}))}{\sqrt{Var(\hat{Z}(\mathbf{x}))}} \tag{6.11}$$

so that $E(Z_1) = 0$ and $Var(Z_1) = 1$. Therefore, the asymptotic values of e_{12} and e_{21} associated with $\hat{Z}(\mathrm{x})$ are given by

$$\begin{aligned} e_{12} &= \sum_{\mathrm{x}} p_{\mathrm{x}}^{(1)} \Pr[\hat{Z}(\mathrm{x}) < C | \mathrm{w}_0 \in \Pi_1] \\ &= \sum_{\mathrm{x}} p_{\mathrm{x}}^{(1)} \Pr\left\{ Z_1 < \frac{C - \frac{1}{2}\eta^2(\mathrm{x}) - \ln\left[\frac{p_{\mathrm{x}}^{(1)}}{p_{\mathrm{x}}^{(2)}}\right]}{|\eta(\mathrm{x})|} \right\} \end{aligned} \tag{6.12}$$

and

$$\begin{aligned} e_{21} &= \sum_{\mathrm{x}} p_{\mathrm{x}}^{(2)} \Pr[\hat{Z}(\mathrm{x}) > C | \mathrm{w}_0 \in \Pi_1] \\ &= \sum_{\mathrm{x}} p_{\mathrm{x}}^{(2)} \Pr\left\{ Z_1 > \frac{C - \frac{1}{2}\eta^2(\mathrm{x}) - \ln\left[\frac{p_{\mathrm{x}}^{(1)}}{p_{\mathrm{x}}^{(2)}}\right]}{|\eta(\mathrm{x})|} \right\} \end{aligned} \tag{6.13}$$

The non-robust character of the $\hat{Z}(\mathrm{x})$ procedure is explained by equations (6.12) and (6.13). Since $p_{\mathrm{x}}^{(i)}$, $\eta(\mathrm{x})$ and C are fixed, the values of e_{12} and e_{21} given by (6.12) and (6.13) will clearly change with the distribution of Z_1 (or y_{x_0}). Consider now the $\hat{Z}_R(x)$ procedure. If y_{x_0} has a symmetric distribution, the $\hat{\mu}_{\mathrm{x}}^{(i)}$ and $\hat{s}_{\mathrm{x}}$ converge to their expected values $\mu_{\mathrm{x}}^{(i)}$ and $h\sigma_{\mathrm{x}}$, respectively, where $h = \frac{E(\hat{\sigma}_{\mathrm{x}})}{\sigma_{\mathrm{x}}}$; see Tiku (1980,1982). In that case $\hat{Z}_R(\mathrm{x})$, like $\hat{Z}(\mathrm{x})$, is a linear function of y_{x_0}. Thus, for the $\hat{Z}_R(\mathrm{x})$ procedure we have

$$E(\hat{Z}_R(\mathrm{x})) = (-1)^{i-1}\frac{\eta^2(\mathrm{x})}{2h^2} + \ln\left[\frac{p_{\mathrm{x}}^{(1)}}{p_{\mathrm{x}}^{(2)}}\right] \quad y_0 \in \Pi_i \ (i = 1, 2) \tag{6.14}$$

and

$$Var(Z_R(\mathrm{x})) = \frac{\eta^2(\mathrm{x})}{h^4} \tag{6.15}$$

Define

$$Z_2 = \frac{\hat{Z}_R(\mathrm{x}) - E(\hat{Z}_R(\mathrm{x}))}{\sqrt{Var(\hat{Z}_R(\mathrm{x}))}} \tag{6.16}$$

The asymptotic distribution of Z_2 is exactly the same as that of Z_1 since both are linear in y_{x_0} (asymptotically). Consequently, the asymptotic e_{12} and e_{21} values of

$\hat{Z}_R(\mathrm{x})$ procedure are given by

$$\begin{aligned} e_{12} &= \sum_{\mathrm{x}} p_{\mathrm{x}}^{(1)} \Pr[\hat{Z}_R(\mathrm{x}) < C^* | \mathrm{w}_0 \in \Pi_1] \\ &= \sum_{\mathrm{x}} p_{\mathrm{x}}^{(1)} \Pr\left\{ Z_2 < \frac{C^* - \frac{1}{2h^2}\eta^2(\mathrm{x}) - \ln\left[\frac{p_{\mathrm{x}}^{(1)}}{p_{\mathrm{x}}^{(2)}}\right]}{\frac{|\eta(\mathrm{x})|}{h^2}} \right\} \end{aligned} \tag{6.17}$$

and

$$\begin{aligned} e_{21} &= \sum_{\mathrm{x}} p_{\mathrm{x}}^{(2)} \Pr[\hat{Z}_R(\mathrm{x}) > C^* | \mathrm{w}_0 \in \Pi_1] \\ &= \sum_{\mathrm{x}} p_{\mathrm{x}}^{(2)} \Pr\left\{ Z_2 > \frac{C^* - \frac{1}{2h^2}\eta^2(\mathrm{x}) - \ln\left(\frac{p_{\mathrm{x}}^{(1)}}{p_{\mathrm{x}}^{(2)}}\right)}{\frac{|\eta(\mathrm{x})|}{h^2}} \right\} \end{aligned} \tag{6.18}$$

Since h assumes different values for different distributions, the cutoff points, given in (6.17) or (6.17), change with the distribution of Z_2 (or y_0). Suppose that Z_2 has the following symmetric family of distributions with probability density function

$$f(z) \propto \left[1 + \frac{z^2}{k}\right]^{-p} \quad -\infty < z < \infty \quad (k = 2p - 3, p \geq 2) \tag{6.19}$$

This family represents a wide class of non-normal distributions with kurtosis ranging between 3 ($p = \infty$) and ∞ (p=2) and it has been extensively used in robustness studies; see Tiku , Tan and Balakrishnan (1986). For the family (6.19) the values of h are also given in Tiku, Tan and Balakrishnan (1986). Using these, the values of e_{12} and e_{21} can be calculated from (6.17) and (6.18).

6.6 Comparison for small samples

It is very difficult to evaluate the e_{12} and e_{21} values of the procedure described in Section 5 analytically for small sample sizes n_1 and n_2. However, to compare the $\hat{Z}(\mathrm{x})$ and $\hat{Z}_R(\mathrm{x})$ procedures, we carried out an extensive Monte Carlo investigation and simulated their e_{12} and e_{21} values for the following distributions which represent

a wide class of symmetric (normal and non-normal) models:

(1) Normal $N(\mu, \sigma^2)$

(2) $y_1, y_2, \ldots, y_{n-2}$ come from $N(\mu, \sigma^2)$ and y_{n-1}, y_n from $N(\mu, (10\sigma)^2)$

(3) $y_1, y_2, \ldots, y_{n-6}$ come from $N(\mu, \sigma^2)$ and $y_{n-5}, \ldots, y_n$ from $N(\mu, (4\sigma)^2)$

(4) $y_1, y_2, \ldots, y_{n-6}$ come from $N(\mu, \sigma^2)$ and $y_{n-5}, \ldots, y_n$ from $N(\mu, (10\sigma)^2)$

(5) Logistic: $\frac{\exp - \frac{y-\mu}{\sigma}}{\sigma\left(1+\exp - \frac{y-\mu}{\sigma}\right)^2}, \quad -\infty < y < \infty,$

(6) Double exponential: $\frac{1}{2\sigma} \exp - \frac{|y-\mu|}{\sigma}, \quad -\infty < y\infty,$

(7) Mixture model: $0.90N(\mu, \sigma^2) + 0.10N(\mu, (4\sigma)^2)$

(8) Mixture model: $0.90N(\mu, \sigma^2) + 0.10N(\mu, (10\sigma)^2)$

and

(9) Student's t_4 : $\frac{3}{4\sigma\sqrt{\pi}\left(1+\left(\frac{y-\mu}{\sigma}\right)^2\right)^{5/2}}, \quad -\infty < y < \infty$

(6.20)

Then, with e_{12} fixed as 0.05, the values of e_{12} and $1 - e_{21}$ were simulated (based on 1,000 Monte Carlo runs) for both the procedures under all 9 models given in (6.20). These simulated values are presented in Table 6.1 for

$$\begin{array}{llllll}
\mathbf{a} & \mu_1 = 0 & \mu_2 = 1 & \sigma^2 = 1 & \underline{\delta}' = (0.25, 0.25) & \underline{\gamma}' = (1,1) \\
\mathbf{b} & \mu_1 = 0 & \mu_2 = 2 & \sigma^2 = 1 & \underline{\delta}' = (0.25, 0.25) & \underline{\gamma}' = (1,1) \\
\mathbf{c} & \mu_1 = 0 & \mu_2 = 1 & \sigma^2 = 4 & \underline{\delta}' = (0.25, 0.25) & \underline{\gamma}' = (1,1) \ , \\
\mathbf{d} & \mu_1 = 0 & \mu_2 = 1 & \sigma^2 = 1 & \underline{\delta}' = (0.25, 0.25) & \underline{\gamma}' = (4,1) \\
\mathbf{e} & \mu_1 = 0 & \mu_2 = 1 & \sigma^2 = 1 & \underline{\delta}' = (0.25, 0.25) & \underline{\gamma}' = (1,4)
\end{array}$$

$$P = \begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 0.25 \end{bmatrix}$$

and $n_1 = n_2 = 100$. It is quite clear from Table 6.1 that the robust linear procedure has more stable e_{12} values than the corresponding classical linear procedure. It is also very clear from Table 6.1 that the robust linear classification procedure has considerably larger $1 - e_{21}$ values than the corresponding classical linear classification procedure under all the non-normal models in (6.20). Thus, we see that the robust linear classification procedure proposed in Section 4 is very efficient and quite robust to departures from normality.

Table 6.1: Simulated values of e_{12} and $1 - e_{21}$ for the classical and the robust linear classification procedures for the models in (6.20).

		e_{12} values					$1 - e_{21}$ values				
Model	Procedure	a	b	c	d	e	a	b	c	d	e
(1)	$\hat{Z}$	.054	.059	.054	.058	.049	.195	.434	.113	.159	.168
	$\hat{Z}_R$	.066	.063	.068	.067	.068	.209	.442	.136	.187	.187
(2)	$\hat{Z}$	.043	.040	.034	.030	.038	.127	.275	.068	.098	.117
	$\hat{Z}_R$	.063	.058	.064	.064	.064	.202	.424	.122	.175	.176
(3)	$\hat{Z}$	.042	.031	.034	.034	.033	.131	.305	.068	.099	.113
	$\hat{Z}_R$	.062	.052	.060	.063	.060	.187	.408	.113	.167	.164
(4)	$\hat{Z}$	.028	.015	.022	.016	.016	.077	.160	.034	.055	.073
	$\hat{Z}_R$	.060	.049	.059	.063	.058	.183	.396	.111	.160	.157
(5)	$\hat{Z}$	.054	.058	.052	.053	.048	.186	.436	.105	.153	.160
	$\hat{Z}_R$	.070	.068	.068	.071	.068	.216	.470	.141	.104	.201
(6)	$\hat{Z}$	.048	.055	.056	.052	.061	.164	.450	.101	.161	.145
	$\hat{Z}_R$	.070	.074	.089	.077	.086	.229	.565	.155	.206	.205
(7)	$\hat{Z}$	.038	.038	.040	.044	.045	.106	.224	.063	.089	.096
	$\hat{Z}_R$	.072	.073	.079	.067	.081	.192	.363	.125	.159	.165
(8)	$\hat{Z}$	.031	.032	.024	.031	.029	.044	.081	.029	.039	.036
	$\hat{Z}_R$	.076	.072	.081	.069	.079	.173	.327	.123	.150	.152
(9)	$\hat{Z}$	.056	.068	.056	.068	.053	.285	.473	.112	.179	.169
	$\hat{Z}_R$	.062	.071	.072	.079	.064	.196	.470	.132	.185	.179

6.7 Illustrative example

In this section we consider a data set from the following situation: At a certain university, Course A can be taken

1. if a student has credit in at least one of Course B and C or
2. with the instructor's permission

There are two instructors for this course and group 1 denotes the marks of students from first instructor and group 2 denotes the marks of students from the second instructor. The marks were recorded in the following form:

1. $\mathbf{X}^T = (0,0)$ if the student does not have credits for both courses B and C.
2. $\mathbf{X}^T = (0,1)$ if the student has credits for course C but not for course B
3. $\mathbf{X}^T = (1,0)$ if the student has credits for course B but not for course C
4. $\mathbf{X}^T = (1,1)$ if the student has credits for both courses B and C.
5. Y denotes the raw marks of the student in course A.

These data are presented in Tables 6.2. and Table 6.3. From these data we obtain the following quantities:

$\mathbf{x}^T$	$\bar{y}_{\mathbf{x}}^{(1)}$	$\bar{y}_{\mathbf{x}}^{(2)}$	$\hat{s}_{\mathbf{x}}^{(1)}$	$\hat{s}_{\mathbf{x}}^{(2)}$	$\hat{s}_{\mathbf{x}}$
0 0	60.52241	64.96532	7.38649	8.74589	7.87382
1 0	64.25967	71.19374	25.22479	15.49498	21.44097
0 1	64.14256	70.99187	19.38939	16.55971	18.18892
1 1	69.23428	75.15727	21.43854	17.87649	19.97181

$\mathbf{x}^T$	$\hat{\mu}_{\mathbf{x}}^{(1)}$	$\hat{\mu}_{\mathbf{x}}^{(2)}$	$\hat{\sigma}_{\mathbf{x}}^{(1)}$	$\hat{\sigma}_{\mathbf{x}}^{(2)}$	$\hat{\sigma}_{\mathbf{x}}$
0 0	60.60252	64.94766	6.75722	6.95518	6.82965
1 0	64.11650	71.13015	22.00745	14.25641	18.96240
0 1	64.50060	70.93756	14.58025	12.21315	13.57878
1 1	69.44048	74.57758	14.18861	10.81115	12.78820

x^T		$\hat{p}_x^{(1)}$	$\hat{p}_x^{(2)}$
0	0	0.23333	0.13333
1	0	0.30000	0.19333
0	1	0.26000	0.19333
1	1	0.20667	0.14667

From these values we obtain $\hat{C}$ to be -0.91472 and $\hat{C}_R$ to be -0.91658. So, with e_{12} (the probability of wrongly classifying an individual from group 1 into group 2) fixed as 0.05, by applying the classical procedure in (6.7) and reclassifying each individual in Tables 6.2 and 6.3, we obtain estimates of e_{12} and $1 - e_{21}$ (the probability of correctly classifying an individual from group 2 into itself) to be 0.060 and 0.580, respectively. Similarly, by applying the robust linear classification procedure in (6.9) and reclassifying each individual in Tables 6.2 and 6.3, we obtain estimates of e_{12} and $1 - e_{21}$ to be 0.073 and 0.630, respectively.

Table 6.2: Marks records of group 1

i	$\mathbf{x}^T$		y	i	$\mathbf{x}^T$		y	i	$\mathbf{x}^T$		y	i	$\mathbf{x}^T$		y	i	$\mathbf{x}^T$		y
1	0	1	59	31	1	0	65	61	1	1	70	91	1	1	58	121	1	0	62
2	1	1	57	32	1	0	65	62	1	0	56	92	1	0	55	122	0	0	62
3	1	1	71	33	1	1	66	63	0	0	64	93	0	1	64	123	0	0	63
4	1	0	64	34	0	1	69	64	1	0	63	94	0	0	60	124	1	1	71
5	0	1	66	35	1	1	70	65	0	0	63	95	1	1	71	125	1	0	66
6	0	0	59	36	1	0	63	66	0	1	67	96	1	1	67	126	1	0	66
7	0	1	62	37	0	0	60	67	0	1	64	97	1	0	71	127	1	1	74
8	0	0	62	38	1	0	64	68	1	1	78	98	0	1	68	128	1	0	70
9	0	0	62	39	0	1	66	69	0	0	55	99	1	0	62	129	0	0	60
10	1	0	64	40	0	0	56	70	0	0	63	100	0	1	58	130	1	0	66
11	0	0	61	41	1	0	69	71	0	0	61	101	1	0	57	131	0	0	58
12	0	1	65	42	0	1	67	72	0	0	58	102	1	1	70	132	1	0	57
13	0	1	63	43	1	0	59	73	0	1	58	103	0	0	61	133	1	0	57
14	1	0	66	44	0	1	64	74	0	0	57	104	1	0	63	134	1	0	64
15	1	0	68	45	0	1	64	75	0	1	70	105	1	0	61	135	0	1	66
16	0	0	64	46	1	0	69	76	0	1	64	106	0	1	66	136	0	0	58
17	1	1	64	47	0	0	60	77	1	0	64	107	1	0	64	137	0	1	64
18	0	1	65	48	0	1	61	78	1	1	67	108	0	1	62	138	1	1	71
19	0	0	59	49	1	1	72	79	1	0	62	109	0	1	66	139	0	0	54
20	0	0	63	50	1	1	67	80	0	1	64	110	1	1	71	140	1	0	63
21	1	1	74	51	0	0	57	81	1	0	72	111	1	0	62	141	0	0	60
22	1	1	63	52	1	0	59	82	1	0	63	112	1	1	73	142	1	1	78
23	0	0	59	53	0	0	61	83	0	1	63	113	0	0	61	143	0	1	68
24	0	0	59	54	1	0	62	84	1	1	71	114	0	1	68	144	0	0	59
25	0	1	65	55	1	0	68	85	1	1	69	115	0	0	67	145	1	1	64
26	1	1	68	56	0	1	64	86	1	1	68	116	1	0	63	146	0	1	60
27	0	0	60	57	1	1	68	87	1	0	64	117	0	1	61	147	1	0	71
28	0	1	67	58	1	0	57	88	0	0	55	118	1	1	65	148	0	1	45
29	1	1	70	59	1	0	66	89	0	1	69	119	1	0	67	149	1	0	80
30	0	1	63	60	0	1	59	90	0	1	59	120	1	1	69	150	1	0	52

Table 6.3: Marks records of group 1

i	$\mathbf{x}^T$		y	i	$\mathbf{x}^T$		y	i	$\mathbf{x}^T$		y	i	$\mathbf{x}^T$		y	i	$\mathbf{x}^T$		y
1	1	1	74	21	1	0	73	41	0	1	71	61	0	0	66	81	0	1	66
2	1	0	76	22	1	0	73	42	1	1	77	62	0	1	70	82	0	1	73
3	1	0	75	23	0	1	72	43	0	1	76	63	0	0	64	83	0	1	79
4	0	1	71	24	0	1	73	44	0	0	63	64	1	0	72	84	0	0	64
5	1	1	70	25	1	0	69	45	1	0	69	65	0	0	64	85	1	1	75
6	0	0	64	26	1	1	72	46	1	1	77	66	0	0	70	86	1	0	66
7	0	1	67	27	1	0	73	47	0	0	61	67	1	0	67	87	1	0	79
8	1	0	72	28	0	1	68	48	1	1	73	68	1	1	76	88	1	1	76
9	1	1	77	29	0	0	66	49	0	0	58	69	1	1	83	89	0	1	70
10	1	1	70	30	1	1	71	50	1	0	72	70	0	0	64	90	1	0	73
11	1	0	61	31	1	0	72	51	0	1	72	71	0	1	65	91	0	0	68
12	0	1	69	32	0	1	70	52	0	0	63	72	1	1	86	92	1	1	76
13	0	0	63	33	0	1	69	53	0	1	74	73	1	0	70	93	1	0	67
14	0	1	62	34	0	1	75	54	0	0	66	74	1	0	74	94	1	0	69
15	1	0	72	35	1	0	70	55	0	0	66	75	0	1	72	95	0	0	60
16	0	1	80	36	1	1	72	56	1	0	67	76	1	1	77	96	0	1	69
17	1	0	68	37	1	1	75	57	1	1	69	77	0	1	69	97	0	1	70
18	0	1	69	38	0	1	67	58	1	1	74	78	1	1	74	98	0	1	63
19	1	0	66	39	0	0	66	59	0	0	65	79	1	1	69	99	1	0	66
20	0	0	70	40	1	0	78	60	1	0	70	80	0	1	72	100	1	0	74

Chapter 7

Multivariate Classical and Robust Linear Classification Procedures Based On k–Dichotomous and Continuous Variables

7.1 Introduction

In Chapter 3, we discussed the classical and the robust linear classification procedures based on an univariate dichotomous variable and an associated univariate normal variable. These results were generalized in Chapter 4 to the case where the data contains an univariate dichotomous variable followed by a multivariate normal variable. For this situation, the classical and the robust linear classification procedures were derived and discussed in detail. In the last chapter, we discussed yet another generalization of the results in Chapter 3 by considering the classification problem based on a multivariate dichotomous variable and an associated univariate normal variable. In this chapter, we consider this classification problem in a very general set-up , namely, when the data contains a multivariate dichotomous variable followed by a multivariate normal variable.

In Section 2, we first describe the basic model of the general classification

problem based on $k-$ variate dichotomous variable and an associated $p-$variate normal variable. In Section 3, we derive the classical likelihood-ratio classification procedure for the case when all the parameters in the model are known and derive exact and explicit formulas for the two errors of misclassification. In Section 4, we consider the case when all the parameters in the model are unknown and describe the classical classification procedure and its asymptotic determination of the cut-off point when one of the errors of misclassification is at a pre-fixed level. In Section 5, we make use of the multivariate MML estimators developed in Chapter 2 to propose a robust linear procedure for this classification problem which is analogous to the classical linear procedure described in Section 4. We also describe the asymptotic determination of the cut-off point of this robust procedure when one of the errors of misclassification is at a pre-fixed level and discuss some asymptotic optimality properties of this procedure. It should be pointed out here that this development considers the classification problem in a very general form and hence generalizes the results presented in Chapters 3, 4 and 6 in different ways.

7.2 The model

Let $\mathbf{X}$ be a $k-$variate dichotomous variable with probability mass function

$$\Pr(\mathbf{X} = \mathbf{x}) = p_{\mathbf{x}},$$

where each element of $\mathbf{x}$ takes 0 or 1. Following Olkin and Tate (1961), it is assumed that $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}$ has p-variate continuous variable such that the conditional distribution of $\mathbf{Y}$, given $\mathbf{X} = \mathbf{x}$, is a p-variate normal with mean $\mu_{\mathbf{x}}$ and variance-covariance matrix $\Sigma_{\mathbf{x}}$. Then the joint density function of $\mathbf{W} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$

$$f(\mathbf{w}) = p_{\mathbf{x}}\phi_p(\mathbf{y}; \mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}),$$

where $\phi_p(\mathrm{y};\mu_\mathrm{x},\Sigma_\mathrm{x})$ is the density function of a $p-$variate normal distribution with mean vector μ_x and variance-covariance matrix Σ_x. The marginal distribution of Y is

$$g(\mathrm{y}) \;=\; \sum_{\mathrm{x}} p_\mathrm{x}\phi_p(\mathrm{y};\mu_\mathrm{x},\Sigma_\mathrm{x}), \tag{7.1}$$

i.e. a 2^k component multivariate mixed normal distribution. Here, $\sum_\mathrm{x}$ denotes the summation over all possible vectors of x. It is noted that X and Y are independent if and only if $\mu_\mathrm{x} = \mu$ and $\Sigma_\mathrm{x} = \Sigma$, for all x. For the problem of classifying an observation w into one of two populations, Π_1 and Π_2, it is assumed that if w is from Π_i, then its density function is

$$f_i(\mathrm{w}) = p_\mathrm{x}^{(i)}\phi_p(\mu_\mathrm{x}^{(i)},\Sigma_\mathrm{x}), \quad i = 1,2. \tag{7.2}$$

It should be noted that the covariance matrix of W is not the same in Π_1 and Π_2. However, the conditional covariance matrices of Y given X are equal for the two populations. Thus Σ_x does not change with i, but $p_\mathrm{x}^{(i)}$ and $\mu_\mathrm{x}^{(i)}$ do.

7.3 Classification when the population parameters are known

Based on the model described above, we derive in this section the classical likelihood-ratio classification procedure for the case when all the parameters of the model are known and then derive exact and explicit expressions for the errors of misclassification of this procedure.

7.3.1 Likelihood ratio classification procedure

For the model specified by (7.2), the likelihood ratio is

$$\begin{aligned} V(\mathrm{w}) &= \ln\left[\frac{f_1(\mathrm{w})}{f_2(\mathrm{w})}\right] \\ &= -\frac{1}{2}(\mathrm{y}-\mu_\mathrm{x}^{(1)})^T\Sigma_\mathrm{x}{}^{-1}(\mathrm{y}-\mu_\mathrm{x}^{(1)}) + \frac{1}{2}(\mathrm{y}-\mu_\mathrm{x}^{(2)})^T\Sigma_\mathrm{x}{}^{-1}(\mathrm{y}-\mu_\mathrm{x}^{(2)}) \end{aligned}$$

$$
\begin{aligned}
& + \ln \left[\frac{p_{\mathbf{x}}^{(1)}}{p_{\mathbf{x}}^{(2)}} \right] \\
= \; & -\frac{1}{2}(\mathbf{y}^T \Sigma_{\mathbf{x}}^{-1} \mathbf{y} - 2\mathbf{y}^T \Sigma_{\mathbf{x}}^{-1} \mu_{\mathbf{x}}^{(1)} + \mu_{\mathbf{x}}^{(1)^T} \Sigma_{\mathbf{x}}{}^{-1} \mu_{\mathbf{x}}^{(1)}) \\
& + \frac{1}{2}(\mathbf{y}^T \Sigma_{\mathbf{x}}^{-1} \mathbf{y} - 2\mathbf{y}^T \Sigma_{\mathbf{x}}^{-1}(x) \mu_{\mathbf{x}}^{(2)} + \mu_{\mathbf{x}}^{(2)^T} \Sigma_{\mathbf{x}}^{-1} \mu_{\mathbf{x}}^{(2)}) \\
& + \ln \left[\frac{p_{\mathbf{x}}^{(1)}}{p_{\mathbf{x}}^{(2)}} \right] \\
= \; & \mathbf{y}^T \Sigma_{\mathbf{x}}{}^{-1} (\mu_{\mathbf{x}}^{(1)} - \mu_{\mathbf{x}}^{(2)}) \\
& - \frac{1}{2} \mu_{\mathbf{x}}^{(1)^T} \Sigma_{\mathbf{x}}^{-1} \mu_{\mathbf{x}}^{(1)} + \frac{1}{2} \mu_{\mathbf{x}}^{(2)^T} \Sigma_{\mathbf{x}}^{-1} \mu_{\mathbf{x}}^{(2)} \\
& + \ln \left[\frac{p_{\mathbf{x}}^{(1)}}{p_{\mathbf{x}}^{(2)}} \right] \\
= \; & \left\{ \mathbf{y} - \frac{1}{2}(\mu_{\mathbf{x}}^{(1)} + \mu_{\mathbf{x}}^{(2)}) \right\} \Sigma_{\mathbf{x}}^{-1} (\mu_{\mathbf{x}}^{(1)} - \mu_{\mathbf{x}}^{(2)}) \\
& + \ln \left[\frac{p_{\mathbf{x}}^{(1)}}{p_{\mathbf{x}}^{(2)}} \right]
\end{aligned}
\tag{7.3}
$$

Thus the classification procedure is to classify an observation $\mathbf{w}_0$ into Π_1 or Π_2 according as

$$
V(\mathbf{w}) \underset{<}{>} C, \tag{7.4}
$$

where the cut-off point C is determined so that the error of misclassification e_{12} (for example) is at a pre-specified level. It is noted that when $\mathbf{X}$ and $\mathbf{Y}$ are independent (i.e., $\mu_{\mathbf{x}}^{(i)} = \mu^{(i)}$, $i = 1, 2$, and $\Sigma_{\mathbf{x}} = \Sigma$) and $\mathbf{X}$ has the same distribution in Π_1 and Π_2 (i.e., $p_{\mathbf{x}}^{(1)} = p_{\mathbf{x}}^{(2)}$), then for all $\mathbf{x}$'s $V(\mathbf{w}) = V(\mathbf{y})$. Thus all the information for classification comes from $\mathbf{Y}$ and the value of the dichotomous variable $\mathbf{X}$ is not involved in the decision. In this case, the conditional and the marginal distributions of $\mathbf{Y}$ are identically normal and the classical classification procedure based on the p continuous variables only is the same as the above procedure.

7.3.2 Errors of misclassification

Since $V(\mathrm{w})$ is a linear combination of normal random variate, given w is from Π_1, the conditional distribution of $V(\mathrm{w})$ given $\mathbf{X} = \mathrm{x}$ is normal with mean

$$\frac{1}{2}\left\{\mu_{\mathrm{x}}^{(1)} - \mu_{\mathrm{x}}^{(2)}\right\}^T \Sigma_{\mathrm{x}}^{-1}(\mu_{\mathrm{x}}^{(1)} - \mu_{\mathrm{x}}^{(2)}) + \ln\left[\frac{p_{\mathrm{x}}^{(1)}}{p_{\mathrm{x}}^{(2)}}\right]$$

and variance

$$\eta^2(\mathrm{x}) = \left\{\mu_{\mathrm{x}}^{(1)} - \mu_{\mathrm{x}}^{(2)}\right\}^T \Sigma_{\mathrm{x}}^{-1}\left\{\mu_{\mathrm{x}}^{(1)} - \mu_{\mathrm{x}}^{(2)}\right\}.$$

Also, given that w is from Π_2 the conditional distribution of $V(\mathrm{w})$ given $\mathbf{X} = \mathrm{x}$ is normal with mean $-\frac{1}{2}\eta^2(\mathrm{x}) + \ln\left[\frac{p_{\mathrm{x}}^{(1)}}{p_{\mathrm{x}}^{(2)}}\right]$ and variance $\eta^2(\mathrm{x})$. The probability of misclassifying an observation from Π_1 into Π_2 is therefore given by

$$\begin{aligned} e_{12} &= Pr(V(\mathrm{w}) < C|\mathrm{w} \in \Pi_1) \\ &= \sum_{\mathrm{x}} p_{\mathrm{x}}^{(1)} P(V(\mathrm{w}) < C|\mathrm{y} \sim \phi_p(\mu_{\mathrm{x}}^{(1)}, \Sigma_{\mathrm{x}})) \\ &= \sum_{\mathrm{x}} p_{\mathrm{x}}^{(1)} \Phi\left\{\frac{C - \frac{1}{2}\eta^2(\mathrm{x}) - \ln\left[\frac{p_{\mathrm{x}}^{(1)}}{p_{\mathrm{x}}^{(2)}}\right]}{|\eta(\mathrm{x})|}\right\}, \end{aligned} \tag{7.5}$$

and the probability of misclassifying an observation from Π_2 into Π_1 is given by

$$\begin{aligned} e_{21} &= Pr(V(\mathrm{w}) > C|\mathrm{w} \in \Pi_2) \\ &= \sum_{\mathrm{x}} p_{\mathrm{x}}^{(2)} P(V(\mathrm{w}) > C|\mathrm{y} \sim \phi_p(\mu_{\mathrm{x}}^{(2)}, \Sigma_{\mathrm{x}}) \\ &= \sum p_{\mathrm{x}}^{(2)} \Phi\left\{-\frac{C + \frac{1}{2}\eta^2(\mathrm{x}) - \ln\left[\frac{p_{\mathrm{x}}^{(1)}}{p_{\mathrm{x}}^{(2)}}\right]}{|\eta(\mathrm{x})|}\right\} \end{aligned} \tag{7.6}$$

7.4 Classical Classification procedure when the population parameters are unknown

Since the population parameters are not known in most applications, the usual practice is to substitute estimates of the population parameters $p_{\mathrm{x}}^{(i)}$, $\mu_{\mathrm{x}}^{(i)}$, Σ_{x}, for

all x into $V(\mathrm{x})$. Thus, let $\begin{pmatrix} \mathrm{x}_{1i} \\ \mathrm{y}_{1i} \end{pmatrix}$, $i = 1, 2, \ldots, n_1$, be a random sample of size n_1 from population Π_1 which is the model specified in (7.2) with $p_{\mathrm{x}}^{(1)}$ and $\mu_{\mathrm{x}}^{(1)}$. Similarly, let $\begin{pmatrix} \mathrm{x}_{2i} \\ \mathrm{y}_{2i} \end{pmatrix}$, $i = 1, 2, \ldots, n_2$, be a random sample of size n_2 from population Π_2 which is the model specified in (7.2) with $p_{\mathrm{x}}^{(2)}$ and $\mu_{\mathrm{x}}^{(2)}$. Then the classical estimators (MLEs) of the parameters are

$$\hat{p}_{\mathrm{x}}^{(i)} = \frac{n_i(\mathrm{x})}{n_i}, \quad (i = 1, 2),$$

where $n_i(\mathrm{x})$ is the number of observations in the ith sample that correspond to $\mathrm{X} = \mathrm{x}$, and

$$\hat{\mu}_{\mathrm{x}}^{(i)} = \bar{\mathrm{y}}_i(\mathrm{x})$$

and

$$\hat{\Sigma}_{\mathrm{x}} = \mathrm{S}(\mathrm{x}) = \frac{(n_1(\mathrm{x}) - 1)\mathrm{S}_1(\mathrm{x}) + (n_2(\mathrm{x}) - 1)\mathrm{S}_2(\mathrm{x})}{n_1(\mathrm{x}) + n_2(\mathrm{x}) - 2}$$

for all $\mathrm{X} = \mathrm{x}$, and $i = 1$ and 2; here, $\mathrm{S}_i(x)$ is the sample variance-covariance matrix from the ith sample and is given by

$$\mathrm{S}_i(\mathrm{x}) = \frac{1}{n_i(\mathrm{x}) - 1} \sum_{l=1}^{n_i(\mathrm{x})} (\mathrm{y}_{il}(\mathrm{x}) - \bar{\mathrm{y}}_i(\mathrm{x}))(\mathrm{y}_{il}(\mathrm{x}) - \bar{\mathrm{y}}_i(\mathrm{x}))^T.$$

Let $\mathrm{w}_0 = \begin{pmatrix} \mathrm{x}_0 \\ \mathrm{y}_0 \end{pmatrix}$ be an independent new observation that needs to be classified into either Π_1 or Π_2. Then the classical linear classification procedure is to classify $\mathrm{w}_0 = \begin{pmatrix} \mathrm{x}_0 \\ \mathrm{y}_0 \end{pmatrix}$ into Π_1 or Π_2 according as

$$\begin{aligned} V(\mathrm{w}_0) &= \left\{\mathrm{y}_0 - \frac{1}{2}(\bar{\mathrm{y}}_1(\mathrm{x}_0) + \bar{\mathrm{y}}_2(\mathrm{x}_0))\right\}^T \mathrm{S}^{-1}(\mathrm{x}_0)(\bar{\mathrm{y}}_1(\mathrm{x}_0) - \bar{\mathrm{y}}_2(\mathrm{x}_0)) \\ &\quad + \ln\left[\frac{\hat{p}_{\mathrm{x}}^{(1)}}{\hat{p}_{\mathrm{x}}^{(2)}}\right] \gtrless C^*, \end{aligned} \tag{7.7}$$

where the cut-off point C^* is determined so that the error of misclassification e_{12} (for example) is at a pre-specified level. The cut-off point may be obtained as the

solution $\hat{C}^*$ of the equation (with e_{12} fixed as α) by substituting the appropriate estimators for the parameters $\eta^2(\mathrm{x})$, $p_{\mathrm{x}}^{(1)}$ and $p_{\mathrm{x}}^{(2)}$ in (7.5). The classical estimator of $\eta^2(\mathrm{x})$ is

$$\hat{\eta}^2(\mathrm{x}) = \frac{n_1(\mathrm{x}) + n_2(\mathrm{x}) - p - 3}{n_1(\mathrm{x}) + n_2(\mathrm{x}) - 2}(\bar{\mathrm{y}}_1(\mathrm{x}) - \bar{\mathrm{y}}_2(\mathrm{x}))^T \mathrm{S}^{-1}(\mathrm{x})(\bar{\mathrm{y}}_1(\mathrm{x}) - \bar{\mathrm{y}}_2(\mathrm{x})), \quad \text{for } \mathrm{X} = \mathrm{x};$$

i.e. Mahalanobis' sample squared distance. Note that this estimator is not unbiased but the unbiased estimator of $\eta^2(\mathrm{x})$ may take up inadmissible negative values as we mentioned in Chapter 2. Therefore the cut-off point $\hat{C}^*$ is the solution of

$$\sum_{\mathrm{x}} \hat{p}_{\mathrm{x}}^{(1)} \Phi \left\{ \frac{\hat{C}^* - \frac{1}{2}\hat{\eta}^2(\mathrm{x}) - \ln\left(\frac{\hat{p}_{\mathrm{x}}^{(1)}}{\hat{p}_{\mathrm{x}}^{(2)}}\right)}{|\hat{\eta}(\mathrm{x})|} \right\} = \alpha. \tag{7.8}$$

In the case when both the errors of misclassification are allowed to float, the classical linear classification procedure is to classify the new independent observation $\mathrm{w}_0 = \begin{pmatrix} \mathrm{x}_0 \\ \mathrm{y}_0 \end{pmatrix}$ into Π_1 or Π_2 according as

$$\begin{aligned} V(\mathrm{w}_0) &= \left\{ \mathrm{y}_0 - \frac{1}{2}(\bar{\mathrm{y}}_1(\mathrm{x}_0) + \bar{\mathrm{y}}_2(\mathrm{x}_0)) \right\}^T \mathrm{S}^{-1}(\mathrm{x}_0)(\bar{\mathrm{y}}_1(\mathrm{x}_0) - \bar{\mathrm{y}}_2(\mathrm{x}_0)) \\ &\quad + \ln\left[\frac{\hat{p}_{\mathrm{x}}^{(1)}}{\hat{p}_{\mathrm{x}}^{(2)}}\right] \gtrless = 0. \end{aligned} \tag{7.9}$$

7.5 Robust Linear Classification Procedure

In the last section, the classical ML estimators were used to develop a linear classification procedure for classifying an observation, based on a $k-$variate dichotomous variable and an associated $p-$variate normal variable, into one of two populations. In this section, we make use of the multivariate MML estimators derived in Chapter 2 to propose a robust linear classification procedure which is analogous to the classical linear classification procedure presented in the last section. We also describe the asymptotic determination of the cut-off point of this procedure and some asymptotic optimality properties. To this end, let $(\hat{\mu}_{\mathrm{x}}^{(1)}, \hat{\Sigma}_{\mathrm{x}}^{(1)})$ and $(\hat{\mu}_{\mathrm{x}}^{(2)}, \hat{\Sigma}_{\mathrm{x}}^{(2)})$ be the multivariate MML estimators of $(\mu_{\mathrm{x}}^{(1)}, \Sigma_{\mathrm{x}}^{(1)})$ and $(\mu_{\mathrm{x}}^{(2)}, \Sigma_{\mathrm{x}}^{(2)})$ obtained from the

samples $\mathrm{y}_{1k}(\mathrm{x})$, $k = 1, 2, \ldots, n_1(\mathrm{x})$, and $\mathrm{y}_{2k}(\mathrm{x})$, $k = 1, 2, \ldots, n_2(\mathrm{x})$, respectively, for $\mathrm{X} = \mathrm{x}$. Let $\hat{\Sigma}_{\mathrm{x}}$ be the pooled MML estimator of Σ_{x} given by

$$\hat{\Sigma}_{\mathrm{x}} = \frac{(A_1(\mathrm{x}) - 1)\hat{\Sigma}_{\mathrm{x}}^{(1)} + (A_2(\mathrm{x}) - 1)\hat{\Sigma}_{\mathrm{x}}^{(2)}}{A_1(\mathrm{x}) + A_2(\mathrm{x}) - 2}. \tag{7.10}$$

Then the robust classification procedure for classifying a new observation $\mathrm{w}_0 = \begin{pmatrix} \mathrm{x}_0 \\ \mathrm{y}_0 \end{pmatrix}$ is to classify it into Π_1 or Π_2 according as

$$\begin{aligned} V_R(\mathrm{x}) &= \left\{\mathrm{y}_0 - \frac{1}{2}(\hat{\mu}_{\mathrm{x}_0}^{(1)} + \hat{\mu}_{\mathrm{x}_0}^{(2)})\right\}^T \hat{\Sigma}_{\mathrm{x}_0}^{-1}(\hat{\mu}_{\mathrm{x}}^{(1)} - \mu_{\mathrm{x}}^{(2)}) \\ &\quad + \ln\left[\frac{\hat{p}_{\mathrm{x}_0}^{(1)}}{\hat{p}_{\mathrm{x}_0}^{(2)}}\right] \gtrless C_R^*, \end{aligned} \tag{7.11}$$

where the cut-off point C_R^* is determined so that the error of misclassification e_{12} is fixed as α. Therefore, cut-off point C_R^* can be obtained as a solution of (7.5) after substituting multivariate MML estimators of $\mu_{\mathrm{x}}^{(i)}$ and Σ_{x} which is

$$\sum_{\mathrm{x}} \hat{p}_{\mathrm{x}}^{(1)} \Phi\left\{\frac{\hat{C}_R^* - \frac{1}{2}\hat{\eta}_R^2(\mathrm{x}) - \ln\left(\frac{\hat{p}_{\mathrm{x}}^{(1)}}{\hat{p}_{\mathrm{x}}^{(2)}}\right)}{|\hat{\eta}_R(\mathrm{x})|}\right\} = \alpha, \tag{7.12}$$

where

$$\hat{\eta}_R^2(\mathrm{x}) = \frac{A_1(\mathrm{x}) + A_2(\mathrm{x}) - p - 3}{A_1(\mathrm{x}) + A_2(\mathrm{x}) - 2}(\hat{\mu}_{\mathrm{x}}^{(1)} - \hat{\mu}_{\mathrm{x}}^{(2)})^T \Sigma_{\mathrm{x}}^{-1}(\hat{\mu}_{\mathrm{x}}^{(1)} - \hat{\mu}_{\mathrm{x}}^{(2)}) \tag{7.13}$$

In the case when both the errors of misclassification are allowed to float, the robust linear classification procedure is to classify the new independent observation $\mathrm{w}_0 = \begin{pmatrix} \mathrm{x}_0 \\ \mathrm{y}_0 \end{pmatrix}$ into Π_1 or Π_2, or arbitrarily into Π_1 or Π_2, according as

$$V_R(\mathrm{w}_0) \gtrless 0, \tag{7.14}$$

where the statistic $V_R(\mathrm{w})$ is as given in (7.11).

Theorem 13 *The statistic $V_R(\mathrm{w}_0)$ in (7.11) asymptotically has exactly the same distribution as the statistic $V(\mathrm{w}_0)$ in (7.7), under normality.*

Proof: From the properties of multivariate MML estimators discussed in Chapter 12 we have for large n_1 and n_2 the distribution of the statistic $V_R(\mathrm{w}_0)$ in (7.11) to be the same as the distribution of the statistic $V(\mathrm{w}_0)$ in (7.3) under normality. Also from the classical theory of ML estimators the distribution of the statistic $V(\mathrm{w}_0)$ is the same as the distribution of the statistics $V(\mathrm{w}_0)$ in (7.3) for large n_1 and n_2 under normality. Hence the theorem. ∎

Theorem 14 *Asymptotically $V(\mathrm{w}_0)$ and $V_R(\mathrm{w}_0)$ procedures have exactly the same $1 - e_{21}$ value, for a pre-assigned value of e_{12}, under normality.*

Proof: The result follows from Theorem 13 and the fact that $\hat{C}$ and $\hat{C}_R$ both converge to C given by (7.5) as n_1 and n_2 become large, since $\hat{p}_{\mathrm{x}}^{(i)}$ converges to $p_{\mathrm{x}}^{(i)}$, $\bar{\mathrm{y}}_i(\mathrm{x})$ and $\hat{\mu}_{\mathrm{x}}^{(i)}$ both converge to $\mu_{\mathrm{x}}^{(i)}$, $i = 1, 2$, and $\mathrm{S}(\mathrm{x})$ and $\hat{\Sigma}_{\mathrm{x}}$ both converge to Σ_{x}. ∎

As mentioned earlier in the introduction of this chapter, the classification problem considered in this chapter is in a very general form and, therefore, the methods presented here generalize the results of Chapter 3, 4 and 6. The methodology presented in this chapter will enable one to classify data comprising of a $k-$ variate dichotomous variable and an associated $p-$ variate continuous variable and, hence, will be quite useful in many practical situations some of which have been indicated in previous chapters.

Chapter 8

Robustness of the classical univariate linear classification procedure under heterogeneity of variances

8.1 Introduction

In the preceding chapters, our discussion has been regarding the robustness of the classical classification procedures (basically linear classification procedures) for the variety of situations considered and then in developing "robust" classification procedures based on MML estimators for these situations. In all these discussions, the robustness is assessed through the insensitivity of the procedures under departures from the presumed normal distribution of the populations.

It should be remembered that yet another assumption made applying the linear classification procedure is that two populations have an unknown but common variance. One, therefore, would like to examine the performance of the linear classification procedure when this assumption of homogeneity of variances is violated. In this chapter, we carry out such a study for the classical univariate linear classification procedure. We derive the exact distribution functions and expected

values of the errors of misclassification under this heterogeneity of variances for the normal as well as the outlier-normal models for the populations. Through these results, we assess the performance of the linear classification procedure.

Given the training samples $x_{11}, x_{12}, \ldots, x_{1n_1}$ and $x_{21}, x_{22}, \ldots, x_{2n_2}$ from populations Π_1 and Π_2, respectively, and an independent observation x_0 to be allocated, the classical linear classification procedure allocates x_0 into Π_1 or Π_2 or randomly into Π_1 or Π_2 according as

$$\left\{x - \frac{\bar{x}_1 + \bar{x}_2}{2}\right\}(\bar{x}_1 - \bar{x}_2) \gtrless 0 \tag{8.1}$$

where

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}.$$

This is the most commonly used procedure for the two group classification problem when the errors of misclassification are both allowed to float (that is, neither is fixed at a pre-specified level). If sampling is from normality with the means and equal variance being known, then Fisher's linear classification procedure (the linear classification procedure in (8.1) with the known parameters μ_1 and μ_2 being present in place of their estimates $\bar{x}_1$ and $\bar{x}_2$) is optimal in the sense that it minimizes the total probability of misclassification (Lachenbruch, 1975). By considering the case when all the parameters are known, Gilbert (1969) has examined the effect of heterogeneity of variances on Fisher's linear classification procedure. Then, by comparing this case to that of the quadratic classification procedure, Gilbert (1969) has shown that the quadratic classification procedure performs slightly better than the linear classification procedure. It is not known, however, what effect the heterogeneity of variances has on the linear classification procedure in (8.1) when the parameters are all unknown. In this chapter, we precisely examine this situation.

In Section 2, by assuming normal distributions for the two populations with unequal variances we derive the distribution functions and expected values of the errors of misclassification of the linear procedure in (8.1). Next, in Section 3 we derive similar results by assuming outlier-normal models for the samples from the two populations. In Section 4, we make use of the results developed in Section 2

and 3 to study the effect of the heterogeneity of variances on the linear classification procedure.

8.2 Sampling from the normal model

In this section, we derive the distribution functions and expected values of the errors of misclassification of the linear classification procedure in (8.1) in the case of heterogeneity of variances. Assume the mean and variance of population Π_i to be μ_i and σ_i^2 for $i = 1, 2$. Let us denote $\frac{\sigma_2}{\sigma_1}$ to be equal to ϕ. Define random variables U and V to be

$$\begin{aligned} U &= \bar{X}_1 + \bar{X}_2 \\ V &= \bar{X}_2 - \bar{X}_1. \end{aligned}$$

Then

$$\begin{aligned} U &\sim N(\mu_u, \sigma_u^2), \\ V &\sim N(\mu_v, \sigma_v^2), \end{aligned}$$

where

$$\begin{aligned} \mu_u &= \mu_1 + \mu_2, \\ \mu_v &= \mu_2 - \mu_1, \\ \sigma_u^2 = \sigma_v^2 &= \sigma^2 \left(\frac{1}{n_1} + \frac{\phi^2}{n_2} \right) \quad (\text{with } \sigma_1 = \sigma_2), \end{aligned}$$

and the correlation coefficient ρ between U and V is

$$\rho = \left(\frac{\frac{\phi^2}{n_2} - \frac{1}{n_1}}{\frac{\phi^2}{n_2} + \frac{1}{n_1}} \right).$$

i.e. the joint distribution of (U, V) is bivariate normal. If we write u and v for realization of U and V then the linear classification rule in (8.1) becomes: Classify X_0 into Π_1 or Π_2 according as

$$V(X_0) = \left[X_0 - \frac{u}{2} \right] v \gtrless 0.$$

From this we can write the error of misclassification as

$$e_{12} = \begin{cases} \int_{U/2}^{\infty} f_1(x)dx & \text{for} \quad V > 0 \\ \int_{-\infty}^{U/2} f_1(x)dx & \text{for} \quad V < 0 \end{cases}$$

or

$$e_{12} = \begin{cases} 1 - \Phi\left[\frac{U/2-\mu_1}{\sigma}\right] & \text{for} \quad V > 0 \\ \Phi\left[\frac{U/2-\mu_1}{\sigma}\right] & \text{for} \quad V < 0 \end{cases} \tag{8.2}$$

and

$$e_{21} = \begin{cases} \int_{-\infty}^{U/2} f_1(x)dx & \text{for} \quad V > 0 \\ \int_{U/2}^{\infty} f_1(x)dx & \text{for} \quad V < 0 \end{cases}$$

or

$$e_{21} = \begin{cases} \Phi\left[\frac{U/2-\mu_2}{\sigma_2}\right] & \text{for} \quad V > 0 \\ 1 - \Phi\left[\frac{U/2-\mu_2}{\sigma_2}\right] & \text{for} \quad V < 0 \end{cases}, \tag{8.3}$$

where $\Phi(.)$ denotes the cumulative distribution function of a standard normal variable.

8.2.1 Distribution functions of errors of misclassification

From the above expressions, we can write the distribution function of e_{12} as

$$\begin{aligned} G_1(z) &= \Pr[e_{12} \leq z] \\ &= \Pr\left[1 - \Phi\left[\frac{U/2-\mu_1}{\sigma}\right] \leq z, V > 0\right] + \Pr\left[\Phi\left[\frac{U/2-\mu_1}{\sigma}\right] \leq z, V < 0\right] \\ &= \Pr\left[\frac{U/2-\mu_1}{\sigma} \geq \Phi^{-1}(1-z), V > 0\right] + \Pr\left[\frac{U/2-\mu_1}{\sigma} \leq \Phi^{-1}(z), V < 0\right] \\ &= \Pr[U \leq k_1, V < 0] + \Pr[U \geq k_2, V > 0] \end{aligned}$$

where

$$\begin{aligned} k_1 &= 2[\mu_1 + \sigma\Phi^{-1}(z)] \\ \text{and} \quad k_2 &= 2[\mu_1 - \sigma\Phi^{-1}(z)] \quad (\text{since } \Phi^{-1}(1-z) = -\Phi^{-1}(z)). \end{aligned}$$

From this we can write the distribution function of e_{12} as

$$G_1(z) = \Pr\left[\frac{U-\mu_u}{\sigma_u} \leq \frac{k_1-\mu_u}{\sigma_u}, \frac{V-\mu_v}{\sigma_v} < -\frac{\mu_v}{\sigma_v}\right] +$$

$$\begin{aligned}
&\Pr\left[\frac{U-\mu_u}{\sigma_u} \geq \frac{k_2-\mu_u}{\sigma_u}, \frac{V-\mu_v}{\sigma_v} > -\frac{\mu_v}{\sigma_v}\right] \\
= \; & H\left[\frac{k_1-\mu_u}{\sigma_u}, -\frac{\mu_v}{\sigma_v}; \rho\right] + H\left[-\frac{k_2-\mu_u}{\sigma_u}, \frac{\mu_v}{\sigma_v}; \rho\right] \qquad (8.4)
\end{aligned}$$

But

$$\begin{aligned}
\frac{k_1-\mu_u}{\sigma_u} &= \frac{2[\mu_1+\sigma\Phi^{-1}(z)]-(\mu_1+\mu_2)}{\sigma\left[\frac{1}{n_1}+\frac{\phi^2}{n_2}\right]^{1/2}} \\
&= \frac{\frac{\mu_1-\mu_2}{\sigma}+2\Phi^{-1}(z)}{\tau} \\
&= \frac{-\delta+2\Phi^{-1}(z)}{\tau},
\end{aligned}$$

where

$$\begin{aligned}
\tau^2 &= \frac{1}{n_1}+\frac{\phi^2}{n_2} \\
\delta &= \frac{\mu_2-\mu_1}{\sigma}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\frac{\mu_v}{\sigma_v} &= \frac{\mu_2-\mu_1}{\tau\sigma} \\
&= \frac{\delta}{\tau}
\end{aligned}$$

and

$$\begin{aligned}
\frac{k_2-\mu_u}{\sigma_u} &= \frac{2[\mu_1-\sigma\Phi^{-1}(z)]-(\mu_1+\mu_2)}{\sigma\tau} \\
&= -\frac{\delta+2\Phi^{-1}(z)}{\tau}
\end{aligned}$$

Therefore, the cumulative distribution function of the error of misclassification e_{12} can be written as

$$G_1(z) = H\left[\frac{-\delta+2\Phi^{-1}(z)}{\tau}, -\frac{\delta}{\tau}; \rho\right] + H\left[\frac{\delta+2\Phi^{-1}(z)}{\tau}, \frac{\delta}{\tau}; \rho\right], \qquad (8.5)$$

where $H(a,b;\rho)$ is the bivariate cumulative distribution function of a standard bivariate normal distribution with correlation coefficient ρ.

Proceeding similarly, we can also derive the cumulative distribution function of the error of misclassification e_{21} to be

$$G_2(z) = H\left[\frac{-\delta^* + 2\Phi^{-1}(z)}{\tau^*}, -\frac{\delta^*}{\tau^*}; \rho^*\right] + H\left[\frac{\delta^* + 2\Phi^{-1}(z)}{\tau^*}, \frac{\delta^*}{\tau^*}; \rho^*\right], \tag{8.6}$$

where

$$\begin{aligned}
\delta^* &= \frac{\mu_1 - \mu_2}{\sigma_2} \\
\phi^* = \frac{\sigma_1}{\sigma_2} &= \frac{1}{\phi} \\
\rho^* &= \left(\frac{\frac{\phi^{*2}}{n_1} - \frac{1}{n_2}}{\frac{\phi^{*2}}{n_1} + \frac{1}{n_2}}\right), \\
\text{and} \quad \tau^{*2} &= \frac{1}{n_2} + \frac{\phi^{*2}}{n_1},
\end{aligned}$$

8.2.2 Expected values of errors of misclassification

Since the expected value is the unconditional probabilities of error of misclassification we can write

$$E[e_{12}] = \Pr[U < 2X_0, V > 0 | X_0 \in \Pi_1] + \Pr[U > 2X_0, V < 0 | X_0 \in \Pi_2] \tag{8.7}$$

By using **Result 3** of Appendix A, Eq. (8.7) can be simplified to obtain

$$\begin{aligned}
E[e_{12}] &= H\left[\frac{2\mu_1 - \mu_u}{(4\sigma^2 + \sigma_u^2)^{1/2}}, \frac{\mu_v}{\sigma_v}; -\frac{\rho\sigma_u}{(4\sigma^2 + \sigma_u^2)^{1/2}}\right] + \\
&\quad H\left[-\frac{2\mu_1 - \mu_u}{(4\sigma^2 + \sigma_u^2)^{1/2}}, -\frac{\mu_v}{\sigma_v}; -\frac{\rho\sigma_u}{(4\sigma^2 + \sigma_u^2)^{1/2}}\right].
\end{aligned} \tag{8.8}$$

But

$$\begin{aligned}
\frac{2\mu_1 - \mu_u}{(4\sigma^2 + \sigma_u^2)^{1/2}} &= \frac{2\mu_1 - \mu_1 - \mu_2}{(4\sigma^2 + \tau^2\sigma^2)^{1/2}} \\
&= -\frac{\delta}{(4 + \tau^2)^{1/2}}, \\
\frac{\sigma_u}{(4\sigma^2 + \sigma_u^2)^{1/2}} &= \frac{\tau\sigma}{(4\sigma^2 + \tau^2\sigma^2)^{1/2}} \\
&= \frac{\tau}{(4 + \tau^2)^{1/2}},
\end{aligned}$$

and

$$\begin{aligned}\frac{\mu_v}{\sigma_v} &= \frac{\mu_2 - \mu_1}{\tau\sigma} \\ &= \frac{\delta}{\tau}.\end{aligned}$$

Therefore, the expected value of the error of misclassification e_{12} can be written as

$$E[e_{12}] = H\left[-\frac{\delta}{(4+\tau^2)^{1/2}}, \frac{\delta}{\tau}; \rho_1\right] + H\left[\frac{\delta}{(4+\tau^2)^{1/2}}, -\frac{\delta}{\tau}; \rho_1\right],$$

where

$$\rho_1 = -\frac{\rho\tau}{[4+\tau^2]^{1/2}}.$$

Proceeding similarly, we can also derive the expected value of the error of misclassification e_{21} to be

$$E[e_{21}] = H\left[-\frac{\delta^*}{(4+\tau^{*2})^{1/2}}, \frac{\delta^*}{\tau^*}; \rho_1^*\right] + H\left[\frac{\delta^*}{(4+\tau^{*2})^{1/2}}, -\frac{\delta^*}{\tau^*}; \rho_1^*\right],$$

where

$$\rho_1^* = -\frac{\rho^*\tau^*}{[4+\tau^{*2}]^{1/2}}.$$

8.3 Sampling from the normal-outlier model

In this section, we develop results similar to the ones presented in the last section for the case when the two samples are from general outlier-normal models. In particular, we derive the distribution functions and expected values of the errors of misclassification of the linear classification procedure in (8.1) under the heterogeneity of variances. These results will enable us to study the compounded effects of the heterogeneity of variances and the presence of outliers on the linear classification procedure.

Let us assume the ith training sample contains s_i general outliers; that is,

$$\begin{aligned} X_{i,j} &\sim N(\mu_i, \sigma_i^2), \qquad j = 1, 2, \ldots, n_i - s_i, \\ X_{i,j} &\sim N(\mu_i + a_i\sigma_i, b_i^2\sigma_i^2), \qquad j = n_i - s_i + 1, n_i - s_i + 2, \ldots, n_i. \end{aligned}$$

Then, it is easy to show that

$$\bar{X}_i \sim N\left(\mu_i + q_i a_i \sigma_i, \frac{\sigma_i^2}{n_i}[1 + q_i(b_i^2 - 1)]\right),$$

where $q_i = \frac{s_i}{n_i}$ is the proportion of outliers in the ith sample. Since $\sigma_1 = \sigma$ and $\sigma_2 = \phi\sigma$, we have

$$\begin{aligned}
\bar{X}_1 &\sim N\left(\mu_1 + q_1 a_1 \sigma, \frac{\sigma^2}{n_1}[1 + q_1(b_1^2 - 1)]\right) \\
\bar{X}_2 &\sim N\left(\mu_2 + q_2 a_2 \phi\sigma, \frac{\phi^2\sigma^2}{n_2}[1 + q_2(b_2^2 - 1)]\right)
\end{aligned}$$

As before, by defining the random variables U and V to be $U = \bar{X}_1 + \bar{X}_2$ and $V = \bar{X}_2 - \bar{X}_1$, we see immediately that

$$\begin{aligned}
U &\sim N(\mu_u, \sigma_u^2), \\
V &\sim N(\mu_v, \sigma_v^2),
\end{aligned}$$

where

$$\begin{aligned}
\mu_u &= \mu_1 + \mu_2 + \sigma[q_1 a_1 + q_2 a_2 \phi], \\
\mu_v &= \mu_2 - \mu_1 + \sigma[q_2 a_2 \phi - q_1 a_1], \\
\sigma_u^2 = \sigma_v^2 &= \sigma^2\left[\frac{1 + q_1(b_1^2 - 1)}{n_1} + \phi^2\left(\frac{1 + q_2(b_2^2 - 1)}{n_2}\right)\right], \\
\text{and} \quad \rho &= \frac{\phi^2\left(\frac{1+q_2(b_2^2-1)}{n_2}\right) - \frac{1+q_1(b_1^2-1)}{n_1}}{\phi^2\left(\frac{1+q_2(b_2^2-1)}{n_2}\right) + \frac{1+q_1(b_1^2-1)}{n_1}}
\end{aligned}$$

8.3.1 Distribution functions of errors of misclassification

In this case also the joint distribution of (U, V) is bivariate normal with means μ_u, μ_v, variances σ_u^2, σ_v^2 and the correlation coefficient ρ. As a result, we may use (8.4) to obtain the cumulative distribution function of the error of misclassification e_{12}. For this let us denote

$$\begin{aligned}
\tau_1^2 &= \frac{1 + q_1(b_1^2 - 1)}{n_1} + \phi^2\left(\frac{1 + q_2(b_2^2 - 1)}{n_2}\right) \\
a_1^* &= q_1 a_1 + \phi q_2 a_2, \\
\text{and} \quad a_2^* &= \phi q_2 a_2 - q_1 a_1.
\end{aligned}$$

By using these quantities we can simplify the arguments in the expression on the right hand side of (8.4) as follows:

$$\begin{aligned}
\frac{k_1 - \mu_u}{\sigma_u} &= \frac{2(\mu_1 + \Phi^{-1}(z)) - \mu_1 - \mu_2 - a_1^{*}\sigma}{\tau_1 \sigma} \\
&= \frac{-\delta + 2\Phi^{-1}(z) - a_1^{*}}{\tau_1}, \\
\frac{k_2 - \mu_u}{\sigma_u} &= \frac{2(\mu_1 - \Phi^{-1}(z)) - \mu_1 - \mu_2 - a_1^{*}\sigma}{\tau_1 \sigma} \\
&= \frac{-\delta - 2\Phi^{-1}(z) - a_1^{*}}{\tau_1}, \\
\text{and} \quad \frac{\mu_v}{\sigma_v} &= \frac{\mu_2 - \mu_1 + a_2^{*}\sigma}{\tau_1 \sigma} \\
&= \frac{\delta + a_2^{*}}{\tau_1}
\end{aligned}$$

Therefore, the cumulative density function of e_{12} in this case can be written as

$$\begin{aligned}
G_1(z) &= H\left[\frac{-\delta + 2\Phi^{-1}(z) - a_1^{*}}{\tau_1}, -\frac{\delta + a_2^{*}}{\tau_1}; \rho\right] + \\
& \quad H\left[\frac{\delta + 2\Phi^{-1}(z) + a_1^{*}}{\tau_1}, \frac{\delta + a_2^{*}}{\tau_1}; \rho\right],
\end{aligned} \tag{8.9}$$

where $H(a, b; \rho)$ denotes, as before, the cumulative distribution function of a standard bivariate normal distribution with correlation coefficient ρ.

Proceeding similarly, we can derive the cumulative distribution of the error of misclassification e_{21} as

$$\begin{aligned}
G_2(z) &= H\left[\frac{-\delta^{*} + 2\Phi^{-1}(z) - a_1^{**}}{\tau_1^{*}}, -\frac{\delta^{*} + a_2^{**}}{\tau_1^{*}}; \rho^{*}\right] + \\
& \quad H\left[\frac{\delta^{*} + 2\Phi^{-1}(z) + a_1^{**}}{\tau_1^{*}}, \frac{\delta^{*} + a_2^{**}}{\tau_1^{*}}; \rho^{*}\right],
\end{aligned} \tag{8.10}$$

where

$$\begin{aligned}
\phi^{*} &= \frac{\sigma_1}{\sigma_2} = \frac{1}{\phi} \\
\rho^{*} &= \frac{\phi^{*2}\left(\frac{1+q_1(b_1^2-1)}{n_1}\right) - \frac{1+q_2(b_2^2-1)}{n_1}}{\phi^{*2}\left(\frac{1+q_1(b_1^2-1)}{n_1}\right) + \frac{1+q_2(b_2^2-1)}{n_2}}
\end{aligned}$$

$$
\begin{aligned}
\tau_1^{*2} &= \frac{1+q_2(b_2^2-1)}{n_2} + \phi^{*2}\left(\frac{1+q_1(b_1^2-1)}{n_1}\right) \\
a_1^{**} &= q_2 a_2 + \phi^* q_1 a_1, \\
a_2^{**} &= \phi^* q_1 a_1 - q_2 a_2, \\
\text{and} \quad \delta^* &= \frac{\mu_1 - \mu_2}{\sigma_2}
\end{aligned}
$$

8.3.2 Expected values of errors of misclassification

To derive the expected value of the error of misclassification e_{12} in this case, we could use the expression in Eq. (8.8). By using the expressions of $\mu_u, \mu_v, \sigma_u, \sigma_v$ and ρ given earlier, we can simplify the arguments in the expression on the right-hand side of (8.8) as follows:

$$
\begin{aligned}
\frac{2\mu_1 - \mu_u}{(4\sigma^2 + \sigma_u^2)^{1/2}} &= \frac{2\mu_1 - \mu_1 - \mu_2 - a_1^*\sigma}{(4\sigma^2 + \sigma^2\tau_1^2)^{1/2}} \\
&= -\frac{\delta + a_1^*}{(4+\tau_1^2)^{1/2}} \\
\frac{\mu_v}{\sigma_v} &= \frac{\delta + a_2^*}{\tau_1}, \\
\text{and} \quad \rho_1 &= -\frac{\rho\sigma_u}{(4\sigma^2 + \sigma_u^2)^{1/2}} \\
&= -\frac{\rho\tau_1}{(4+\tau_1^2)^{1/2}}
\end{aligned}
$$

Therefore, the expected value of the error of misclassification e_{12} in this case is given by

$$
E(e_{12}) = H\left[-\frac{\delta + a_1^*}{(4+\tau_1^2)^{1/2}}, \frac{\delta + a_2^*}{\tau_1}; \rho_1\right] + H\left[\frac{\delta + a_1^*}{(4+\tau_1^2)^{1/2}}, -\frac{\delta + a_2^*}{\tau_1}; \rho_1\right].
$$

Following the same lines, we can derive the expected value of the error of misclassification e_{21} as

$$
E(e_{21}) = H\left[-\frac{\delta^* + a_1^{**}}{(4+\tau_1^{*2})^{1/2}}, \frac{\delta^* + a_2^{**}}{\tau_1^*}; \rho_1^*\right] + H\left[\frac{\delta^* + a_1^{**}}{(4+\tau_1^{*2})^{1/2}}, -\frac{\delta^* + a_2^{**}}{\tau_1^*}; \rho_1^*\right],
$$

where

$$
\rho_1^* = -\frac{\rho^*\tau_1^*}{(4+\tau_1^{*2})^{1/2}}. \tag{8.11}
$$

8.4 Numerical results and assessment of robustness

The distribution function $G_1(z)$ and the expected value $E[e_{12}]$ have been evaluated for a variety of values of the parameters. While $G_1(z)$ is depicted graphically, the expected values are tabled. The Figures 8.1 - 8.5 give $G_1(z)$ for the normal distribution when $n_1 = n_2 = 20$ and Table 8.1 gives the expected values of e_{12}. These are presented for $\delta = 1.0\ (0.25)\ 2.0$ when the ratio of standard deviations $\phi = 1.0\ (0.25)\ 2.0,\ 3.0,\ 5.0$.

It is readily seen from Figs. 8.1-8.5 that $G_1(z)$ for the case when $\phi > 1$ (under the heterogeneity of variances) is larger than $G_1(z)$ for the case when $\phi = 1$ (under the homogeneity of variances) when $z \leq E(e_{12})$, but the former tends to be flatter than the latter when $z > E(e_{12})$. This means that the probability of e_{12} being large becomes higher in the case of heterogeneity than in the case of homogeneity. This tendency of the cumulative distribution function of e_{12} to become flat for large values of z becomes quite pronounced for large values of ϕ. The resultant of this is that the expected value of e_{12} also increases with increasing values of ϕ as may be seen from Table 8.1. In conclusion, through this study of the distribution function as well as the expected value of the errors of misclassification, we would like to mention that the linear classification procedure in (8.1) does not affect very much when the homogeneity assumption of variances does not get violated severely (that is, when ϕ is close to 1, say up to 1.5); however, when ϕ gets larger (that is, when the homogeneity assumption of variances is seriously violated), then the errors of misclassification of the linear classification procedure tend to become large. As will be seen in the next chapter, it will be advisable in this case to use a quadratic classification procedure.

In Figures 8.6-8.10, the distribution function $G_1(z)$ has been presented for the outlier-normal case when $n_1 = n_2 = 20$ and for a wide variety of values of the parameters. In Table 8.2, the expected value of e_{12} is presented for these situations. The number of outliers present in the two samples (s_1, s_2) is taken to be (0,0), (0,1), (1,1), (0,2) and (2,2).

Table 8.1: $E(e_{12})$ values under normal distribution

ϕ	$\delta = 1.00$	$\delta = 1.25$	$\delta = 1.50$	$\delta = 2.00$
1.00	0.31100	0.26853	0.22941	0.16164
1.25	0.31206	0.26929	0.23018	0.16247
1.50	0.31345	0.27031	0.23114	0.16347
1.75	0.31498	0.27161	0.23230	0.16465
2.00	0.31637	0.27312	0.23370	0.16600
3.00	0.31700	0.27820	0.24039	0.17310

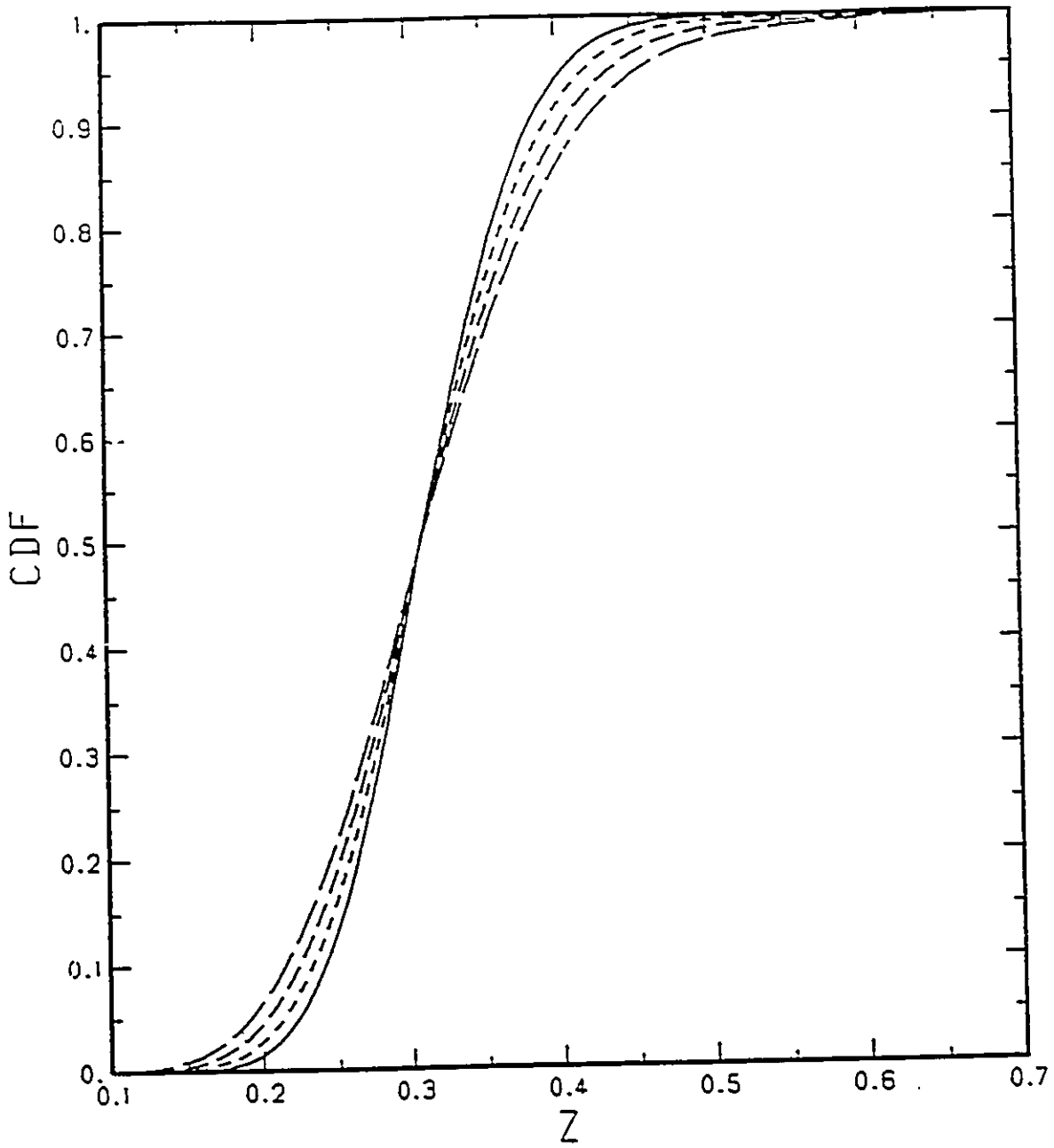
Table 8.2: $E(e_{12})$ values for $a_1 = a_2 = 1$ and $\delta = 2.0$ under outlier-normal

s_1	s_2	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
0	1	0.15240	0.16643	0.15274	0.18834	0.15391	0.23127	0.18946	0.47776
1	1	0.14780	0.18676	0.14847	0.22498	0.15008	0.27678	0.19569	0.50631
0	2	0.14670	0.22844	0.14790	0.30766	0.15202	0.39327	0.25616	0.65470
2	2	0.13920	0.30925	0.14247	0.38173	0.15168	0.45154	0.27805	0.65865

In this case, we observe from Figs 8.6-8.10 that $G_1(z)$ for the case when $\phi > 1$ (under the heterogeneity of variances) is larger than $G_1(z)$ for the normal case when $\phi = 1$ (under the homogeneity of variances) for small values of z, but the former tends to be flatter than the latter for large values of z. As the number of outliers present in the two samples increases, the distribution tends to get quite flat for large values of z. This simply means that the probability of e_{12} being large in the case of heterogeneity becomes much higher due to the presence of few outliers in the two samples. This is reflected in Table 8.2 through dramatically increasing values of the expected value of e_{12} when ϕ increases and simultaneously the number of outliers in the samples also increases. For this reason, we will propose in the next chapter a robust quadratic classification procedure based on MML estimators which can take into account the heterogeneity of variances of the two populations and simultaneously be able to handle departures from normality of the population distributions.

- Model 1 $\phi = 1.50$, $b_1 = b_2 = 5$
- Model 2 $\phi = 1.50$, $b_1 = b_2 = 10$
- Model 3 $\phi = 1.75$, $b_1 = b_2 = 5$
- Model 4 $\phi = 1.75$, $b_1 = b_2 = 10$
- Model 5 $\phi = 2.00$, $b_1 = b_2 = 5$
- Model 6 $\phi = 2.00$, $b_1 = b_2 = 10$
- Model 7 $\phi = 3.00$, $b_1 = b_2 = 5$
- Model 8 $\phi = 3.00$, $b_1 = b_2 = 10$

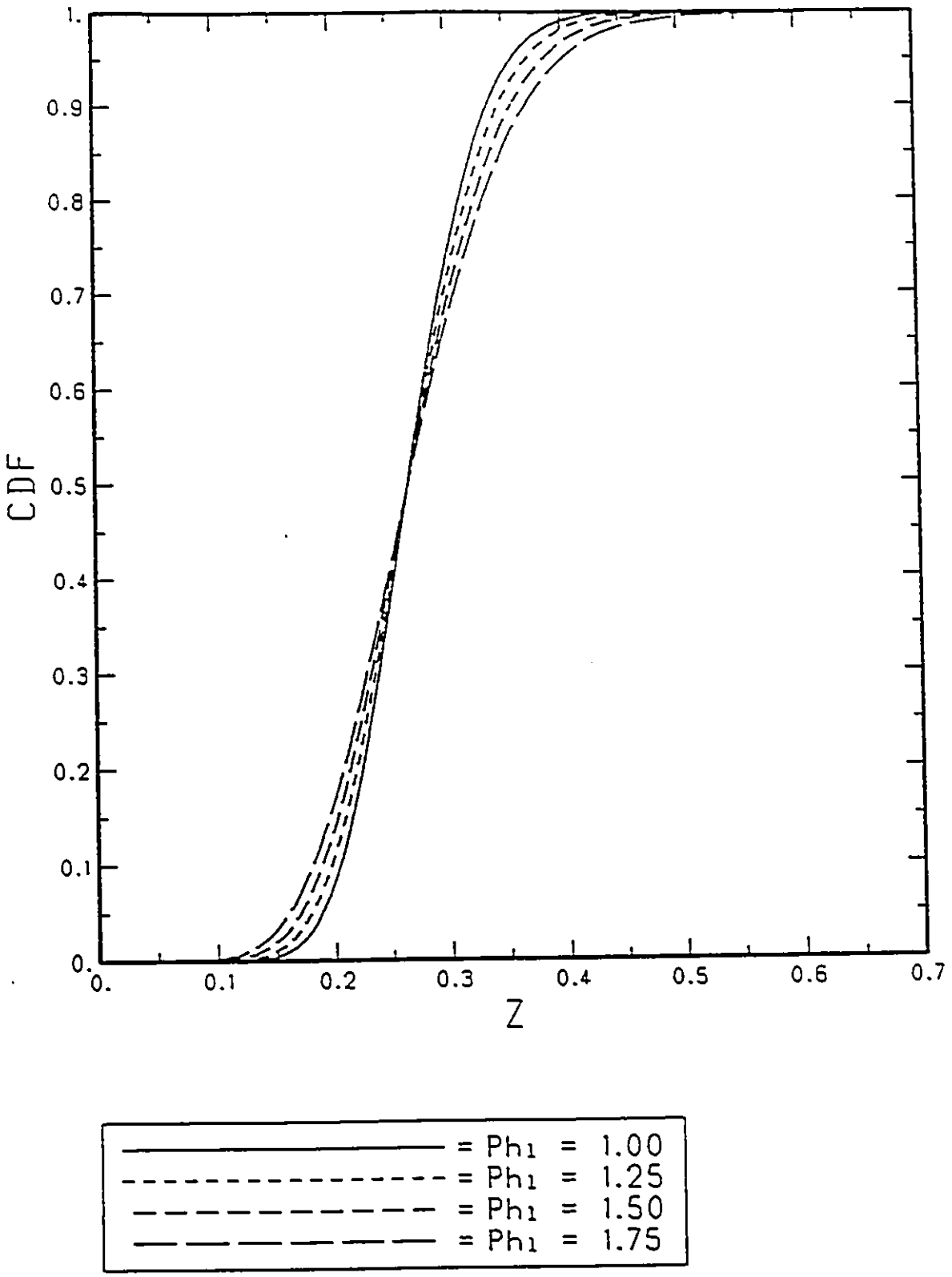
Figure 8.1: Distribution function of error of misclassification e_{12}: Normal model 1



= Ph1 = 1.00
= Ph1 = 1.25
= Ph1 = 1.50
= Phi = 1.75

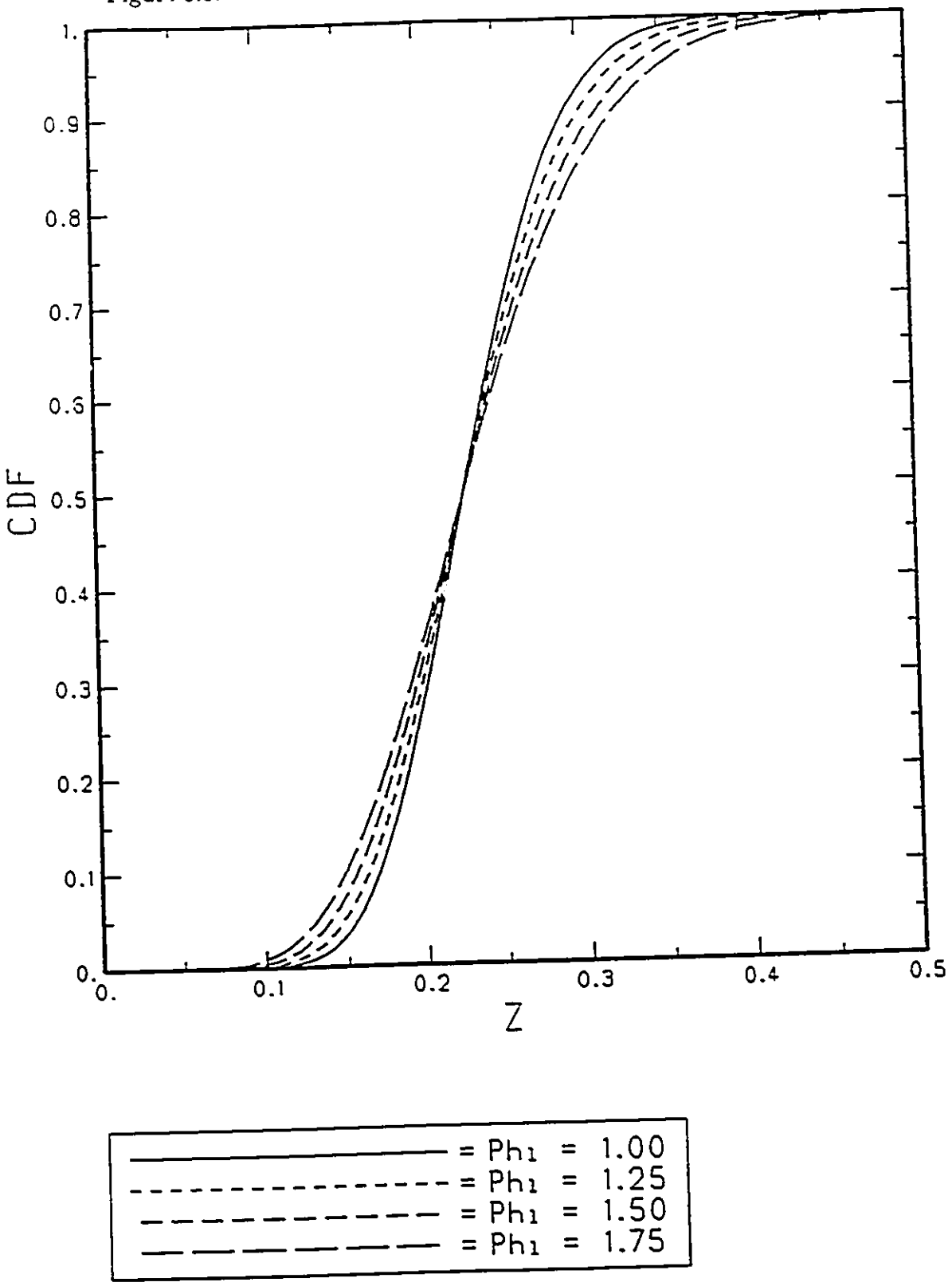
Graph for Delta = 1.00

Figure 8.2: Distribution function of error of misclassification e_{12}: Normal model 2



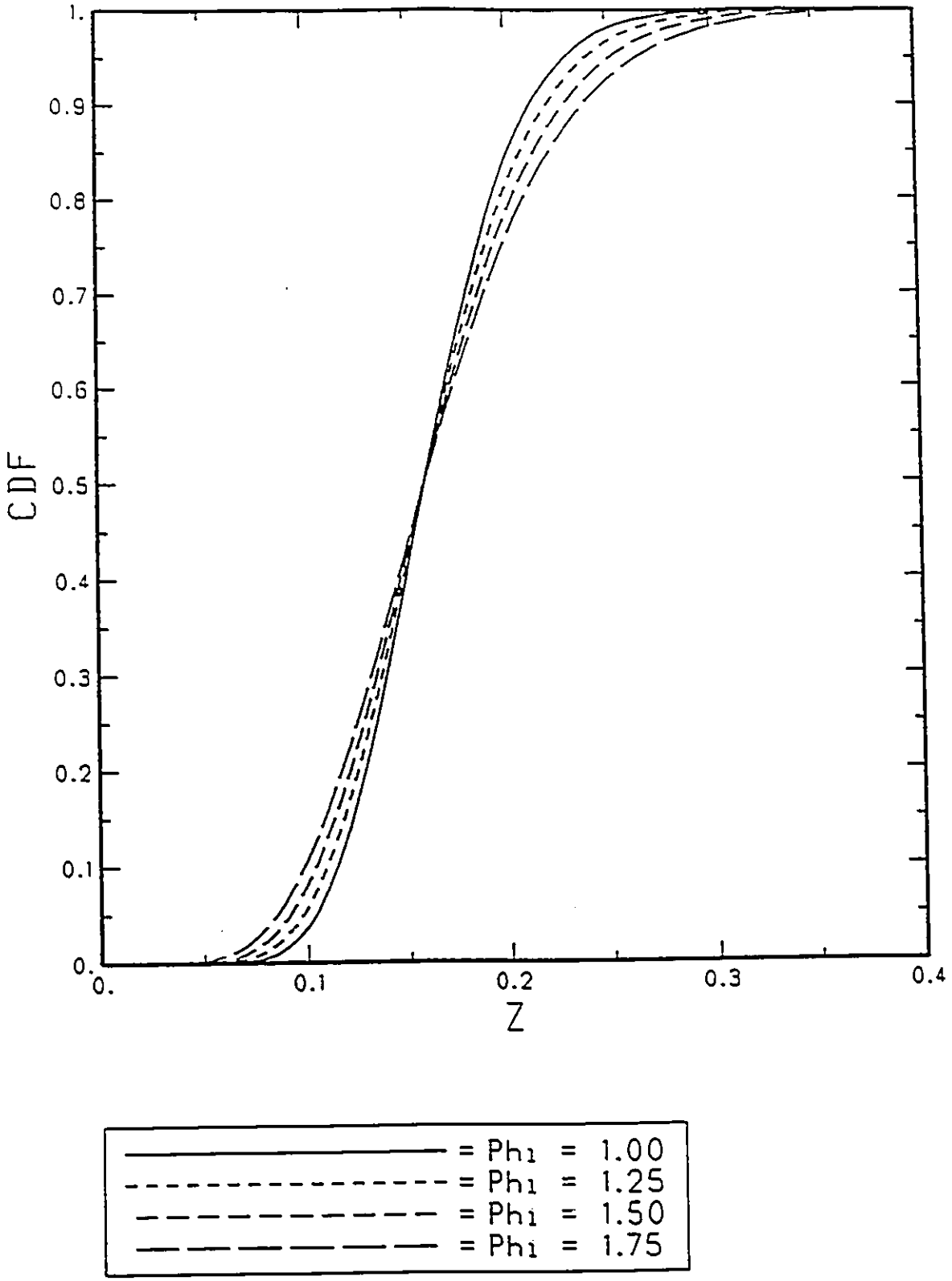
Graph for Delta = 1.25

Figure 8.3: Distribution function of error of misclassification c_{12}: Normal model 3



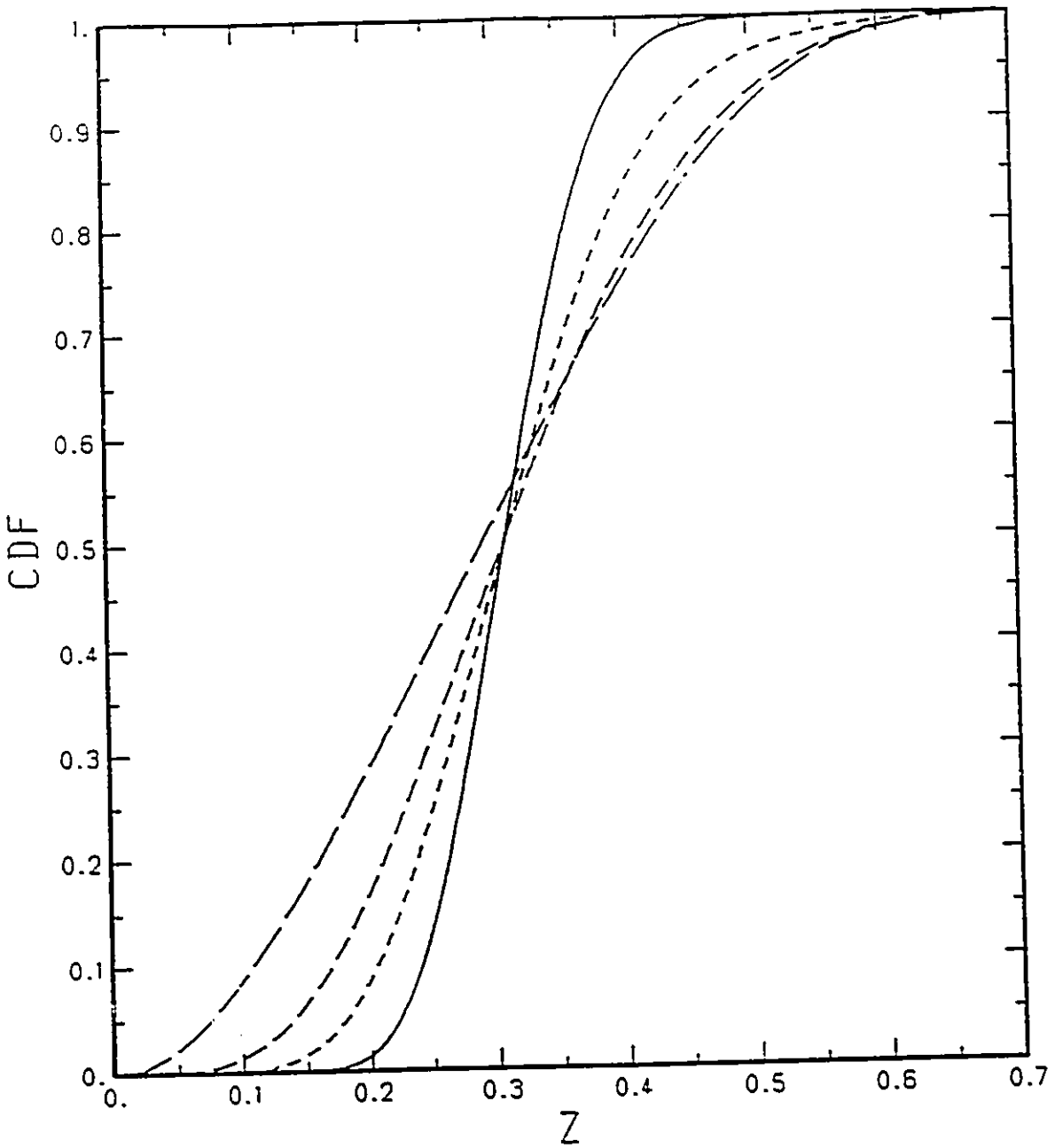
Graph for Delta = 1.50

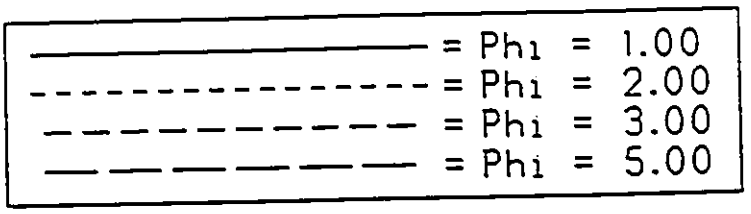
Figure 8.4: Distribution function of error of misclassification e_{12}: Normal model 4



Graph for Delta = 2.00

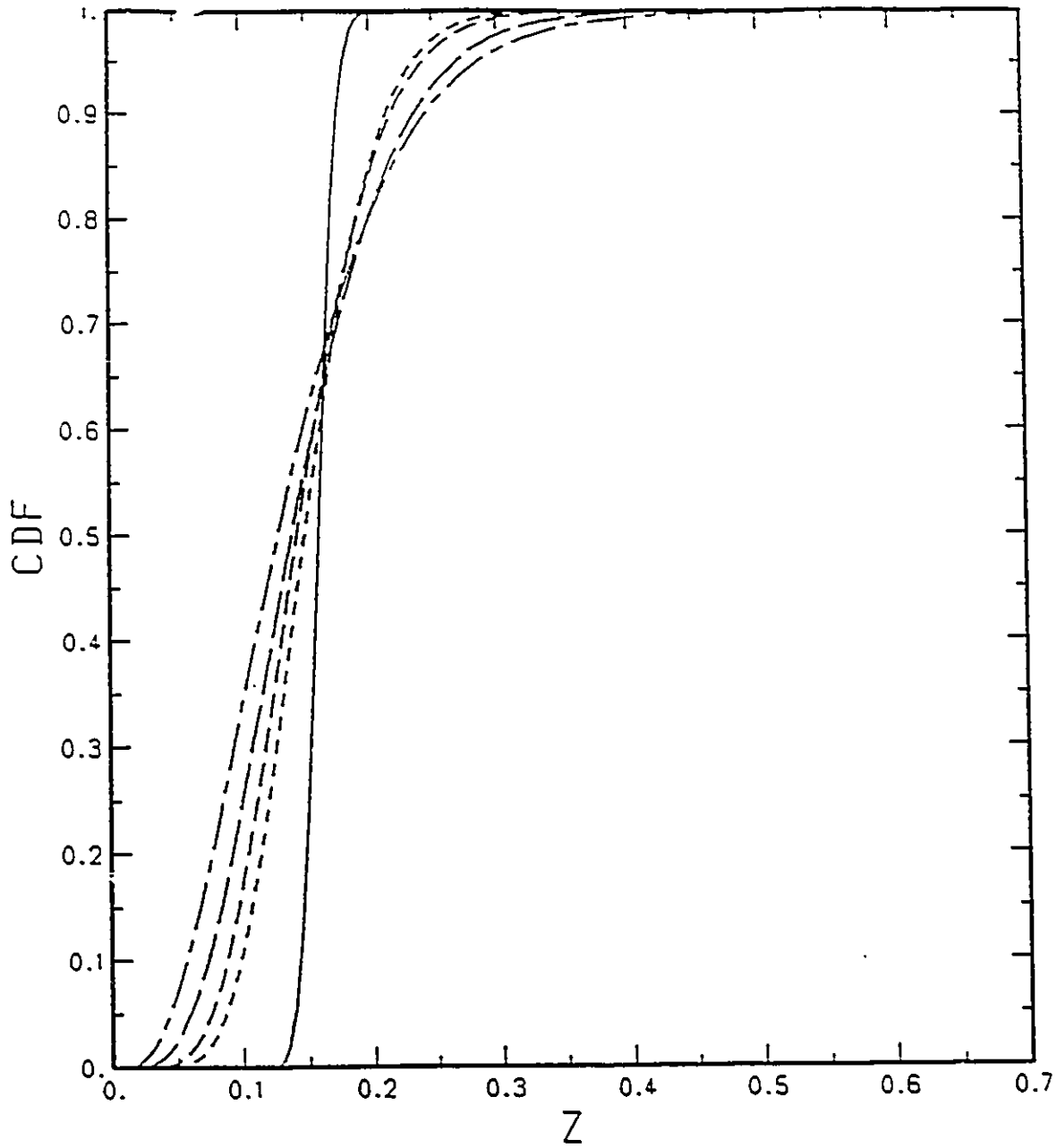
Figure 8.5: Distribution function of error of misclassification e_{12}: Normal model 5





Graph for Delta = 1.00

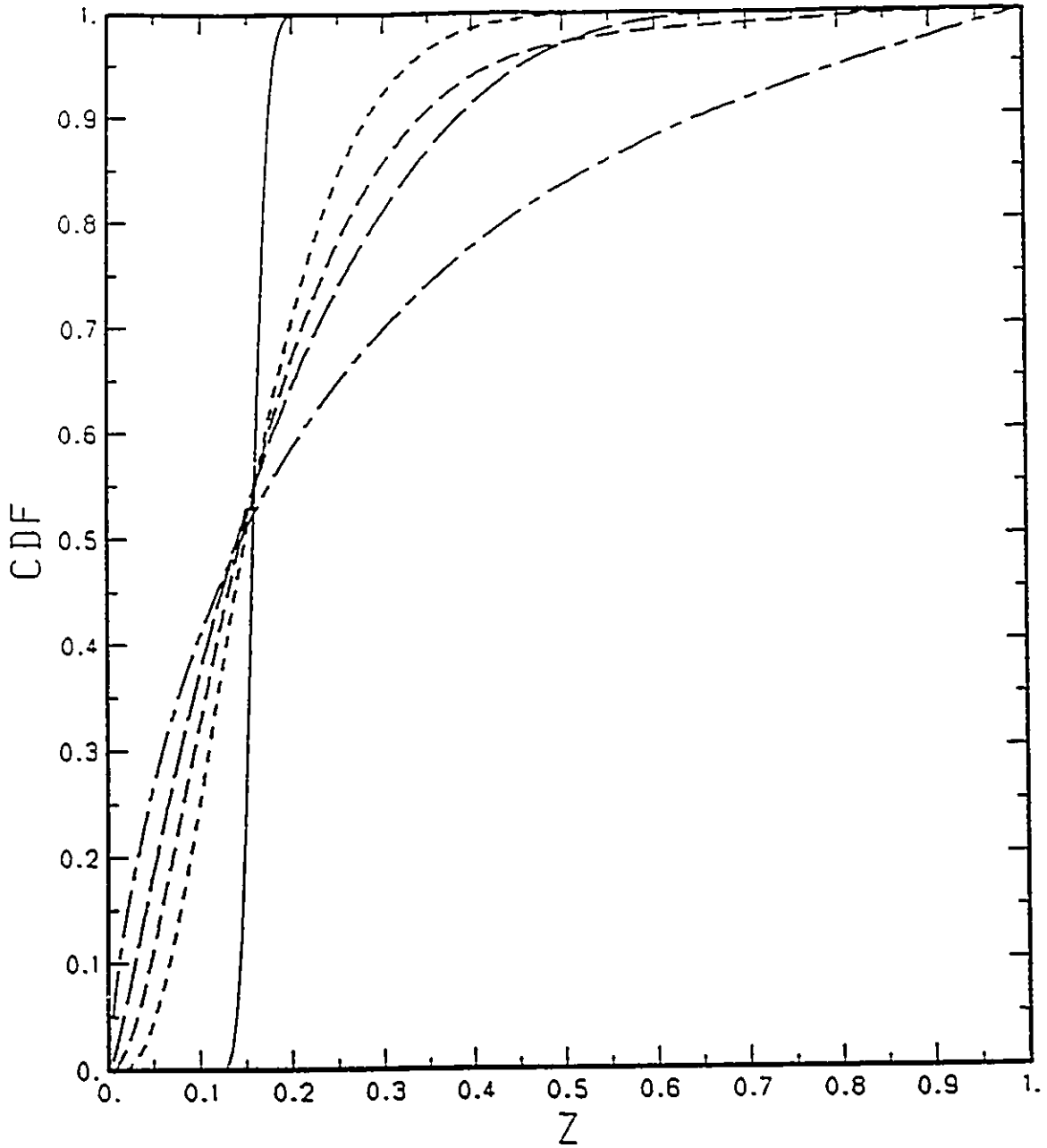
Figure 8.6: Distribution function of error of misclassification e_{12}: Outlier-Normal model 1



————————	= S1=0 S2=0	Ph1=1.0
- - - - - - - - - - - - - - -	= S1=0 S2=1	Ph1=2.0
— — — — — — — — — — —	= S1=1 S2=1	Ph1=2.0
— — — — — — — —	= S1=0 S2=2	Ph1=2.0
— - — - — - — - — -	= S1=2 S2=2	Ph1=2.0

Graph for Delta = 2.0
a1=a2=1 b1=b2=5

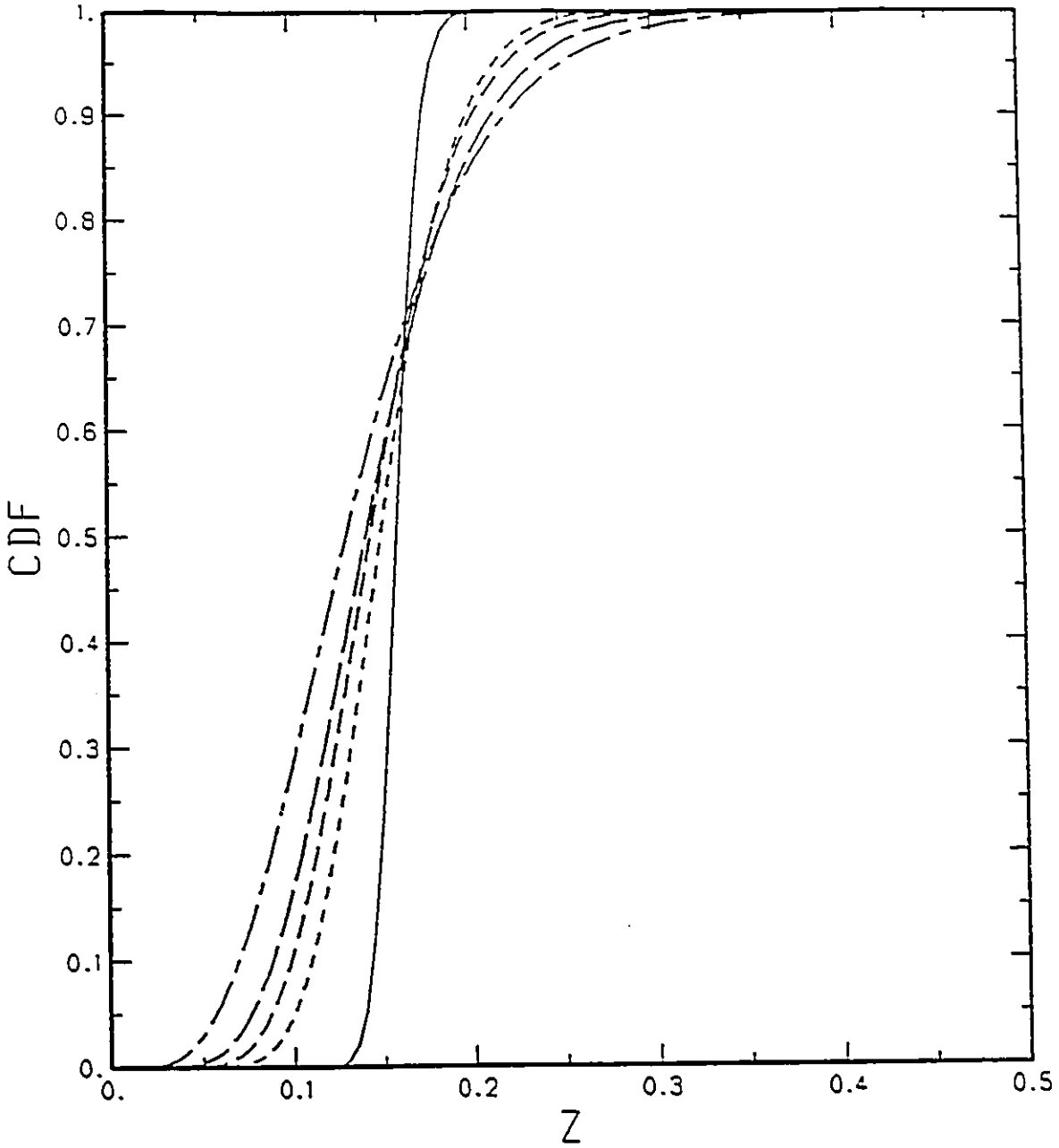
Figure 8.7: Distribution function of error of misclassification e_{12}: Outlier-Normal model 2



———	= S1=0	S2=0	Ph1=1.0
- - - - -	= S1=0	S2=1	Ph1=1.5
— — — —	= S1=1	S2=1	Ph1=1.5
— — — —	= S1=0	S2=2	Ph1=1.5
— - — -	= S1=2	S2=2	Ph1=1.5

Graph for Delta = 2.0
a1=a2=1 b1=b2=10

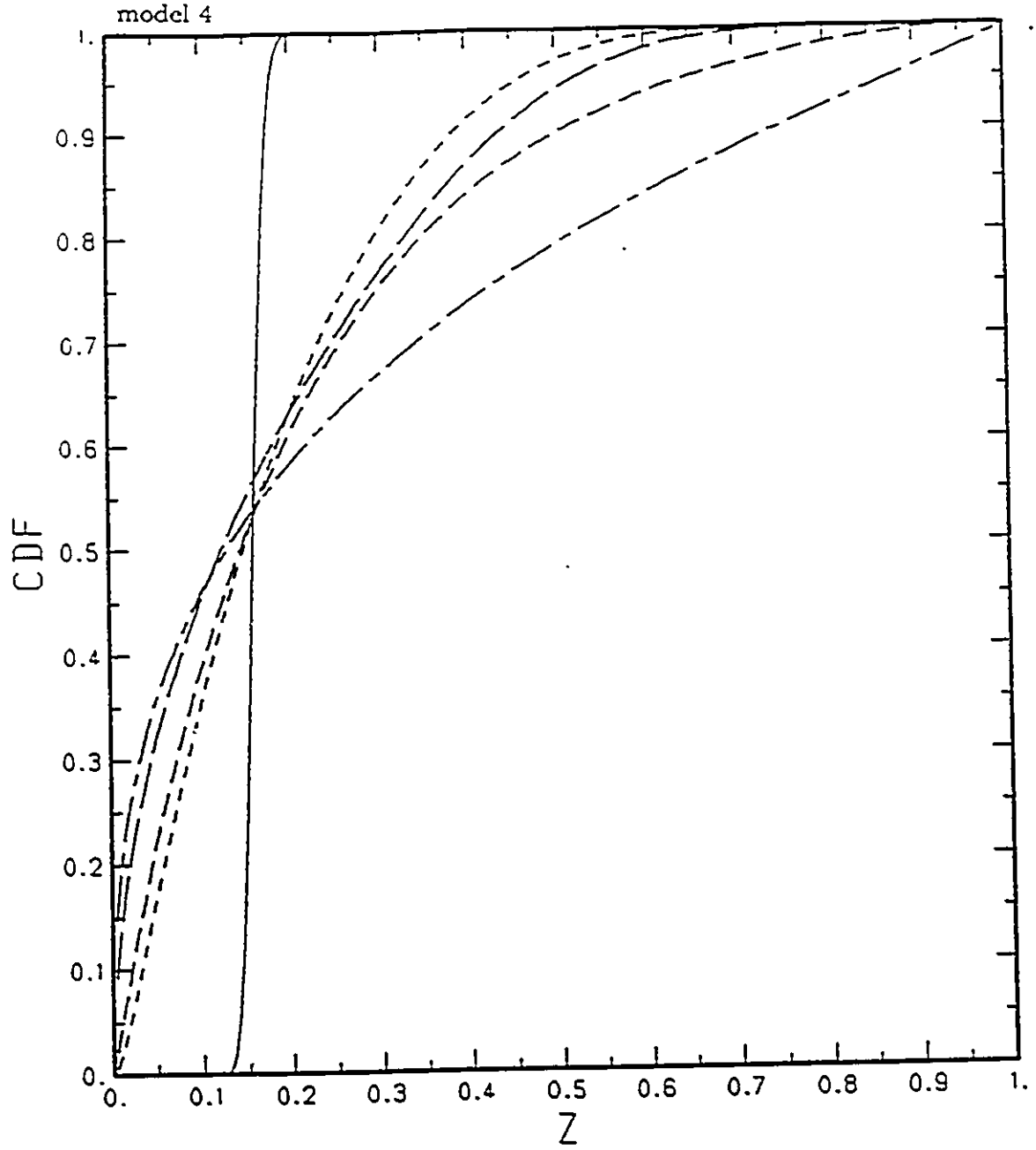
Figure 8.8: Distribution function of error of misclassification e_{12}: Outlier-Normal model 3



————————	= S1=0	S2=0	Ph1=1.0
- - - - - - - - - - - - - - - -	= S1=0	S2=1	Ph1=1.5
— — — — — — — — — — —	= S1=1	S2=1	Ph1=1.5
— — — — — — — — —	= S1=0	S2=2	Ph1=1.5
— - — - — - — - — -	= S1=2	S2=2	Ph1=1.5

Graph for Delta = 2.0
a1=a2=1 b1=b2=5

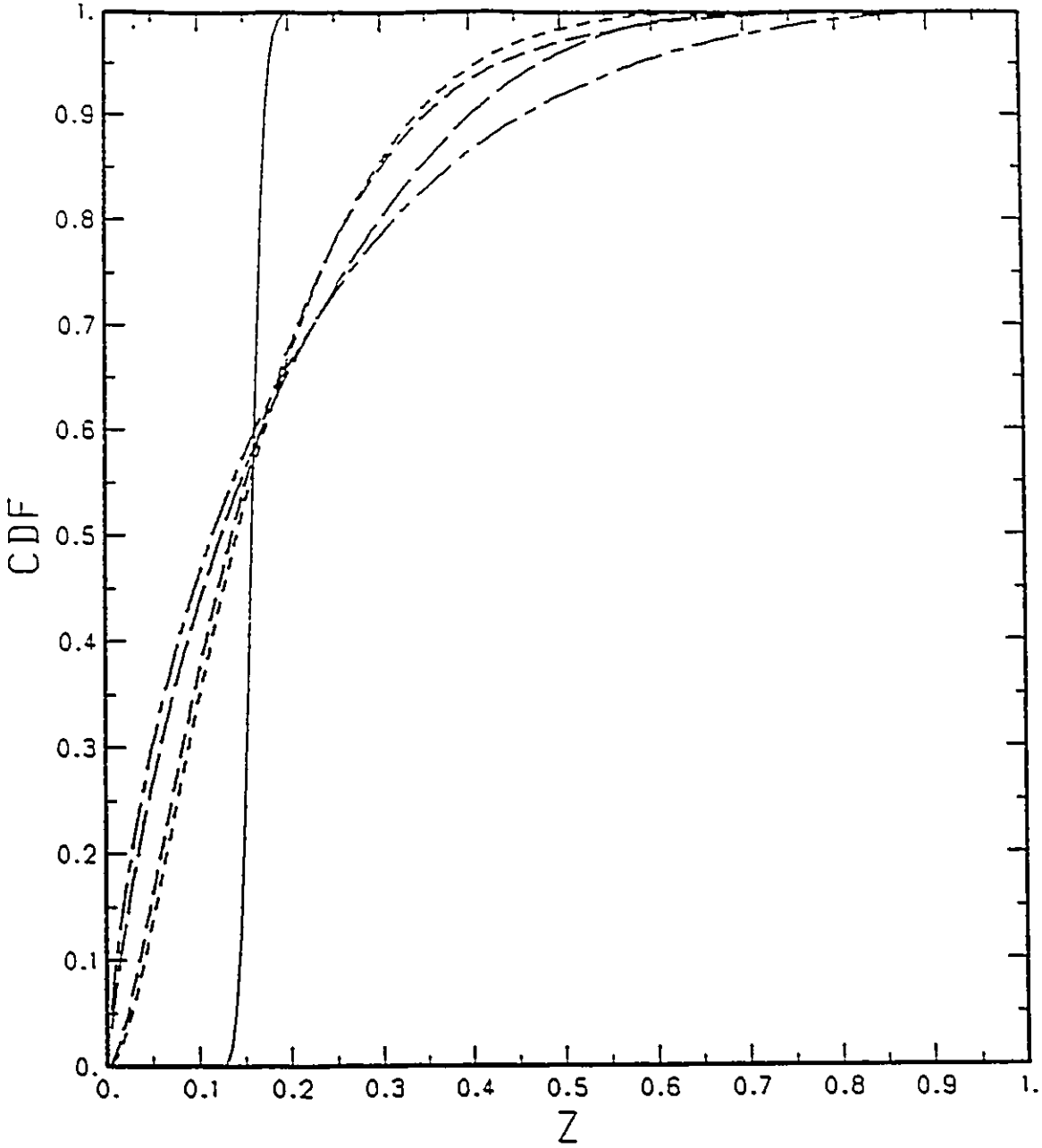
Figure 8.9: Distribution function of error of misclassification e_{12}: Outlier-Normal model 4



————————	= S1=0	S2=0	Ph1=1.0
- - - - - - - - - - - - - - -	= S1=0	S2=1	Ph1=2.0
— — — — — — — — — — —	= S1=1	S2=1	Ph1=2.0
— — — — — — — —	= S1=0	S2=2	Ph1=2.0
— - — - — - — - — -	= S1=2	S2=2	Ph1=2.0

Graph for Delta = 2.0
a1=a2=1 b1=b2=10

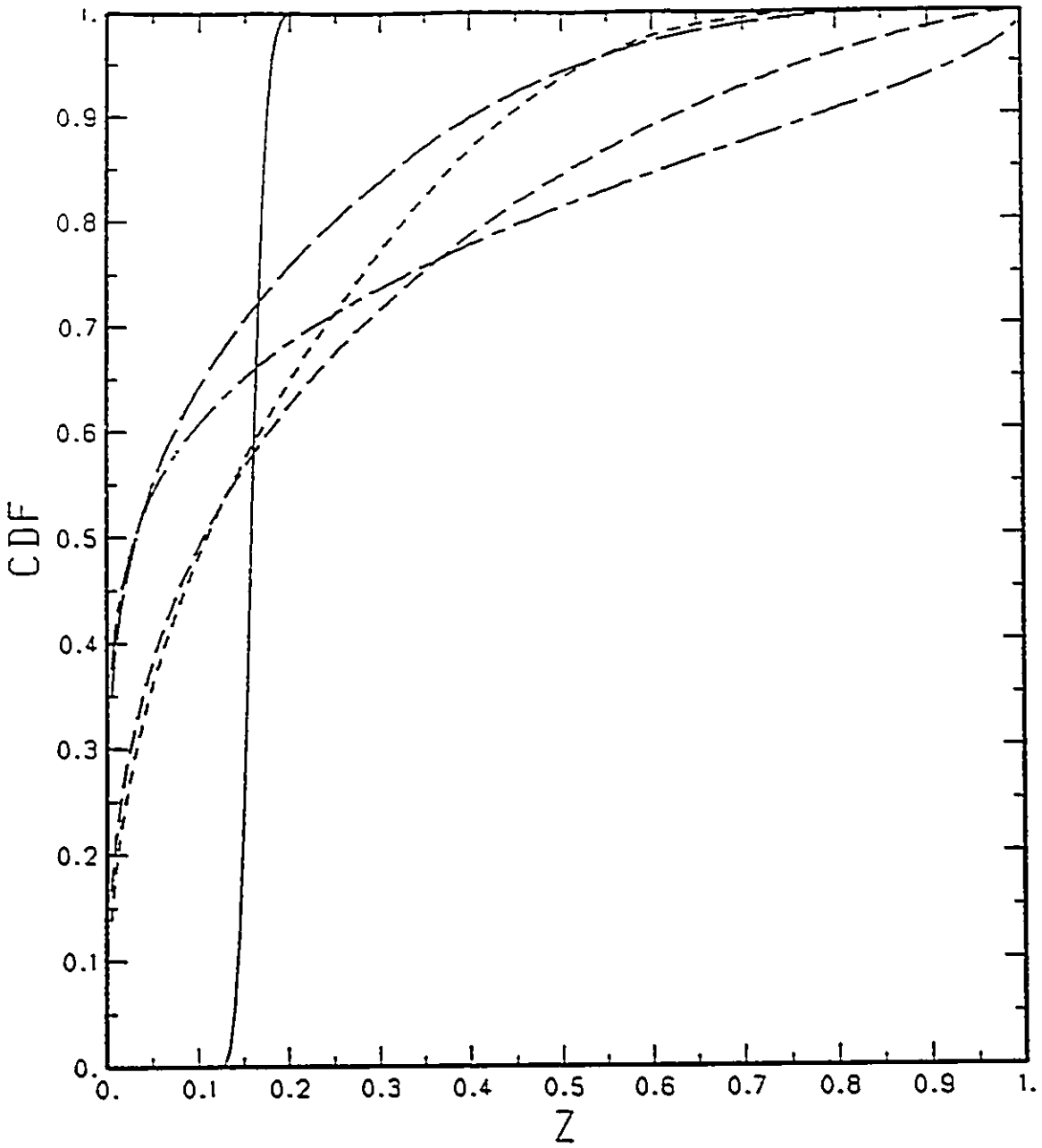
Figure S.10: Distribution function of error of misclassification e_{12}: Outlier-Normal model 5



————————	= S1=0 S2=0	Ph1=1.0
- - - - - - - - - - - - - - - -	= S1=0 S2=1	Ph1=3.0
— — — — — — — — — — —	= S1=1 S2=1	Ph1=3.0
— — — — — — — —	= S1=0 S2=2	Ph1=3.0
— - — - — - — - — - —	= S1=2 S2=2	Ph1=3.0

Graph for Delta = 2.0
a1=a2=1 b1=b2=5

Figure 8.11: Distribution function of error of misclassification: Outlier-Normal model 6



————————	= S1=0 S2=0 Ph1=1.0
- - - - - - - - - - - - - - - -	= S1=0 S2=1 Ph1=3.0
— — — — — — — — — — — —	= S1=1 S2=1 Ph1=3.0
— — — — — — — — — —	= S1=0 S2=2 Ph1=3.0
— - — - — - — - — - —	= S1=2 S2=2 Ph1=3.0

Graph for Delta = 2.0
a1=a2=1 b1=b2=10

Chapter 9

Classical and Robust Univariate Quadratic and Transformed Linear Classification Procedures Based On Dichotomous and Continuous Variables

9.1 Introduction

In the last chapter, we examined the robustness of the univariate linear classification procedure under the heterogeneity of variances and observed that the errors of misclassification tend to become large as the variances of the two populations get apart. In addition to the heterogeneity of variances if the samples also contain some outliers, we further noted that the errors of misclassification become very large. Hence, one may be interested in developing some classification procedures by starting with the assumption that the two population variances are unequal and also in developing robust analogues of them. In this chapter, by considering the classification problem based on a dichotomous and normal variables and assuming that the conditional variances of the two populations are unequal, we develop the

classical quadratic classification procedure and a transformed linear classification procedure. We also propose robust analogues of these procedures and carry out a comparative study of all these procedures.

It is important to realize that sets of measurements made on two classes of subjects may differ not only in their mean values but also in their degrees of dispersion. We illustrate this phenomenon through two practical situations considered by Penrose (1947).

1. In psychological data cases frequently arise where variance differences are significant. The distributions of intelligence quotients for males and females do not differ significantly in their means but have reportedly been found to differ in variance. Male scores have a greater degree of dispersion than female scores, and this fact can be used to help discriminate male from female scores. A measurement taken from the common mean value in either direction measures the relative probability of male and female scores. Extreme measurements are more likely to represent male scores and minimal deviations are more likely to be female scores.

2. In reaction experiments such as the word association test originally developed by Jung, a response which consists of a repetition of the stimulus word or too similar a word is abnormal as a response which consists of very unusual word. Similarly, a very brief time interval between stimulus and response is abnormal just as a very long interval or a failure to reply is abnormal. Nevertheless, peculiar word responses, long intervals or failures are more often obtained from mentally abnormal subjects. The distribution of reaction time intervals, therefore, for psychotic subjects has a longer mean and a larger variance than the distribution for normals. In order to use such a test as this for classification efficiently between normals and abnormal either by itself or in a battery of measurements, means and variances must both be taken into account.

Cavalli (1945) and Penrose (1947) have studied the problem of classification in the univariate case of unequal variances, and Smith (1947) has proposed the use of the likelihood ratio in the multivariate case of unequal covariance matrices. Anderson

and Bahadur (1961) have presented a few linear classification procedures in the multivariate case of unequal covariance matrices.

Tiku and Balakrishnan (1985, 1989) have also proposed a robust univariate quadratic classification procedure based on MML estimators and have shown that their procedure is robust to departures from normality in comparison to the classical quadratic classification procedure. By comparing these two procedures, when the error of misclassification e_{12} is at a pre-fixed level, under normal and some non-normal models, they have shown that the robust quadratic procedure is quite efficient and robust to departures from normality. The simulated e_{12} values of these two procedures, however, are usually somewhat larger than the pre-fixed level. So, Tiku and Balakrishnan (1985,1989) have also developed a transformed linear classification procedure and a robust analogue of it as alternatives to the quadratic classification procedures; these procedures have simpler asymptotic distribution theory, simulated e_{12} values closer to the presumed level, and $1-e_{21}$ values only slightly smaller than the corresponding quadratic classification procedures.

In this chapter, we consider the classification problem based on a dichotomous and a univariate normal variable when the conditional variances of the two populations are unequal. The basic model of this classification problem is described in Section 2. In Section 3, we derive the likelihood-ratio classification procedure in the case when all the parameters of the model are known. We discuss some approximations to the distribution of the classification statistic in this case. In Section 4, we consider the case when all the parameters of the model are unknown and then describe the classical quadratic and the robust quadratic classification procedures. The asymptotic determination of the cut-off points of these two procedures, when one of the errors of misclassification is at a pre-fixed level, is also discussed. In Section 5, we develop a transformed linear classification procedure and a robust analogue of it based on the MML estimators. In Section 6, we carry out a comparison of these four procedures by considering normal and also various non-normal models and some selected choices of the parameters. The robust transformed linear classification procedure proposed in Section 5 turns out to be the best of these four procedures as they possesses a simple asymptotic distribution theory, has its e_{12}

values to be quite close to the presumed level, has its $1 - e_{21}$ values almost the same as that of the corresponding quadratic classification procedure, and also remains robust to departures from normality.

The results presented in this chapter, therefore, generalize the results of Tiku and Balakrishnan (1985, 1989) to the case when the data contains a dichotomous variable and an associated continuous variable.

9.2 The Model

Let X be a Bernoulli variate with probability mass function

$$\Pr(X = x) = \theta^x(1-\theta)^{1-x}, \quad x = 0,\ 1,\ \ 0 < \theta < 1.$$

Let Y be a continuous random variable such that the conditional distribution of Y given $X = x$ is normal with mean $\mu(x) = \mu + x\delta$ and variance $\sigma^2(x) = \sigma^2 + x\gamma^2$, where μ, δ, σ, and γ are finite but unknown constants. Clearly, the probability density function of

$$\mathbf{W} = \begin{bmatrix} X \\ Y \end{bmatrix}$$

is given by

$$f(\mathbf{w}) = \theta^x(1-\theta)^{1-x}\phi(y; \mu(x), \sigma^2(x)) \tag{9.1}$$

where for given $X = x$, $\phi(y; \mu(x), \sigma^2(x))$ denotes the normal distribution described above. As we have shown in Chapter 3, the expected value of the random vector $\mathbf{W}$ is

$$E(\mathbf{W}) = \begin{bmatrix} \theta \\ \mu + \theta\delta \end{bmatrix}$$

and the variance-covariance matrix is

$$Cov(\mathbf{W}) = \begin{bmatrix} \theta(1-\theta) & \theta(1-\theta)\delta \\ \theta(1-\theta)\delta & \sigma^2 + \gamma^2\theta + \theta(1-\theta)\delta^2 \end{bmatrix}$$

The marginal distribution of y is

$$\begin{aligned} g(y) &= \sum_{x=0}^{1} \theta^x(1-\theta)^{1-x}\phi(\mu(x), \sigma^2(x)) && (9.2) \\ &= (1-\theta)\phi(\mu(0), \sigma^2(0)) + \theta\phi(\mu(1), \sigma^2(1)) && (9.3) \end{aligned}$$

i.e. a two-component univariate mixture normal. Note that X and Y are independent if and only if $\mu(0) = \mu(1)$ and $\sigma^2(0) = \sigma^2(1)$, i.e. $\delta = \gamma^2 = 0$. For the problem of classifying an observation w_0 into one of two populations, Π_1 and Π_2, it is assumed that if $\mathrm{w}_0 = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ is from Π_i, then its density function is given by

$$f_i(\mathrm{w}_0) = \theta_i^{x_0}(1-\theta_i)^{1-x_0}\phi(\mu_i(x_0), \sigma_i^2(x_0)), \quad i = 1, 2. \tag{9.4}$$

9.3 Classification when the population parameters are known

In this section, by assuming that the parameters of the two populations are known we proceed to derive the likelihood ratio classification procedure and the associated probabilities of misclassification.

For the model specified by (9.4), the log likelihood ratio is

$$\begin{aligned}
V(\mathrm{w}_0) &= \log\left[\frac{f_1(\mathrm{w}_0)}{f_2(\mathrm{w}_0)}\right] \\
&= \log\left[\frac{\sigma_2(x_0)}{\sigma_1(x_0)}\exp\left\{-\frac{1}{2}\left[\frac{y_0-\mu_1(x_0)}{\sigma_1(x_0)}\right]^2 + \frac{1}{2}\left[\frac{y_0-\mu_2(x_0)}{\sigma_2(x_0)}\right]^2\right\}\right. \\
&\quad \left.\left[\frac{\theta_1}{\theta_2}\right]^{x_0}\left[\frac{1-\theta_1}{1-\theta_2}\right]^{(1-x_0)}\right] \\
&= \left[\frac{y_0-\mu_2(x_0)}{\sigma_2(x_0)}\right]^2 - \left[\frac{y_0-\mu_1(x_0)}{\sigma_1(x_0)}\right]^2 + 2x_0 log\left[\frac{\theta_1}{\theta_2}\right] + 2(1-x_0)log\left[\frac{1-\theta_1}{1-\theta_2}\right] \\
&\quad -log\left[\frac{\sigma_1^2(x_0)}{\sigma_2^2(x_0)}\right] \quad (x_0 = 0, 1)
\end{aligned} \tag{9.5}$$

Therefore, the classification rule classifies $\mathrm{w}_0 = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}$ into Π_1 or Π_2 according as

$$V(\mathrm{w}_0) \gtrless C. \tag{9.6}$$

Let e_{12} and e_{21} denote the errors of misclassification, i.e., e_{12} is the probability of classifying w_0 into Π_2 when in fact it has come from Π_1, and similarly e_{21} is the

probability of classifying w_0 into Π_1 when in fact it has come from Π_2. The cutoff point C in (9.6) is so determined that e_{12} assumes a pre-assigned value. To determine the constant C for a specified level of e_{12}, we require the distribution of $V(w_0)$ which unfortunately does not assume a manageable form. Therefore, we use an approximation similar to the one given by Tiku and Balakrishnan (1985, 1989).

Theorem 15 *Under the assumption of normality, $V(w)$ has exact normal distribution for $\frac{\sigma_1^2(x)}{\sigma_2^2(x)} = 1$ and can be approximated by a normal distribution if the quantity $\frac{\sigma_1^2(x)}{\sigma_2^2(x)}$ takes values in the vicinity of 1.*

Proof: If $w \in \Pi_1$ then

$$Z = \frac{y - \mu_1(x)}{\sigma_1(x)} \sim N(0,1).$$

This implies

$$\begin{aligned}
\left[\frac{y-\mu_2(x)}{\sigma_2(x)}\right]^2 &= \left[\frac{y-\mu_1(x)}{\sigma_2(x)} + \frac{\mu_1(x)-\mu_2(x)}{\sigma_2(x)}\right]^2 \\
&= \left[\frac{\sigma_1(x)}{\sigma_2(x)}\frac{y-\mu_1(x)}{\sigma_1(x)} + \frac{\mu_1(x)-\mu_2(x)}{\sigma_2(x)}\right]^2 \\
&= [\phi_x Z + \delta_x]^2 \\
&= \phi_x^2 Z^2 + 2\phi_x\delta_x Z + \delta_x^2 \qquad (9.7)
\end{aligned}$$

where $\phi_x^2 = \frac{\sigma_1^2(x)}{\sigma_2^2(x)}$, and $\delta_x^2 = \frac{(\mu_1(x)-\mu_2(x))^2}{\sigma_2^2(x)}$. Hence, we can write

$$V(w|w \in \Pi_1) = (\phi_x^2 - 1)Z^2 + 2\phi_x\delta_x Z + \delta_x^2 + 2\lambda(x,\theta_1,\theta_2) - log(\phi_x^2), \qquad (9.8)$$

where $\lambda(x,\theta_1,\theta_2) = x log\left[\frac{\theta_1}{\theta_2}\right] + (1-x) log\left[\frac{1-\theta_1}{1-\theta_2}\right]$. If $w \in \Pi_2$ then

$$Z = \frac{y - \mu_2(x)}{\sigma_2(x)} \sim N(0,1).$$

This implies

$$\begin{aligned}
\left[\frac{y-\mu_1(x)}{\sigma_1(x)}\right]^2 &= \left[\frac{y-\mu_2(x)}{\sigma_1(x)} + \frac{\mu_2(x)-\mu_1(x)}{\sigma_1(x)}\right]^2 \\
&= \left[\frac{\sigma_2(x)}{\sigma_1(x)}\frac{y-\mu_2(x)}{\sigma_2(x)} + \frac{\mu_2(x)-\mu_1(x)}{\sigma_1(x)}\right]^2 \\
&:= [\phi_x' Z - \delta_x']^2 \\
&= \phi_x'^2 Z^2 - 2\phi_x'\delta_x' Z + \delta_x'^2, \qquad (9.9)
\end{aligned}$$

where $\phi'_x = \frac{\sigma_2(x)}{\sigma_1(x)}$, and $\delta'_x = \frac{(\mu_1(x)-\mu_2(x))}{\sigma_1(x)}$. Hence, we can write

$$V(\mathrm{w}|\mathrm{w} \in \Pi_2) = (1-\phi'^2_x)Z^2 + 2\phi'_x\delta'_x Z - \delta'^2_x + 2\lambda(x,\theta_1,\theta_2) - log(\phi^2_x) \qquad (9.10)$$

Let $V = V(\mathrm{w}|\mathrm{w} \in \Pi_1) - E(V(\mathrm{w}|\mathrm{w} \in \Pi_1))$ and let $a = \phi^2_x - 1, \quad b = \phi_x\delta_x$. Then, we have

$$\begin{aligned}
V &= aZ^2 + 2bZ - a \\
V^2 &= a^2Z^4 + 4abZ^3 - 2a^2Z^2 + 4b^2Z^2 - 4abZ + a^2 \\
V^3 &= a^3Z^6 + 6a^2bZ^5 - 3a^3Z^4 + 12ab^2Z^4 - 12a^2bZ^3 \\
&\quad +3a^3Z^2 + 8b^3Z^3 - 12b^2aZ^2 + 6ba^2Z - a^3 \\
V^4 &= a^4Z^8 - 4a^4Z^6 + 6a^4z^4 - 4a^4Z^2 + 16b^4Z^4 + a^4 + 8a^3bZ^7 \\
&\quad +24a^2b^2Z^6 - 24a^3bZ^5 + 32ab^3Z^5 - 48a^2b^2Z^4 + 24a^3bZ^3 \\
&\quad -32b^3aZ^3 + 24b^2a^2Z^2 - 8ba^3Z \qquad (9.11)
\end{aligned}$$

Since for $r = 1, 2, \ldots\ldots$

$$\begin{aligned}
E\left[Z^{2r}\right] &= \frac{(2r)!}{2^r(r!)} \\
\text{and} \quad E\left[Z^{2r+1}\right] &= 0,
\end{aligned}$$

we have

$$\begin{aligned}
E(V^2) &= 3a^2 - 2a^2 + 4b^2 + a^2 \\
E(V^3) &= 15a^3 - 9a^3 + 36ab^2 + 3a^3 - 12ab^2 - a^3 \\
E(V^4) &= 105a^4 - 60a^4 + 18a^4 - 4a^4 + 48b^4 + a^4 + 360a^2b^2 - 144a^2b^2 + 24a^2b^2
\end{aligned}$$

i.e.

$$\begin{aligned}
E(V^2) &= 2a^2 + 4b^2 \\
E(V^3) &= 8a^3 + 24ab^2 \\
E(V^4) &= 60a^4 + 240a^2b^2 + 48b^4
\end{aligned}$$

Therefore, we obtain

$$E(V(\mathrm{w})|\mathrm{w} \in \Pi_1) = (\phi^2_x - 1) + \delta^2_x + 2\lambda(x,\theta_1,\theta_2) - log(\phi^2_x)$$

$$
\begin{aligned}
Var(V(\mathrm{w})|\mathrm{w} \in \Pi_1) &= 2(\phi_x^2 - 1)^2 + 4\delta_x^2\phi_x^2, \\
\mu_3(V(\mathrm{w})|\mathrm{w} \in \Pi_1) &= 8(\phi_x^2 - 1)^3 + 24\delta_x^2\phi_x^2(\phi_x^2 - 1), \\
\text{and} \quad \mu_4(V(\mathrm{w})|\mathrm{w} \in \Pi_1) &= 60(\phi_x^2 - 1)^4 + 240\delta_x^2\phi_x^2(\phi_x^2 - 1)^2 + 48\delta_x^4\phi_x^4, \quad (9.12)
\end{aligned}
$$

Hence, the coefficients of skewness and kurtosis, κ_1 and κ_2, are given by

$$
\begin{aligned}
\kappa_1 &= 36(\phi_x - 1)^2\delta_x^2 + \bigcirc((\phi_x - 1)^2) \quad (9.13) \\
\text{and} \quad \kappa_2 &= 3 + \left[-\left(\frac{3}{2}\right)\frac{4\delta_x^2 + 8}{\delta_x^2} - 12 + \left(\frac{1}{16}\right)\frac{960\delta_x^2 + 288\delta_x^4}{\delta_x^4}\right](\phi_x - 1)^2 + \\
&\quad \bigcirc((\phi_x - 1)^2) \quad (9.14)
\end{aligned}
$$

upto the second term of Taylor series expansion at $\phi_x = 1$. Since the coefficient of skewness and kurtosis of $V(\mathrm{w})$ are exactly equal to 0 and 3, respectively, when $\phi_x^2 = \frac{\sigma_1^2(x)}{\sigma_2^2(x)} = 1$ (the equal conditional variance case), and reasonably close to 0 and 3 if ϕ_x is in the vicinity of 1, the distribution of $V(\mathrm{w})$ may be approximated reasonably by a normal distribution, if $\mathrm{w} \in \Pi_1$.

Let $V_1 = V(\mathrm{w}|\mathrm{w} \in \Pi_2) - E(V(\mathrm{w}|\mathrm{w} \in \Pi_2))$ and let $a = 1 - \phi_x'^2$, $b = \phi_x'\delta_x'$. Then from (9.10) we have

$$
V_1 = aZ^2 + 2bZ - a. \quad (9.15)
$$

From this and (9.12), we obtain

$$
\begin{aligned}
E(V(\mathrm{w})|\mathrm{w} \in \Pi_2) &= (1 - \phi_x'^2) - \delta_x'^2 + 2\lambda(x, \theta_1, \theta_2) - log(\phi_x^2) \\
Var(V(\mathrm{w})|\mathrm{w} \in \Pi_2) &= 2(1 - \phi_x'^2)^2 + 4\delta_x'^2\phi_x'^2, \\
\mu_3(V(\mathrm{w})|\mathrm{w} \in \Pi_2) &= 8(1 - \phi_x'^2)^3 + 24\delta_x'^2\phi_x'^2(1 - \phi_x'^2), \\
&\text{and} \\
\mu_4(V(\mathrm{w})|\mathrm{w} \in \Pi_1) &= 60(1 - \phi_x'^2)^4 + 240\delta_x'^2\phi_x'^2(1 - \phi_x'^2)^2 + 48\delta_x'^4\phi_x'^4, \quad (9.16)
\end{aligned}
$$

Hence, the coefficients of skewness and kurtosis, κ_1 and κ_2, are given by

$$
\begin{aligned}
\kappa_1 &= 36(1 - \phi_x')^2\delta_x^2 + \bigcirc((1 - \phi_x')^2) \quad (9.17) \\
\text{and} \quad \kappa_2 &= 3 + \left[-\left(\frac{3}{2}\right)\frac{4\delta_x'^2 + 8}{\delta_x'^2} - 12 + \left(\frac{1}{16}\right)\frac{960\delta_x'^2 + 288\delta_x'^4}{\delta_x'^4}\right](1 - \phi_x')^2 + \\
&\quad \bigcirc((1 - \phi_x')^2) \quad (9.18)
\end{aligned}
$$

upto the second term of Taylor series expansion at $\phi'_x = 1$. Since the coefficient of skewness and kurtosis of $V(\mathrm{w})$ are exactly equal to 0 and 3, respectively, when $\phi'^2_x = \frac{\sigma_2^2(x)}{\sigma_1^2(x)} = 1$ (the equal conditional variance case), and reasonably close to 0 and 3 if ϕ'_x is in the vicinity of 1,, the distribution of $V(\mathrm{w})$ may be approximated reasonably by a normal distribution, if $\mathrm{w} \in \Pi_2$. ∎

The normal approximation may be suitable only in some range of δ_x^2 and ϕ_x^2, and in general a four-moment Pearson curve may be fitted by making use of the tables of Johnson et al. (1963). Suppose we use the normal approximation, and we can write

$$
\begin{aligned}
e_{12} &= \Pr[V(\mathrm{w}_0) < C | \mathrm{w}_0 \in \Pi_1] \\
&= (1-\theta_1)\Pr\left[V\left(\begin{bmatrix} 0 \\ y_0 \end{bmatrix}\right) < C | \mathrm{w}_0 \in \Pi_1\right] + \theta_1 \Pr\left[V\left(\begin{bmatrix} 1 \\ y_0 \end{bmatrix}\right) < C | \mathrm{w}_0 \in \Pi_1\right] \\
&= (1-\theta_1)\Phi\left[\frac{C - (\phi_0^2 + \delta_0^2 - 1) - 2\lambda(0,\theta_1,\theta_2) + \log(\phi_0^2)}{\sqrt{2(\phi_0^2-1)^2 + 4\phi_0^2\delta_0^2}}\right] \\
&\quad + \theta_1 \Phi\left[\frac{C - (\phi_1^2 + \delta_1^2 - 1) - 2\lambda(1,\theta_1,\theta_2) + \log(\phi_1^2)}{\sqrt{2(\phi_1^2-1)^2 + 4\phi_1^2\delta_1^2}}\right],
\end{aligned}
$$

where $\Phi(z) = \int_{-\infty}^{z} \phi(z)dz$, and $\phi(z)$ is the standard normal density function. Similarly, we can write

$$
\begin{aligned}
e_{21} &= \Pr[V(\mathrm{w}_0) > C | \mathrm{w}_0 \in \Pi_2] \\
&= (1-\theta_2)\Pr\left[V\left(\begin{bmatrix} 0 \\ y_0 \end{bmatrix}\right) > C | \mathrm{w}_0 \in \Pi_2\right] + \theta_2 \Pr\left[V\left(\begin{bmatrix} 1 \\ y_0 \end{bmatrix}\right) > C | \mathrm{w}_0 \in \Pi_2\right], \\
&= (1-\theta_2)\Phi\left[-\frac{C - (1 - \phi_0'^2 - \delta_0'^2) - 2\lambda(0,\theta_1,\theta_2) + \log(\phi_0^2)}{\sqrt{2(1-\phi_0'^2)^2 + 4\phi_0'^2\delta_0'^2}}\right] \\
&\quad + \theta_2 \Phi\left[-\frac{C - (1 - \phi_1'^2 - \delta_1'^2) - 2\lambda(1,\theta_1,\theta_2) + \log(\phi_1^2)}{\sqrt{2(1-\phi_1'^2)^2 + 4\phi_1'^2\delta_1'^2}}\right].
\end{aligned}
$$

The cut-off point C may be obtained as a solution of the equation

$$
H[C;\theta_1,\theta_2,\phi_0,\phi_1,\delta_0,\delta_1] = (1-\theta_1)\Phi\left[\frac{C - (\phi_0^2 + \delta_0^2 - 1) - 2\lambda(0,\theta_1,\theta_2) - \log(\phi_0^2)}{\sqrt{2(\phi_0^2-1)^2 + 4\phi_0^2\delta_0^2}}\right]
$$

$$+\theta_1\Phi\left[\frac{C-(\phi_1^2+\delta_1^2-1)-2\lambda(1,\theta_1,\theta_2)-\log(\phi_1^2)}{\sqrt{2(\phi_1^2-1)^2+4\phi_1^2\delta_1^2}}\right]$$
$$= \alpha, \tag{9.19}$$

when the error of misclassification e_{12} is pre-fixed as α. Note that the right-hand side of (9.19) is a monotonic increasing function in C and has, therefore, a unique solution.

9.4 Quadratic classification procedures when population parameters are unknown

In practice, however, the population parameters $\mu_i(x)$, $\sigma_i(x)$, θ_i, for $i = 1, 2$, are unknown, and therefore ϕ_x^2 , δ_x^2 are unknown. In this section, we consider two methods of estimation of population parameters, first method being the traditional ML estimation method and the second method being the MML estimation method. For this, let us assume that $\begin{bmatrix} x_{1i} \\ y_{1i} \end{bmatrix}$, $i = 1, 2, \ldots, n_1$, is a random sample of size n_1 from population Π_1 which is the model specified in (9.4) with θ_1, $\mu_1(x)$ and $\sigma_1(x)$ and $\begin{bmatrix} x_{2i} \\ y_{2i} \end{bmatrix}$, $i = 1, 2, \ldots, n_2$, is a random sample of size n_2 from population Π_2 which is the model specified in (9.4) with θ_2, $\mu_2(x)$ and $\sigma_2(x)$.

9.4.1 Classical quadratic classification procedure

In this section, we replace (θ_i, $\mu_i(x)$, $\sigma_i(x)$) by their ML estimators ($\frac{n_i(x)}{n_i}$, $\bar{y}_i(x)$, $s_i^2(x)$). Since the distribution of $\frac{s_1^2(x)/\sigma_1^2(x)}{s_2^2(x)/\sigma_2^2(x)}$ is $F(n_1(x)-1, n_2(x)-1)$, we have

$$E\left(\frac{s_1^2(x)/\sigma_1^2(x)}{s_2^2(x)/\sigma_2^2(x)}\right) = \frac{n_2(x)-1}{n_2(x)-3},$$

and hence we may use

$$\hat{\phi}_x^2 = \left(\frac{n_2(x)-3}{n_2(x)-1}\right)\frac{s_1^2(x)}{s_2^2(x)}, \tag{9.20}$$

as an estimator of ϕ_x^2. An obvious estimator for δ_x^2 would be the statistic obtained by substituting the MLE's of $\mu_1(x)$, $\mu_2(x)$ and $\sigma_2(x)$ in δ_x^2, i.e. $\frac{(\mu_1(x)-\bar{y}_2(x))^2}{s_2^2(x)}$. Derivation of the expected value of this statistic is as follows:

$$E\left(\frac{(\bar{y}_1(x)-\bar{y}_2(x))^2}{s_2^2(x)}\right) = E\left(\bar{y}_1(x)-\bar{y}_2(x)\right)^2 E\left(\frac{1}{s_2^2(x)}\right),$$

which follows from the fact that $\bar{y}_1(x)$, $\bar{y}_2(x)$ and $s_2^2(x)$ are statistically independent. Since $\bar{y}_1(x)-\bar{y}_2(x)$ is normally distributed with mean $\mu_1(x)-\mu_2(x)$ and variance $\frac{\sigma_1^2(x)}{n_1(x)}+\frac{\sigma_2^2(x)}{n_2(x)}$, we immediately have

$$E\left(\bar{y}_1(x)-\bar{y}_2(x)\right)^2 = (\mu_1(x)-\mu_2(x))^2+\frac{\sigma_1^2(x)}{n_1(x)}+\frac{\sigma_2^2(x)}{n_2(x)}.$$

Since $(n_2(x)-1)\frac{s_2^2(x)}{\sigma_2^2(x)}$ is distributed as chi-square with $n_2(x)-1$ degrees of freedom, we have

$$E\left(\frac{1}{(n_2(x)-1)\frac{s_2^2(x)}{\sigma_2^2(x)}}\right) = \frac{1}{n_2(x)-3},$$

$$\text{which implies} \quad E\left(\frac{1}{s_2^2(x)}\right) = \frac{n_2(x)-1}{(n_2(x)-3)\sigma_2^2(x)}$$

As a results, we have

$$E\left(\frac{(\bar{y}_1(x)-\bar{y}_2(x))^2}{s_2^2(x)}\right) = \frac{n_2(x)-1}{n_2(x)-3}\frac{(\mu_1(x)-\mu_2(x))^2}{\sigma_2^2(x)} + \frac{n_2(x)-1}{(n_2(x)-3)}\left(\frac{\sigma_1^2(x)}{\sigma_2(x)}\frac{1}{n_1(x)}+\frac{1}{n_2(x)}\right),$$

i.e.

$$\frac{n_2(x)-3}{n_2(x)-1}E\left(\frac{(\bar{y}_1(x)-\bar{y}_2(x))^2}{s_2^2(x)}\right) = \frac{(\mu_1(x)-\mu_2(x))^2}{\sigma_2^2(x)} + \frac{\sigma_1^2(x)}{\sigma_2^2(x)}\frac{1}{n_1(x)}+\frac{1}{n_2(x)}.$$

Now writing

$$\hat{\delta}_x^2 = \left(\frac{n_2(x)-3}{n_2(x)-1}\right)\frac{(\bar{y}_1(x)-\bar{y}_2(x))^2}{s_2^2(x)}$$

and substituting the estimator for ϕ_x^2 in (9.20) we obtain an unbiased estimator of δ_x

$$\hat{\delta}_x^2 - \frac{\hat{\phi}_x^2}{n_1(x)} - \frac{1}{n_2(x)}. \tag{9.21}$$

But this unbiased estimator may take on inadmissible negative values as mentioned by Lachenbruch and Mickey (1968). So, we may take $\hat{\delta}_x^2$ itself to be an estimator of δ_x^2 and it is of interest to note here from (9.21) that it is an asymptotically unbiased estimator of δ_x^2. By using these estimators in the likelihood-ratio classification procedure given in (9.6), we derive the classical quadratic classification procedure as one which classifies the new observation $\mathbf{w}_0$ into Π_1 or Π_2 according as

$$\Lambda(\mathbf{w}_0) = \Lambda[y_0; \hat{\theta}_1, \hat{\theta}_2, \bar{y}_1(x_0), \bar{y}_2(x_0), s_1^2(x_0), s_2(x_0)] \gtrless \hat{C}, \quad (x_0 = 0, 1), \tag{9.22}$$

where

$$\begin{aligned} \Lambda(\mathbf{w}) \;=\; & \left[\frac{y - \bar{y}_2(x)}{s_2(x)}\right]^2 - \left[\frac{y - \bar{y}_1(x)}{s_1(x)}\right]^2 + 2x log\left[\frac{\hat{\theta}_1}{\hat{\theta}_2}\right] + 2(1-x) log\left[\frac{1-\hat{\theta}_1}{1-\hat{\theta}_2}\right] \\ & -log(\hat{\phi}_x^2) \end{aligned} \tag{9.23}$$

and the cutoff point $\hat{C}$ is obtained as the solution of

$$H[\hat{C}; \hat{\theta}_1, \hat{\theta}_2, \hat{\phi}_0, \hat{\phi}_1, \hat{\delta}_0, \hat{\delta}_1] = \alpha, \tag{9.24}$$

where the function H is as defined in Eq. (9.19) and α is the level at which the error of misclassification e_{12} is pre-fixed.

Theorem 16 *Asymptotically the distribution of statistic* $\Lambda(\mathbf{w})$ *in (9.23) is similar to the distribution of* $V(\mathbf{w})$ *in (9.5).*

Proof: This theorem follows from the fact that when $n_1(x)$ and $n_2(x)$ tend to infinity the ML estimators tend to their population parameters. i.e. $\bar{y}_i(x)$ and $s_i^2(x)$ tend to $\mu_i(x)$ and $\sigma_i^2(x)$. Therefore, the estimators $\hat{\phi}_x^2$ and $\hat{\delta}_x^2$ tend to ϕ_x^2 and δ_x^2 as $n_1(x)$ and $n_2(x)$ tend to infinity. Hence, the proof. ∎

All the developments above have been based on the assumption that the populations are normally distributed. But, the underlying distribution is hardly

ever exactly normal. In fact, the underlying distribution is hardly ever known exactly, and that problem raises the question of how robust the $\Lambda(\mathrm{w}_0)$ procedure in (9.22) is to departures from normality. We will see later that it is not very robust. In fact, the $\Lambda(\mathrm{w}_0)$ procedure has in general e_{12} value lower (or close to) than the pre-specified level which, of course, is not an undesirable feature. However, the $\Lambda(\mathrm{w}_0)$ procedure has high e_{21} values which is indeed an undesirable feature. In the next section, we develop a robust procedure, i.e., a procedure which has e_{12} values close to (at any rate not appreciably larger than) a pre-specified level for normal and numerous symmetric non-normal populations and at the same time has substantially smaller (and stable) e_{21} values than the $\Lambda(\mathrm{w}_0)$ procedure.

9.4.2 Robust quadratic classification procedure

We obtain robust quadratic classification procedure analogous to the classical one in (9.22) by replacing the classical estimators $\bar{y}_1(x)$, $\bar{y}_2(x)$, $s_1^2(x)$ and $s_2^2(x)$ by the corresponding MML estimators $\hat{\mu}_1(x)$, $\hat{\mu}_2(x)$, $\hat{\sigma}_1^2(x)$ and $\hat{\sigma}_2^2(x)$; here, $\hat{\mu}_1(x)$ and $\hat{\sigma}_1(x)$ are the MML estimators of $\mu_1(x)$ and $\sigma_1(x)$ obtained from the sample $y_{1i}(x), i = 1, \ldots, n_1(x)$, and $\hat{\mu}_2(x)$ and $\hat{\sigma}_2(x)$ are the MML estimators of $\mu_2(x)$ and $\sigma_2(x)$ obtained from the sample $y_{2i}(x), i = 1, \ldots, n_2(x)$. Then the robust two-way quadratic classification procedure, with e_{12} pre-fixed as α, is to classify the observation $\mathrm{w}_0 = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ into Π_1 or Π_2 according as

$$\Lambda_R(\mathrm{w}_0) = \Lambda_R[y_0; \hat{\theta}_1, \hat{\theta}_2, \hat{\mu}_1(x_0), \hat{\mu}_2(x_0), \hat{\sigma}_1^2(x_0), \hat{\sigma}_2(x_0)] \gtrless \hat{C}_R, \quad (x_0 = 0, 1), \quad (9.25)$$

where

$$\begin{aligned} \Lambda_R(\mathrm{w}) \;=\; & \left[\frac{y - \hat{\mu}_2(x)}{\hat{\sigma}_2(x)}\right]^2 - \left[\frac{y - \hat{\mu}_1(x)}{\hat{\sigma}_1(x)}\right]^2 + 2x log\left[\frac{\hat{\theta}_1}{\hat{\theta}_2}\right] + 2(1-x) log\left[\frac{1-\hat{\theta}_1}{1-\hat{\theta}_2}\right] \\ & -log(\hat{\phi}_{x,R}^2) \end{aligned} \quad (9.26)$$

The cutoff point $\hat{C}_R$ is obtained as the solution of

$$H[\hat{C}; \hat{\theta}_1, \hat{\theta}_2, \hat{\phi}_{0,R}, \hat{\phi}_{1,R}, \hat{\delta}_{0,R}, \hat{\delta}_{1,R}] = \alpha, \quad (9.27)$$

where $\hat{\phi}^2_{x,R}$ and $\hat{\delta}^2_{x,R}$ are the robust estimators of ϕ^2 and δ^2 given by

$$\hat{\phi}^2_{x,R} = \left(\frac{A_2(x)-3}{A_2(x)-1}\right)\frac{\hat{\sigma}_1^2(x)}{\hat{\sigma}_2^2(x)} \tag{9.28}$$

and

$$\hat{\delta}^2_{x,R} = \left(\frac{A_2(x)-3}{A_2(x)-1}\right)\frac{(\hat{\mu}_1(x)-\hat{\mu}_2(x))^2}{\hat{\sigma}_2^2(x)}. \tag{9.29}$$

Theorem 17 *Asymptotically, $\Lambda(\mathrm{w})$ and $\Lambda_R(\mathrm{w})$ procedures have exactly the same e_{12} and $1-e_{21}$ values, under normality for the population distributions*

Proof: Since the MML estimators $\hat{\mu}(x)$ and $\hat{\sigma}^2(x)$ have asymptotically exactly the same distributions as $\bar{y}(x)$ and $s^2(x)$, viz, $\sqrt{m(x)}\frac{\hat{\mu}(x)-\mu(x)}{\sigma(x)}$ and $(A(x)-1)\frac{\hat{\sigma}^2(x)}{\sigma^2(x)}$ are independently distributed as standard normal and chi-square with $A(x)-1$ degrees of freedom (see Tiku, Tan and Balakrishnan (1986)), respectively, the distribution of $\Lambda_R(\mathrm{w})$ in (9.25) is the same as that of $\Lambda(\mathrm{w})$ (for large $n_1(x)$ and $n_2(x)$) under the assumption of normality. Therefore, for large $n_1(x)$ and $n_2(x)$, the classical and the robust quadratic classification procedures in (9.22) and (9.25), respectively, will both have the same $1-e_{21}$ value for a fixed e_{12}, under the assumption of normality.

9.5 Transformed Linear Classification Procedures

In addition to the fact that the classical and the robust quadratic classification procedures developed in that last section have a complicated asymptotic theory, simulated results indicate that these two procedures also have their actual e_{12} values to be somewhat larger than the pre-fixed level of e_{12}; see, for example, Tables 9.1 and 9.2. Therefore, in this section we develop transformed linear classification procedures which, firstly have simple asymptotic theory and secondly have their actual e_{12} values to be close to the pre-fixed level of e_{12} (see, for example, Tables 9.3 and 9.4). This development is similar to the one used by Tiku and Balakrishnan (1985, 1989) and hence can be considered to be a generalization of their method

to the case when the data consists of a dichotomous variable and an associated continuous variable.

The classical transformed linear classification procedure is based on the statistic

$$\Lambda^*(\mathrm{w}) = \left\{ \frac{y}{s_1(x)} - \frac{1}{2}\left(\frac{\bar{y}_1(x)}{s_1(x)} + \frac{\bar{y}_2(x)}{s_2(x)}\right)\right\}\left(\frac{\bar{y}_1(x)}{s_1(x)} - \frac{\bar{y}_2(x)}{s_2(x)}\right) + \lambda(x, \hat{\theta}_1, \hat{\theta}_2). \quad (9.30)$$

where $\lambda(x, \theta_1, \theta_2) = x \log\left[\frac{\theta_1}{\theta_2}\right] + (1-x)\log\left[\frac{1-\theta_1}{1-\theta_2}\right]$ as defined earlier in Section 3.

The classical transformed linear classification procedure is based on the statistic $\Lambda^*(\mathrm{w})$ in (9.30) and is to classify $\mathrm{w}_0 = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ into Π_1 or Π_2 according as

$$\Lambda^*(\mathrm{w}_0) \gtrless \hat{C}^*, \quad (9.31)$$

where the cut-off point $\hat{C}^*$ is derived by fixing the error of misclassification e_{12} at a specified level and using the following asymptotic argument. Since the distributional properties of $\Lambda^*(\mathrm{w})$ in (9.35) is unknown it is difficult to determine the cut-off point $\hat{C}^*$ in (9.31) exactly for a specified e_{12}. Asymptotically (as $n_1(x)$, $n_2(x) \to \infty$), $\hat{\theta}_1$, $\hat{\theta}_2$, $\bar{y}_1(x)$, $\bar{y}_2(x)$, $s_1(x)$ and $s_2(x)$ tend to θ_1, θ_2, $\mu_1(x)$, $\mu_2(x)$, $\sigma_1(x)$ and $\sigma_2(x)$, respectively, and hence the distribution of $\Lambda^*(\mathrm{w})$ in (9.30) is exactly the same as the distribution of $V(\mathrm{w})$, where

$$V(\mathrm{w}) = \left\{ \frac{y}{\sigma_1(x)} - \frac{1}{2}\left(\frac{\mu_1(x)}{\sigma_1(x)} + \frac{\mu_2(x)}{\sigma_2(x)}\right)\right\}\left(\frac{\mu_1(x)}{\sigma_1(x)} - \frac{\mu_2(x)}{\sigma_2(x)}\right) + \lambda(x, \theta_1, \theta_2). \quad (9.32)$$

Since $V(\mathrm{w})$ is a linear function of a normal variable y, we find immediately that

$$E(V(\mathrm{w})|\mathrm{w} \in \Pi_1) = \frac{1}{2}\left(\frac{\mu_1(x)}{\sigma_1(x)} - \frac{\mu_2(x)}{\sigma_2(x)}\right)^2 + \lambda(x, \theta_1, \theta_2),$$

and

$$Var(V(\mathrm{w})|\mathrm{w} \in \Pi_1) = \left(\frac{\mu_1(x)}{\sigma_1(x)} - \frac{\mu_2(x)}{\sigma_2(x)}\right)^2.$$

The cutoff point C^* is then a solution of the equation (with e_{12} pre- fixed as α)

$$\begin{aligned} \alpha &= \Pr[V(\mathrm{w}_0) < C^* | \mathrm{w}_0 \in \Pi_1] \\ &= (1-\theta_1)\Pr\left[V\left(\begin{bmatrix} 0 \\ y_0 \end{bmatrix}\right) < C^* | \mathrm{w}_0 \in \Pi_1\right] + \theta_1 \Pr\left[V\left(\begin{bmatrix} 1 \\ y_0 \end{bmatrix}\right) < C^* | \mathrm{w}_0 \in \Pi_1\right]; \end{aligned}$$

that is, the cut-off point C^* may be obtained as a solution of the equation

$$(1-\theta_1)\Phi\left\{\frac{C^* - \frac{1}{2}\eta^2(0) - \ln\left(\frac{1-\theta_1}{1-\theta_2}\right)}{|\eta(0)|}\right\} + \theta_1\Phi\left\{\frac{C^* - \frac{1}{2}\eta^2(1) - \ln\left(\frac{\theta_1}{\theta_2}\right)}{|\eta(1)|}\right\} = \alpha,$$

where $\eta^2(x) = \left\{\frac{\mu_1(x)}{\sigma_2(x)} - \frac{\mu_2(x)}{\sigma_2(x)}\right\}^2$ for $x = 0, 1$. Since $\mu_i(x)$, $\sigma_i(x)$ are all unknown, we will replace $\mu_i(x)$ by $\bar{y}_i(x)$ and $\frac{1}{\sigma_i(x)}$ by $\frac{a_i(x)}{s_i(x)}$, where $a_i(x) = \frac{\Gamma(\frac{n_i(x)-1}{2})}{(\frac{n_i(x)-1}{2})^{\frac{1}{2}}\Gamma(\frac{n_i(x)-2}{2})}$. Note that $E(\frac{a_i(x)}{s_i(x)}|X = x) = \frac{1}{\sigma_i(x)}$ (This follows from the fact that if Z is distributed as chi-squared with $n-1$ degrees of freedom, then $E(Z^{-1/2}) = \frac{(\frac{1}{2})^{1/2}\Gamma(\frac{n-2}{2})}{\Gamma(\frac{n-1}{2})}$). Therefore, the cut-off point $\hat{C}^*$ of the classical transformed linear classification procedure in (9.31) may be determined approximately as a solution of the equation (with e_{12} fixed as α)

$$(1-\hat{\theta}_1)\Phi\left\{\frac{\hat{C}^* - \frac{1}{2}\hat{\eta}^2(0) - \ln\left(\frac{1-\hat{\theta}_1}{1-\hat{\theta}_2}\right)}{|\hat{\eta}(0)|}\right\} + \theta_1\Phi\left\{\frac{\hat{C}^* - \frac{1}{2}\hat{\eta}^2(1) - \ln\left(\frac{\hat{\theta}_1}{\hat{\theta}_2}\right)}{|\hat{\eta}(1)|}\right\} = \alpha, \quad (9.33)$$

where $\hat{\eta}^2(x) = \left\{a_1(x)\frac{\bar{y}_1(x)}{s_1(x)} - a_2(x)\frac{\bar{y}_2(x)}{s_2(x)}\right\}^2$ for $x = 0, 1$.

Similarly, by using the MML estimators instead of the classical ML estimators in place of the parameters, we obtain the robust transformed linear classification procedure to be one that is based on the statistic

$$\Lambda_R^*(\mathrm{w}) = \left\{\frac{y}{\hat{\sigma}_1(x)} - \frac{1}{2}\left(\frac{\hat{\mu}_1(x)}{\hat{\sigma}_1(x)} + \frac{\hat{\mu}_2(x)}{\hat{\sigma}_2(x)}\right)\right\}\left(\frac{\hat{\mu}_1(x)}{\hat{\sigma}_1(x)} - \frac{\hat{\mu}_2(x)}{\hat{\sigma}_2(x)}\right) + \lambda(x, \hat{\theta}_1, \hat{\theta}_2). \quad (9.34)$$

The robust transformed linear classification procedure is then based on the statistic $\Lambda_R^*(\mathrm{w})$ in (9.34) and classifies an observation $\mathrm{w}_0 = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ into Π_1 or Π_2 according as

$$\Lambda_R^*(\mathrm{w}_0) \gtrless \hat{C}_R^*, \quad (9.35)$$

where the cut-off point $\hat{C}_R^*$, with e_{12} pre-fixed as α, can be determined approximately as a solution of the equation

$$(1-\hat{\theta}_1)\Phi\left\{\frac{\hat{C}_R^* - \frac{1}{2}\eta_R^2(0) - \ln\left(\frac{1-\hat{\theta}_1}{1-\hat{\theta}_2}\right)}{|\eta_R(0)|}\right\} + \hat{\theta}_1\Phi\left\{\frac{\hat{C}_R^* - \frac{1}{2}\eta_R^2(1) - \ln\left(\frac{\hat{\theta}_1}{\hat{\theta}_2}\right)}{|\eta_R(1)|}\right\} = \alpha, \quad (9.36)$$

where $\hat{\eta}_R^2(x) = \left\{b_1 \frac{\hat{\mu}_1(x)}{\hat{\sigma}_2(x)} - b_2 \frac{\hat{\mu}_2(x)}{\hat{\sigma}_2(x)}\right\}^2$ for $x = 0, 1$ where $b_i(x) = \frac{\Gamma(\frac{A_i(x)-1}{2})}{(\frac{A_i(x)-1}{2})^{\frac{1}{2}}\Gamma(\frac{A_i(x)-2}{2})}$. Note that $E(\frac{a_i(x)}{\hat{\sigma}_i(x)}|X = x) = \frac{1}{\sigma_i(x)}$ asymptotically.

Theorem 18 *The statistic $\Lambda_R^*(\mathrm{w}_0)$ in (9.34) asymptotically has exactly the same distribution as the statistic $\Lambda^*(\mathrm{w}_0)$ in (9.30).*

Proof: For large n_1 and n_2 the distributional properties of MML estimators $\hat{\mu}$ and $\hat{\sigma}$ are similar to those of the ML estimators $\bar{x}$ and s. Therefore, when n_1 and n_2 tend to infinity, the statistics $\Lambda_R^*(\mathrm{w}_0)$ in (9.34) and $\Lambda^*(\mathrm{w}_0)$ in (9.30) both have exactly the same distribution as the statistics $V(\mathrm{w}_0)$ in (9.32). Hence the proof. ∎

Theorem 19 *Asymptotically, $\Lambda^*(\mathrm{w}_0)$ and $\Lambda_R(\mathrm{w}_0)$ procedures have exactly the same e_{12} and $1 - e_{21}$ values, under normality for the population distributions.*

Proof: This theorem follows immediately from **Theorem 13** upon noting that the cut-off points $\hat{C}^*$ and $\hat{C}_R^*$ obtained from Eqs (9.33) and (9.36) both converge to the same point C^*. ∎

9.6 Comparison of the Four Classification Procedures

In order to examine the efficiency and the robustness features of the four classification procedures described in Section 4 and 5, $n_1 = n_2 = n$ pseudo-random observations were simulated for the y_1 and y_2 samples each from each of the following

seven models:

(1) Normal $N(0,1)$

(2) $y_1, y_2, \ldots, y_{n-1}$ come from $N(0,1)$ and y_n from $N(0,100)$

(3) $y_1, y_2, \ldots, y_{n-2}$ come from $N(0,1)$ and y_{n-1} and y_n from $N(0,16)$

(4) Logistic: $\frac{\exp -y}{(1+\exp -y)^2}, \quad -\infty < y < \infty,$

(5) Double exponential: $\frac{1}{2}\exp -|y|, \quad -\infty < y < \infty,$

(6) Mixture model: $0.90N(0,1) + 0.10N(0,16)$

and

(7) Student's t_4 : $\frac{3}{8(1+y^2/4)^{5/2}}, \quad -\infty < y < \infty$

(9.37)

After multiplying the y_i sample by σ_i $(i = 1,2)$, a constant μ_i $(i = 1,2)$ was added to the y_i sample in order to create a scale and location shift between the two populations Π_1 and Π_2. Then, with e_{12} fixed as 0.05, the values of e_{12} and $1-e_{21}$ were simulated (based on 1,000 Monte Carlo runs) for all the four classification procedures under all seven models in (9.37) with $\mu_1 = 0$, $\sigma_1 = 1$ $\mu_2 = 2, 2.25, 2.5$, $\sigma_2 = 2, 3$ and $\delta = \gamma = 1$, $\theta_1 = \theta_2 = 0.25$, $\theta_1 = \theta_2 = 0.5$ and $n_1 = n_2 = 50$. These values are presented in Tables 9.1-9.4.

From Tables 9.1 and 9.2, it is clear that the robust quadratic classification procedure in (9.25) has its $1 - e_{21}$ value (the probability of correctly classifying an individual from population Π_2 into itself) to be very close to that of the classical quadratic classification procedure in (9.22) under normality of the population distributions; at the same time, the robust quadratic procedure has its $1 - e_{21}$ value to be considerably larger than that of the classical quadratic procedure for all the non-normal models. However, the e_{12} values of both these procedures are somewhat larger than the pre-fixed value of e_{12}.

From Tables 9.3 and 9.4, we observe that the robust transformed linear classification procedure in (9.35) has its $1 - e_{21}$ value to be very close to that of the classical transformed linear classification procedure in (9.31) under normality of the population distributions; at the same time, the robust transformed linear procedure has its $1 - e_{21}$ value to be much larger than that of the classical transformed linear

procedure for all the non- normal models considered in this study. Also, we note from Tables 9.3 and 9.4 that both these transformed linear procedures have their e_{12} values to be quite close to the pre-fixed level of e_{12}. Furthermore, upon comparing the $1-e_{21}$ values in Tables 9.3 and 9.4 to the corresponding values in Tables 9.1 and 9.2, we observe that the transformed linear classification procedure described in Section 5 are only slightly less powerful than the corresponding quadratic classification procedures described in Section 4 (less powerful in the sense that they have only slightly smaller $1-e_{21}$ values). So, the numerical results presented in this section suggest that the robust transformed linear classification procedure in (9.35) is the best of these four procedures since it has a simple asymptotic distribution theory, has its e_{12} values to be quite close to the pre-fixed value of e_{12}, has its $1-e_{21}$ value (the probability of correctly classifying an individual from population Π_2 into itself) to be considerably larger than that of the classical transformed linear classification procedure and only slightly smaller than that of the robust quadratic classification procedure, and also remains quite robust to departures from normality.

Table 9.1: Simulated values of e_{12} and $1 - e_{21}$: Case I
Simulated values of e_{12} and $1 - e_{21}$ for the classical and the robust quadratic classification procedures for the models in (9.37) when $n_1 = n_2 = 50$, $\theta_1 = \theta_2 = 0.5$, $\delta = 1.0$ and $\gamma = 1.0$, with e_{12} pre-fixed as 0.05.

M	P	e_{12} values						$1 - e_{21}$ values					
		$\sigma_2 = 2$			$\sigma_2 = 3$			$\sigma_2 = 2$			$\sigma_2 = 3$		
		a	b	c	a	b	c	a	b	c	a	b	c
(1)	Λ	.072	.085	.078	.060	.072	.092	.551	.615	.662	.564	.590	.663
	Λ_R	.072	.091	.079	.063	.081	.095	.547	.616	.660	.558	.600	.652
(2)	Λ	.035	.045	.038	.032	.039	.050	.345	.381	.434	.358	.421	.457
	Λ_R	.065	.083	.076	.059	.077	.088	.538	.611	.652	.547	.592	.637
(3)	Λ	.045	.058	.049	.043	.042	.055	.419	.497	.544	.462	.505	.564
	Λ_R	.061	.083	.070	.056	.072	.084	.521	.600	.645	.539	.582	.633
(4)	Λ	.069	.084	.072	.056	.068	.084	.552	.617	.674	.579	.622	.674
	Λ_R	.082	.094	.084	.072	.086	.102	.575	.648	.710	.591	.639	.693
(5)	Λ	.085	.075	.066	.078	.089	.063	.368	.439	.503	.393	.462	.512
	Λ_R	.099	.095	.089	.099	.117	.084	.438	.511	.583	.467	.546	.583
(6)	Λ	.068	.061	.067	.072	.070	.053	.332	.381	.407	.375	.405	.444
	Λ_R	.102	.100	.091	.112	.098	.087	.476	.555	.602	.500	.564	.596
(7)	Λ	.067	.056	.058	.076	.072	.067	.571	.618	.698	.564	.646	.663
	Λ_R	.087	.072	.080	.106	.097	.089	.642	.700	.764	.631	.726	.726

1. M stands for Model

2. P stands for Procedure

3. $a \Rightarrow \mu_2 = 2.0, \quad b \Rightarrow \mu_2 = 2.25, \quad c \Rightarrow \mu_2 = 2.50$

Table 9.2: Simulated values of e_{12} and $1 - e_{21}$: Case II
Simulated values of e_{12} and $1 - e_{21}$ for the classical and the robust quadratic classification procedures for the models in (9.37) when $n_1 = n_2 = 50$, $\theta_1 = \theta_2 = 0.25$ $\delta = 1.0$ and $\gamma = 1.0$, with e_{12} pre-fixed as 0.05

M	P	e_{12} values						$1 - e_{21}$ values					
		$\sigma_2 = 2$			$\sigma_2 = 3$			$\sigma_2 = 2$			$\sigma_2 = 3$		
		a	b	c	a	b	c	a	b	c	a	b	c
(1)	Λ	.081	.071	.063	.082	.083	.082	.586	.679	.706	.572	.629	.701
	Λ_R	.087	.072	.073	.091	.082	.083	.585	.680	.705	.568	.629	.707
(2)	Λ	.041	.038	.033	.056	.041	.050	.399	.474	.487	.395	.445	.489
	Λ_R	.081	.071	.069	.088	.078	.079	.567	.672	.692	.558	.626	.687
(3)	Λ	.049	.046	.037	.061	.051	.059	.492	.582	.604	.488	.540	.604
	Λ_R	.081	.066	.065	.086	.076	.076	.561	.664	.685	.552	.615	.677
(4)	Λ	.081	.074	.066	.085	.080	.078	.616	.692	.718	.584	.637	.720
	Λ_R	.089	.080	.077	.102	.088	.095	.640	.709	.745	.597	.664	.639
(5)	Λ	.073	.087	.084	.079	.084	.076	.404	.508	.533	.460	.528	.581
	Λ_R	.110	.116	.103	.112	.104	.103	.489	.591	.622	.528	.591	.642
(6)	Λ	.074	.071	.055	.065	.065	.066	.369	.436	.498	.400	.441	.538
	Λ_R	.104	.101	.090	.099	.099	.099	.526	.575	.662	.569	.582	.656
(7)	Λ	.074	.059	.065	.078	.069	.072	.620	.704	.723	.642	.664	.732
	Λ_R	.101	.077	.094	.103	.097	.093	.692	.778	.799	.699	.726	.780

1. M stands for Model
2. P stands for Procedure
3. $a \Rightarrow \mu_2 = 2.0$, $b \Rightarrow \mu_2 = 2.25$, $c \Rightarrow \mu_2 = 2.50$

Table 9.3: Simulated values of e_{12} and $1 - e_{21}$: Case III
Simulated values of e_{12} and $1 - e_{21}$ for the classical and the robust transformed linear classification procedures for the models (9.37) when $n_1 = n_2 = 50$, $\theta_1 = \theta_2 = 0.5$ $\delta = 1.0$ and $\gamma = 1.0$, and e_{12} pre-fixed as 0.05

M	P	e_{12} values						$1 - e_{21}$ values					
		$\sigma_2 = 2$			$\sigma_2 = 3$			$\sigma_2 = 2$			$\sigma_2 = 3$		
		a	b	c	a	b	c	a	b	c	a	b	c
(1)	Λ^*	.054	.071	.055	.060	.061	.072	.503	.576	.635	.510	.565	.616
	Λ_R^*	.059	.073	.058	.067	.060	.078	.512	.576	.633	.516	.563	.618
(2)	Λ^*	.047	.055	.049	.067	.051	.061	.381	.411	.456	.362	.434	.463
	Λ_R^*	.052	.072	.052	.067	.057	.078	.496	.564	.627	.503	.549	.604
(3)	Λ^*	.044	.054	.040	.064	.044	.056	.396	.464	.524	.427	.469	.531
	Λ_R^*	.051	.070	.049	.062	.056	.078	.488	.556	.615	.493	.543	.594
(4)	Λ^*	.054	.072	.054	.057	.063	.074	.503	.584	.651	.513	.566	.630
	Λ_R^*	.067	.085	.064	.073	.073	.087	.534	.609	.678	.539	.596	.648
(5)	Λ^*	.073	.072	.059	.080	.078	.059	.360	.394	.466	.354	.423	.443
	Λ_R^*	.087	.079	.077	.086	.097	.075	.412	.473	.548	.396	.505	.529
(6)	Λ^*	.071	.079	.060	.081	.071	.057	.314	.394	.414	.352	.382	.412
	Λ_R^*	.085	.091	.072	.094	.077	.069	.445	.517	.559	.447	.497	.533
(7)	Λ^*	.055	.040	.058	.068	.060	.053	.537	.527	.672	.500	.606	.602
	Λ_R^*	.071	.062	.074	.089	.074	.072	.603	.656	.740	.563	.679	.685

1. M stands for Model
2. P stands for Procedure
3. $a \Rightarrow \mu_2 = 2.0$, $b \Rightarrow \mu_2 = 2.25$, $c \Rightarrow \mu_2 = 2.50$

Table 9.4: Simulated values of e_{12} and $1 - e_{21}$: Case IV
Simulated values of e_{12} and $1 - e_{21}$ for the classical and the robust transformed linear classification procedures for the models (9.37) when $n_1 = n_2 = 50$, $\theta_1 = \theta_2 = 0.25$ $\delta = 1.0$ and $\gamma = 1.0$, and e_{12} pre-fixed as 0.05

M	P	e_{12} values						$1 - e_{21}$ values					
		$\sigma_2 = 2$			$\sigma_2 = 3$			$\sigma_2 = 2$			$\sigma_2 = 3$		
		a	b	c	a	b	c	a	b	c	a	b	c
(1)	Λ^*	.062	.060	.052	.065	.062	.063	.535	.641	.668	.518	.579	.648
	Λ^*_R	.065	.059	.057	.071	.066	.069	.539	.639	.667	.512	.577	.647
(2)	Λ^*	.041	.038	.033	.056	.041	.050	.407	.484	.489	.373	.413	.462
	Λ^*_R	.063	.056	.055	.064	.064	.066	.523	.634	.658	.508	.564	.631
(3)	Λ^*	.040	.041	.037	.047	.040	.040	.465	.552	.564	.450	.496	.553
	Λ^*_R	.058	.051	.052	.060	.060	.064	.512	.623	.650	.506	.564	.619
(4)	Λ^*	.060	.057	.052	.065	.063	.060	.555	.652	.679	.522	.587	.665
	Λ^*_R	.066	.067	.061	.075	.071	.072	.591	.679	.708	.540	.611	.697
(5)	Λ^*	.052	.069	.071	.060	.062	.051	.364	.466	.496	.396	.455	.529
	Λ^*_R	.083	.096	.091	.076	.082	.068	.462	.552	.580	.473	.523	.594
(6)	Λ^*	.061	.061	.046	.058	.046	.052	.358	.413	.470	.366	.404	.490
	Λ^*_R	.083	.080	.071	.072	.067	.073	.488	.530	.618	.493	.529	.595
(7)	Λ^*	.064	.052	.056	.062	.055	.055	.593	.664	.676	.582	.608	.664
	Λ^*_R	.085	.067	.081	.082	.074	.071	.656	.730	.744	.651	.656	.728

1. M stands for Model
2. P stands for Procedure
3. $a \Rightarrow \mu_2 = 2.0$, $b \Rightarrow \mu_2 = 2.25$, $c \Rightarrow \mu_2 = 2.50$

Chapter 10

MML Estimation for Simple Linear Regression Model

10.1 Introduction

So far in this thesis, the discussion has been on the robustness of classical classification procedures and also in proposing some robust classification procedures. From now on, our discussion will be regarding the robustness of the classical least-squares method of estimation of parameters in a linear regression model and also to propose a robust method of estimation based on the MML estimators.

Regression analysis is one of the most widely used statistical techniques for analyzing a multifactor data. The classical least-squares variable in a linear regression model is normally distributed, the least-squares estimators of the parameters possess several optimal properties; see, for example, Myers (1986). However, when the random error component in the linear regression model has a non-normal distribution, particularly one that has longer or heavier tails than the normal, the classical method of estimation of parameters may not be appropriate. Heavy-tailed distributions usually generate outliers and these outliers may have a very strong influence on the least-squares estimators. In general, the effect of such outliers in the data is to "pull" the least-squares fit too much in their direction and, as a result, the identification of these outliers becomes difficult since their associated residuals

have been made artificially small.

A number of authors have proposed robust and non-parametric regression procedures which usually (quite successfully in many instances) dampen the effect of those observations which would be highly influential if the least-squares method of estimation was employed. That is, a robust procedure tends to leave the residuals associated with the outlying observations large thereby making the identification of such influential points much easier. Adichie (1976a,b) proposed a non-parametric estimation method which is implicit in nature and is based on weighted median of the set of all pairwise slopes. Andrews (1973,1974) developed M-estimation methods based on two ψ functions discussed in Andrews et al. (1972) and Huber (1981). Yale and Forsythe (1976) introduced a Winsorized estimation method which has been modified by Tan and Tabatabai (1988). It is important to mention here that Leone and Moussa-Hamouda (1973), Moussa-Hamouda and Leone (1974, 1977a, 1977b, 1977c), Moussa-Hamouda (1988) and Tiku (1981) have studied method of estimation based on censored samples. But, all the results mentioned above are for the case when repeated measurements are available at each level of the explanatory variable. For more details on these results, one may refer to Balakrishnan and Cohen (1990).

In this chapter we directly adopt the modified maximum likelihood approach of Tiku (1967) for the likelihood function based on the Type II symmetrically censored sample of normalized residuals and derive the MMLEs of the parameters in a simple linear regression model. In Section 2 we derive the MML estimators for the parameters in a simple linear regression model based on symmetrically censored samples. In Section 3 we explain the method of determination of the observations to be censored for this estimation process. In Section 4 we derive approximate expressions for the variances and covariances of these estimators via the information matrix and establish the asymptotic distribution of these estimators. In Section 5 we study the MMLE's under the normal and the mixture-normal models for the error component and compare them with the least square estimators. In Section 6 we compare MMLE procedure with some other prominent robust estimation procedures for departures from normality and also for departures from linearity. Finally,

we consider two examples given by Daniel and Wood (1971) and Mendenhall (1983) and illustrate all the methods of estimation mentioned above.

The estimators presented in this chapter are similar to those given by Tan and Tabatabai (1988). But, the approach presented here identifies the residuals to be censored in a different fashion that conforms with the literature on outlier-detection. This method also seems to be better than the method proposed by Tan and Tabatabai (1988). Our approach also leads to asymptotic variances and covariances of the estimators which could be used to determine the estimated standard errors of the estimates. Furthermore, in this chapter we examine the robustness features of various estimation methods under departures from normality as well as from linearity.

10.2 MML estimation Method

Consider a simple linear regression model

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{10.1}$$

where α and β are the usual regression parameters, $(x_1, x_2, \ldots, x_n)$ are the values of the explanatory variable, and e_i is the random error involved in measuring y_i. We shall assume that $e_i's$ are independently distributed with mean zero and unknown variance σ^2. Write

$$Z_i = (y_i - \alpha - \beta x_i)/\sigma, \quad (i = 1, 2, \ldots, n) \tag{10.2}$$

and denote their order statistics by

$$Z_{(1)} \leq Z_{(2)} \leq Z_{(2)} \leq \ldots \leq Z_{(n)} \tag{10.3}$$

Further, let us denote the (x, y) pair corresponding to the $i'th$ order statistic $Z_{(i)}$ by $(x_{[i]}, y_{[i]})$ for $i = 1, 2, \ldots, n$. Let us consider the Type-II symmetrically censored sample

$$Z_{(r+1)} \leq Z_{(r+2)} \leq \ldots \leq Z_{(n-r)} \tag{10.4}$$

where r is usually chosen to be $[0.5 + 0.1n]$($[g]$ indicating the integer part of g) to achieve robustness to most departures from normality (Tiku, Tan and Balakrishnan, 1986). Under the assumption of normality, the likelihood function based on the symmetrically censored sample (10.4) is

$$L = \frac{n!}{r!^2}\sigma^{-(n-2r)}\left\{\Phi(Z_{(r+1)})\right\}^r\left\{1-\Phi(Z_{(n-r)})\right\}^r\prod_{i=r+1}^{n-r}\phi(z_{(i)}) \tag{10.5}$$

where $\phi(z)$ and $\Phi(z)$ are the density function and cumulative distribution function of a standard normal variable, respectively. Thus

$$\ln L = Const - A\ln\sigma + r\ln\Phi(Z_{(r+1)}) + r\ln\left\{1-\Phi(Z_{(n-r)})\right\} - \frac{1}{2}\sum_{i=r+1}^{n-r}Z_{(i)}^2 \tag{10.6}$$

where $A = n - 2r$. Upon differentiating Eq. (10.6) with respect to α, β and σ and using the following fact

$$\frac{\partial Z_{(i)}}{\partial\alpha} = -\frac{1}{\sigma};\quad \frac{\partial Z_{(i)}}{\partial\beta} = -\frac{x_{[i]}}{\sigma};\quad \frac{\partial Z_{(i)}}{\partial\sigma} = -\frac{Z_{(i)}}{\sigma};$$

we obtain the following likelihood equations:

$$\begin{aligned}\frac{\partial\ln L}{\partial\alpha} &= \frac{1}{\sigma}\left[-r\frac{\phi(Z_{(r+1)})}{\Phi(Z_{(r+1)})} + r\frac{\phi(Z_{(n-r)})}{1-\Phi(Z_{(n-r)})} + \sum_{i=r+1}^{n-r}Z_{(i)}\right]\\ &= 0\end{aligned} \tag{10.7}$$

$$\begin{aligned}\frac{\partial\ln L}{\partial\beta} &= \frac{1}{\sigma}\left[-rx_{[r+1]}\frac{\phi(Z_{(r+1)})}{\Phi(Z_{(r+1)})} + rx_{[r+1]}\frac{\phi(Z_{(n-r)})}{1-\Phi(Z_{(n-r)})} + \sum_{i=r+1}^{n-r}x_{[i]}Z_{(i)}\right]\\ &= 0\end{aligned} \tag{10.8}$$

and

$$\begin{aligned}\frac{\partial\ln L}{\partial\sigma} &= \frac{1}{\sigma}\left[-rZ_{[r+1]}\frac{\phi(Z_{(r+1)})}{\Phi(Z_{(r+1)})} + rZ_{(r+1)}\frac{\phi(Z_{(n-r)})}{1-\Phi(Z_{(n-r)})} + \sum_{i=r+1}^{n-r}Z_{(i)}^2\right]\\ &= 0\end{aligned} \tag{10.9}$$

Eqs. (10.7)-(10.9) do not provide explicit estimators for α, β and σ. However, following the steps of Tiku (1967), we approximate the functions $\frac{\phi(Z_{(r+1)})}{\Phi(Z_{(r+1)})}$ and $\frac{\phi(Z_{(n-r)})}{1-\Phi(Z_{(n-r)})}$:

$$\frac{\phi(Z_{(r+1)})}{\Phi(Z_{(r+1)})} \approx \gamma - \delta Z_{(r+1)} \tag{10.10}$$

and

$$\frac{\phi(Z_{(n-r)})}{1-\Phi(Z_{(n-r)})} \approx \gamma + \delta Z_{(n-r)} \tag{10.11}$$

where γ and δ have been tabulated by Tiku, Tan and Balakrishnan (1986). Tiku (1967) has shown that (10.10) and (10.11) provide very close approximations. In fact, (10.10) and(10.11) are strict equalities for large n (Tiku,1970). Incorporating the linear approximations (10.10) and (10.11) in (10.7) - (10.9), we obtain the following MML equations:

$$\begin{aligned}\frac{\partial \ln L}{\partial \alpha} \approx \frac{\partial \ln L^*}{\partial \alpha} &= \frac{1}{\sigma}\left[-r\left\{\gamma - \delta Z_{(r+1)}\right\} + r\left\{\gamma + \delta Z_{(n-r)}\right\} + \sum_{i=r+1}^{n-r} Z_{(i)}\right] \\ &= 0 \end{aligned} \tag{10.12}$$

$$\begin{aligned}\frac{\partial \ln L}{\partial \beta} \approx \frac{\partial \ln L^*}{\partial \beta} &= \frac{1}{\sigma}\left[-rx_{[r+1]}\left\{\gamma - \delta Z_{(r+1)}\right\} + rx_{[n-r]}\left\{\gamma + \delta Z_{(n-r)}\right\} + \sum_{i=r+1}^{n-r} x_{[i]}Z_{(i)}\right] \\ &= 0 \end{aligned} \tag{10.13}$$

and

$$\begin{aligned}\frac{\partial \ln L}{\partial \sigma} \approx \frac{\partial \ln L^*}{\partial \sigma} &= \frac{1}{\sigma}\left[-A - rZ_{(r+1)}\left\{\gamma - \delta Z_{(r+1)}\right\} + rZ_{(n-r)}\left\{\gamma + \delta Z_{(n-r)}\right\} + \sum_{i=r+1}^{n-r} Z_{(i)}^2\right] \\ &= 0 \end{aligned} \tag{10.14}$$

where $A = n - 2r$. Simplification of first two equations as follows:

$$\begin{aligned}\frac{\partial \ln L^*}{\partial \alpha} &= \frac{1}{\sigma}\left[\sum_{i=r+1}^{n-r} w_i Z_{(i)}\right] \qquad (10.15) \\ &= \frac{1}{\sigma^2}\left[m\bar{y} - m\alpha - m\beta\bar{x}\right] \\ &= \frac{m}{\sigma^2}\left[\bar{y} - \alpha - \beta\bar{x}\right] \\ &= 0 \qquad (10.16)\end{aligned}$$

$$\begin{aligned}\frac{\partial \ln L^*}{\partial \beta} &= \frac{1}{\sigma}\left[-r\gamma(x_{[n-r]} - x_{[r+1]}) + \sum_{i=r+1}^{n-r} w_i x_{[i]} Z_{(i)}\right] \\ &= \frac{1}{\sigma^2}\left[-r\gamma\sigma(x_{[n-r]} - x_{[r+1]}) + \sum_{i=r+1}^{n-r} w_i x_{[i]}(y_{(i)} - \alpha - \beta x_{[i]})\right] \\ &= \frac{1}{\sigma^2}\left[-r\gamma\sigma(x_{[n-r]} - x_{[r+1]}) + S^*_{x,y} + m\bar{x}\bar{y} - \alpha m\bar{x} - \beta[S^*_{x,x} + m\bar{x}^2]\right] \\ &= \frac{1}{\sigma^2}\left[-r\gamma\sigma(x_{[n-r]} - x_{[r+1]}) + S^*_{x,y} - \beta S^*_{x,x} + m\bar{x}(\bar{y} - \alpha - \beta\bar{x})\right] \\ &= 0, \qquad (10.17)\end{aligned}$$

where

$$m = n - 2r + 2r\delta \tag{10.18}$$

$$w_i = \begin{cases} 1 + r\delta & \text{for } i = r+1 \\ 1 & \text{for } r+1 < i < n-r \\ 1 + r\delta & \text{for } i = n-r \end{cases} \tag{10.19}$$

$$\bar{x} = \frac{1}{m}\sum_{i=r+1}^{n-r} w_i x_{[i]} \tag{10.20}$$

and

$$\bar{y} = \frac{1}{m}\sum_{i=r+1}^{n-r} w_i y_{[i]} \tag{10.21}$$

$$S^*_{x,x} = \sum_{i=r+1}^{n-r} w_i x_{[i]}^2 - m\bar{x}^2 \tag{10.22}$$

and

$$S^*_{x,y} = \sum_{i=r+1}^{n-r} w_i x_{[i]} y_{[i]} - m\bar{x}\bar{y}$$

$$S^*_{y,y} = \sum_{i=r+1}^{n-r} w_i y_{[i]}^2 - m\bar{y}^2 \tag{10.23}$$

From Eq. (10.16) we derive the MML estimator of α as

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{10.24}$$

Next, from Eq. (10.17) we derive the MML estimator of β as

$$\hat{\beta} = \frac{S^*_{x,y}}{S^*_{x,x}} + \hat{\sigma}\left\{\frac{r\gamma(x_{[n-r]} - x_{[r+1]})}{S^*_{x,x}}\right\}. \tag{10.25}$$

We can simplify the Eq. (10.14) as follows:

$$\begin{aligned} \frac{\partial \ln L^*}{\partial \sigma} &= \frac{1}{\sigma}\left[-A - rZ_{(r+1)}\left\{\gamma - \delta Z_{(r+1)}\right\} + rZ_{(n-r)}\left\{\gamma + \delta Z_{(n-r)}\right\} + \sum_{i=r+1}^{n-r} Z_{(i)}^2\right] \\ &= \frac{1}{\sigma}\left[-A + r\gamma(Z_{(n-r)} - Z_{(r+1)}) + \sum_{i=r+1}^{n-r} w_i Z_{(i)}^2\right] \\ &= \frac{1}{\sigma^3}\left[-A\sigma^2 + r\gamma\sigma\left(y_{[n-r]} - y_{[r+1]} - \beta(x_{[n-r]} - x_{[r+1]})\right) + \right. \\ &\quad \left. \sum_{i=r+1}^{n-r} w_i(y_{[i]} - \alpha - \beta x_{[i]})^2\right], \end{aligned} \tag{10.26}$$

by substituting the estimators of α and β in

$$\sum_{i=r+1}^{n-r} w_i(y_{(i)} - \alpha - \beta x_{(i)})^2$$

we get

$$\begin{aligned}
\sum_{i=r+1}^{n-r} w_i(y_{(i)} - \alpha - \beta x_{(i)})^2 &= \sum_{i=r+1}^{n-r} w_i(y_{(i)} - \hat{\alpha} - \hat{\beta} x_{(i)})^2 \\
&= \sum_{i=r+1}^{n-r} w_i(y_{(i)} - \bar{y} - \hat{\beta}(x_{(i)} - \bar{x}))^2 \\
&= S^*_{y,y} + \hat{\beta}^2 S^*_{x,x} - 2\hat{\beta} S^*_{x,y} \\
&= S^*_{y,y} + \hat{\beta}(\hat{\beta} S^*_{x,x} - 2S^*_{x,y}) \\
&= S^*_{y,y} + \hat{\beta}(r\gamma\sigma(x_{[n-r]} - x_{[r+1]}) - S^*_{x,y}) \\
&= S^*_{y,y} + \hat{\beta}(r\gamma\sigma)(x_{[n-r]} - x_{[r+1]}) - \hat{\beta} S^*_{x,y} \\
&= S^*_{y,y} + \hat{\beta} r\gamma\sigma(x_{[n-r]} - x_{[r+1]}) - \\
&\quad \frac{S^{2*}_{x,y}}{S^*_{x,x}} - r\gamma\sigma\left\{x_{[n-r]} - x_{[r+1]}\right\}\frac{S^*_{x,y}}{S^*_{x,x}} \qquad (10.27)
\end{aligned}$$

therefore

$$\begin{aligned}
\frac{\partial \ln L^*}{\partial \sigma} &= -\frac{1}{\sigma^3}\left[A\sigma^2 + B\sigma - C\right] \\
&= 0 \qquad (10.28)
\end{aligned}$$

where

$$B = \frac{S^*_{x,y}}{S^*_{x,x}}\left\{r\gamma(x_{[n-r]} - x_{[r+1]})\right\} - r\gamma(y_{[n-r]} - y_{[r+1]}) \qquad (10.29)$$

and

$$C = S^*_{y,y} - \frac{S^{*2}_{x,y}}{S^*_{x,x}} = S^*_{y,y}\left\{1 - \frac{S^{*2}_{x,y}}{S^*_{x,x}S^*_{y,y}}\right\}$$

Upon using the fact that $\delta \geq 0$ (see Tiku, Tan Balakrishnan, 1986) and the result noted by a simple application of the Cauchy-Schwarz inequality, viz.,

$$S^{*2}_{x,y} \leq S^*_{x,x}S^*_{y,y},$$

it follows that $C \geq 0$. Hence, the quadratic equation in (10.28) admits only one positive root thus yielding the MML estimator of σ to be

$$\hat{\sigma} = \frac{-B + \sqrt{(B^2 + 4AC)}}{2A} \tag{10.30}$$

Since $\frac{r\gamma(x_{[n-r]} - x_{[r+1]})}{S^*_{x,x}}$ will usually be small, (10.25) essentially gives the simplified estimator

$$\hat{\beta} \approx \frac{S^*_{x,y}}{S^*_{x,x}}. \tag{10.31}$$

It is of interest to note here that the simplified estimator $\hat{\beta}$ is exactly of the same form as the classical least squares estimator of β. It is this simplified estimator which will be studied in the rest of this chapter. Furthermore, following the suggestion of Tiku and Stewart (1977) and Tiku, Tan and Balakrishnan (1986), we may correct the estimator $\hat{\sigma}$ in (10.30) for its bias by defining

$$\hat{\sigma} = \frac{-B + \sqrt{B^2 + 4AC}}{2\{A(A-2)\}^{1/2}} \tag{10.32}$$

We may observe here that when $r = 0$ the MML estimators $\hat{\alpha}, \hat{\beta}$ and $\hat{\sigma}$ given in (10.24) , (10.31) and (10.32) simply reduce to the corresponding classical estimators.

10.3 Determination of $x_{[i]}$ and $y_{[i]}$

The MML estimators of α, β and σ derived in the preceding section are all based on the order statistics of Z_i defined in (10.2). We realize that the variables Z_i themselves involve the unknown parameters α, β and σ. Hence, one needs to use some initial estimates of these parameters in order to start the ordering of these Z_i. For this purpose, Yale and Forsythe (1976) and Tan and Tabatabai (1988) have used two methods, one based on the least square estimators while the other is based on a procedure similar to the PRESS method of Allen (1971).

To this end, we first note that if the parameters α and β are estimated by least-squares method and the least-squares estimates are $\tilde{\alpha}$ and $\tilde{\beta}$, the estimated residuals are

$$\tilde{\epsilon}_i = y_i - \tilde{\alpha} - \tilde{\beta} x_i, \quad i = 1, 2, \ldots, n \tag{10.33}$$

Under the assumption that the error variables ϵ_i in (10.1) are distributed as normal $N(0,\sigma^2)$, we have the estimated residuals $\tilde{\epsilon}_i$ in (10.33) to be normally distributed with mean 0 and variance given by

$$Var(\tilde{\epsilon}_i) = \sigma^2 \left\{ \frac{n-1}{n} - \frac{(x_i - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \right\} \tag{10.34}$$

We observe from (10.34) that the estimated residuals $\tilde{\epsilon}_i$ have a larger variation if the chosen x_i are closer to their mean $\bar{x}$. As pointed out by Behnken and Draper (1972) and Barnett and Lewis (1978), we need to take this "ballooning effect" of the estimated residuals into account in examining the size of the residuals $\tilde{\epsilon}_i$ for possible identification of outliers among them.

By taking (10.34) into account, we may therefore use the "normalized" estimated residuals given by

$$\frac{\tilde{\epsilon}_i}{s_i} = \frac{\tilde{\epsilon}_i}{s\left\{\frac{n-1}{n} - \frac{(x_i-\bar{x})}{\sum_{j=1}^{n}(x_j-\bar{x})^2}\right\}^{1/2}} \tag{10.35}$$

where $s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\tilde{\epsilon}_i^2$ is an unbiased estimate of σ^2 and,s_i^2 is an unbiased estimate of $Var(\tilde{\epsilon}_i)$ given in (10.34).

Following the notations of Theil (1971), Hoaglin and Welsch (1978), and Belsley, Kuh and Welsch (1980), we may note here that the normalized estimated residuals we have defined in (10.35) are nothing but their "standard residuals" given by

$$\frac{\tilde{\epsilon}_i}{s\sqrt{1-h_i}}, \tag{10.36}$$

where h_i are simply the diagonal elements of their hat matrix.

In order to determine the $x_{[i]}$ and $y_{[i]}$ necessary for the computation of the MML estimators of $\hat{\alpha}, \hat{\beta}$ and $\hat{\sigma}$ derived in the preceding section, we proceed as follows:

1. From the give data, we first obtain the least square estimates of α and β;

2. Next, from (10.36) we compute the normalized estimated residuals for all the n pairs observations and then arrange them in increasing order of magnitude; and

3. $(x_{[i]}, y_{[i]})$ is taken to be the (x, y) pair corresponding to the $i'th$ largest normalized estimated residual for $i = 1, 2, \ldots, n$.

Mentioned should be made of the work Tan and Tabatabai (1988), who has developed similar estimators based on Tiku (1974) MML estimation method. However our work includes an excellent way of determining $x_{[i]}$ and $y_{[i]}$ which we described earlier.

10.4 Asymptotic properties

Let us denote $\mu_{i:n}$ and $\mu_{i:n}^{(2)}$ $(1 \geq i \geq n)$ for the first and second moment, respectively, of the $i'th$ order statistic in sample of size n from the standard normal distribution. These quantities have been quite extensively tabulated by Harter (1961) and Tietjen, Kahaner and Beckman (1971).

Then, we obtain from Equs. (10.12)-(10.14):

$$\begin{aligned}
\frac{\partial^2 \ln L^*}{\partial \alpha^2} &= \frac{1}{\sigma}\frac{\partial}{\partial \alpha}\left[\sum_{i=r+1}^{n-r} w_i Z_{(i)}\right] \\
&= -\frac{\sum_{i=r+1}^{n-r} w_i}{\sigma^2} \\
&= -\frac{m}{\sigma^2}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \ln L^*}{\partial \alpha \partial \beta} &= \frac{1}{\sigma}\frac{\partial}{\partial \beta}\left[\sum_{i=r+1}^{n-r} w_i Z_{(i)}\right] \\
&= -\frac{\sum_{i=r+1}^{n-r} w_i x_{[i]}}{\sigma^2} \\
&= -\frac{m\bar{x}}{\sigma^2}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \ln L^*}{\partial \alpha \partial \sigma} &= -\frac{\sum_{i=r+1}^{n-r} w_i Z_{(i)}}{\sigma^2} - \frac{\sum_{i=r+1}^{n-r} w_i Z_{(i)}}{\sigma^2} \\
&= -\frac{2}{\sigma^2}\sum_{i=r+1}^{n-r} w_i Z_{(i)}
\end{aligned}$$

$$\frac{\partial^2 \ln L^*}{\partial \beta^2} = \frac{1}{\sigma}\frac{\partial}{\partial \beta}\left[\sum_{i=r+1}^{n-r} w_i x_{[i]} Z_{(i)}\right]$$
$$= -\frac{1}{\sigma^2}\sum_{i=r+1}^{n-r} w_i x_{[i]}$$

$$\frac{\partial^2 \ln L^*}{\partial \beta \partial \sigma} = -\frac{1}{\sigma^2}\sum_{i=r+1}^{n-r} w_i Z_{(i)} - \frac{1}{\sigma^2}\sum_{i=r+1}^{n-r} w_i Z_{(i)} - \frac{r\gamma}{\sigma^2}(x_{[n-r]} - x_{[r+1]})$$
$$= -\frac{2}{\sigma^2}\sum_{i=r+1}^{n-r} w_i Z_{(i)} - \frac{r\gamma}{\sigma^2}(x_{[n-r]} - x_{[r+1]})$$

and

$$\frac{\partial^2 \ln L}{\partial \sigma^2} = -\frac{1}{\sigma^2}\left\{3\sum_{i=r+1}^{n-r} w_i Z_{(i)}^2 + 2r\gamma(Z_{(n-r)} - Z_{(r+1)}) - A\right\}.$$

Therefore

$$E\left[-\frac{\partial^2 \ln L}{\partial \alpha^2}\right] \approx E\left[-\frac{\partial^2 \ln L^*}{\partial \alpha^2}\right] = \frac{m}{\sigma^2} \tag{10.37}$$

$$E\left[-\frac{\partial^2 \ln L}{\partial \alpha \beta}\right] \approx E\left[-\frac{\partial^2 \ln L^*}{\partial \alpha \beta}\right] = \frac{m\bar{x}}{\sigma^2} \tag{10.38}$$

$$E\left[-\frac{\partial^2 \ln L}{\partial \alpha \sigma}\right] \approx E\left[-\frac{\partial^2 \ln L^*}{\partial \alpha \sigma}\right] = 0 \tag{10.39}$$

$$E\left[-\frac{\partial^2 \ln L}{\partial \beta^2}\right] \approx E\left[-\frac{\partial^2 \ln L^*}{\partial \beta^2}\right] = \frac{mD_1}{\sigma^2} \tag{10.40}$$

$$E\left[-\frac{\partial^2 \ln L}{\partial \beta \sigma}\right] \approx E\left[-\frac{\partial^2 \ln L^*}{\partial \beta \sigma}\right] = \frac{mD_2}{\sigma^2} \tag{10.41}$$

and

$$E\left[-\frac{\partial^2 \ln L}{\partial \sigma^2}\right] \approx E\left[-\frac{\partial^2 \ln L^*}{\partial \sigma^2}\right] = \frac{mD_3}{\sigma^2} \tag{10.42}$$

where

$$D_1 = \frac{1}{m}\left\{\sum_{i=r+1}^{n-r} w_i x_{[i]}^2\right\} \tag{10.43}$$

$$D_2 = \frac{2}{m}\left\{\sum_{i=r+1}^{n-r} w_i x_{[i]}\mu_{i:n} +\right\} + \frac{r\gamma}{m}(x_{[n-r]} - x_{[r+1]}) \tag{10.44}$$

and

$$D_3 = \frac{3}{m}\sum_{i=r+1}^{n-r} w_i \mu_{i:n}^{(2)} + \frac{2r\gamma}{m}(\mu_{[n-r]} - \mu_{[r+1]}) - \frac{A}{m} \tag{10.45}$$

Since the linear approximations in (10.10) and (10.11) are very close approximates L very closely so that inference procedures for (α, β, σ) based on L^* would be almost as efficient as those based on L. We may note further that asymptotically L^* is identical to L, which establishes the asymptotic optimality of the MML estimators. Hence, by using standard results of maximum likelihood estimation, we have the following theorem which gives the asymptotic distribution of the MML estimators $(\hat{\alpha}, \hat{\beta}, \hat{\sigma})$ of (α, β, σ). For a proof of this theorem, reference may be made to Kendall and Stuart (1973) or Rao (1975).

Theorem 20 *Asymptotically, that is, for large $A = n - 2r$ (hence large n), $(\hat{\alpha}, \hat{\beta}, \hat{\sigma})$ has a trivariate normal distribution with mean vector $(E(\hat{\alpha}) = \alpha, E(\hat{\beta}) = \beta, E(\hat{\sigma}) = \sigma)$ variance-covariance matrix*

$$V = \begin{bmatrix} E\left[-\frac{\partial^2 \ln L^*}{\partial \alpha^2}\right] & E\left[-\frac{\partial^2 \ln L^*}{\partial \alpha \beta}\right] & E\left[-\frac{\partial^2 \ln L^*}{\partial \alpha \sigma}\right] \\ & E\left[-\frac{\partial^2 \ln L^*}{\partial \beta^2}\right] & E\left[-\frac{\partial^2 \ln L^*}{\partial \beta \sigma}\right] \\ & & E\left[-\frac{\partial^2 \ln L^*}{\partial \sigma^2}\right] \end{bmatrix}^{-1} \tag{10.46}$$

$$V = \frac{\sigma^2}{m} \begin{bmatrix} 1 & \bar{x} & 0 \\ & D_1 & D_2 \\ & & D_3 \end{bmatrix}^{-1} \tag{10.47}$$

$$V = \frac{\sigma^2}{m(D_1 D_3 - D_2^2 - \bar{x}^2 D_3)} \begin{bmatrix} D_1 D_3 - D_2^2 & -\bar{x} D_3 & \bar{x} D_2 \\ & D_3 & -D_2 \\ & & D_1 - \bar{x}^2 \end{bmatrix} \tag{10.48}$$

10.5 Comparison with the classical estimators

For assessing the robustness features of the classical estimators and also the MML estimators of (α, β, σ), we follow the lines of Yale and Forsythe (1976) and Tan and Tabatabai (1988) and assume a simple linear regression model $y_i = \alpha + \beta x_i + \epsilon_i$ $(i = 1, 2, \ldots, n)$; here α and β are the regression parameters and ϵ_i's are independent and identically distributed random disturbances with mean 0 and variance σ^2. Then we

let the explanatory variables x_i be taken from a standard normal population and ϵ_i be taken from a normal population with mean 0 and variance σ^2, but $Pr(\sigma = 1) = p$ and $Pr(\sigma = \sigma\prime) = 1 - p$. We make the following choices for these parameters: $\alpha = 0$, $\beta = 1$, $\sigma\prime = 1, 4, 6$ and $p = 1.0, 0.9, 0.8, 0.7$. Two sample sizes $n = 10$ and $n = 20$ have been considered for this simulation study. A somewhat different mixture model has been utilized by Elashoff (1972) in this context.

We then simulated the values of (1) Bias of $\hat{\alpha}$, (2) Bias of $\hat{\beta}$, (3) Bias of $\hat{\sigma}$, (4) Variance of $\hat{\alpha}$, (5) Variance of $\hat{\beta}$, and (6) Variance of $\hat{\sigma}$ for the maximum likelihood estimators of (MLE's) and also for the modified maximum likelihood estimators (MMLE's).

It is of interest to mention here that it can be easily shown that, under normality, the bias in all these estimators is zero for large n. These simulations have all been based on 2,000 Monte Carlo runs. These values for the cases $n = 10$ and $n = 20$ are presented in Table 10.1 and 10.2, respectively. The MML estimators $\hat{\alpha}, \hat{\beta}$ and $\hat{\sigma}$ have been computed as described in Section 5.

Table 10.1: Values of bias and variance of the MLE's and MMLE's of α, β and σ; $n = 10$

p	σ'	Estimator	(1)	(2)	(3)	(4)	(5)	(6)
			Bias			Variance		
1.0	1.0	MLE	-0.0083	0.0023	-0.0338	0.1204	0.1292	0.0596
		MMLE	-0.0076	0.0057	-0.1167	0.1423	0.2577	0.0792
0.9	4.0	MLE	-0.0150	0.0012	0.1903	0.1985	0.2185	0.2862
		MMLE	-0.0073	0.0171	-0.0424	0.1628	0.2867	0.1176
0.9	6.0	MLE	-0.0194	0.0005	0.3789	0.3037	0.3381	0.7461
		MMLE	-0.0089	0.0157	-0.0166	0.1774	0.2950	0.1663
0.8	4.0	MLE	-0.0232	-0.0045	0.4166	0.2819	0.3368	0.4724
		MMLE	-0.0107	0.0142	0.0310	0.1999	0.3193	0.1656
0.7	4.0	MLE	-0.0228	0.0006	0.6276	0.3726	0.4326	0.5897
		MMLE	-0.0176	0.0088	0.1442	0.2545	0.3875	0.2619
0.7	6.0	MLE	-0.0324	-0.0004	1.1624	0.7053	0.8364	1.5579
		MMLE	-0.0225	0.0086	0.2849	0.3666	0.5190	0.5718
0.6	4.0	MLE	-0.0130	-0.0034	0.8188	0.4639	0.5374	0.6805
		MMLE	-0.0110	0.0172	0.2601	0.3289	0.5131	0.3559
0.6	6.0	MLE	-0.0161	-0.0072	1.4949	0.9221	1.0832	1.7561
		MMLE	-0.0099	0.0128	0.5002	0.5376	0.8129	0.8304

Table 10.2: Values of bias and variance of the MLE's and MMLE's of α, β and σ; $n = 20$

p	σ'	Estimator	(1)	(2)	(3)	(4)	(5)	(6)
			Bias			Variance		
1.0	1.0	MLE	-0.0012	0.0003	-0.0141	0.0556	0.0594	0.0282
		MMLE	-0.0015	0.0030	-0.0608	0.0613	0.1173	0.0403
0.9	4.0	MLE	-0.0036	0.0000	0.2405	0.0935	0.0995	0.1829
		MMLE	-0.0012	0.0027	0.0032	0.0705	0.1272	0.0520
0.9	6.0	MLE	-0.0052	-0.0002	0.4720	0.1422	0.1534	0.5193
		MMLE	0.0007	0.0085	0.0842	0.0866	0.1415	0.0731
0.8	4.0	MLE	-0.0015	-0.0126	0.4870	0.1377	0.1430	0.2778
		MMLE	0.0007	0.0085	0.0842	0.0866	0.1415	0.0731
0.8	6.0	MLE	-0.0017	-0.0212	0.9244	0.2437	0.2548	0.7750
		MLE	-0.0012	0.0062	0.1375	0.1020	0.1465	0.1165
0.7	4.0	MLE	0.0003	-0.0132	0.6920	0.1734	0.1913	0.3382
		MMLE	0.0032	0.0072	0.1802	0.1090	0.1618	0.1103
0.7	6.0	MLE	0.0014	-0.0222	1.2882	0.3257	0.3665	0.9052
		MMLE	0.0014	0.0017	0.2876	0.1497	0.1988	0.2274
0.6	4.0	MLE	0.0011	-0.0121	0.8875	0.2139	0.2340	0.3723
		MMLE	0.0023	0.0048	0.2876	0.1414	0.1900	0.1510
0.6	6.0	MLE	0.0026	-0.0204	1.6251	0.4179	0.4660	0.9651
		MMLE	0.0005	-0.0012	0.4696	0.2190	0.2672	0.3656

It is clear from Tables 10.1 and 10.2 that the MML estimators of α and β have bias almost zero under both normal and mixture-normal models. Under normality, the MML estimator of σ has a little bias for small n. We also see that under normality, the MMLE's of α, β and σ have larger variances than the corresponding MLE's as expected. However, the MLE's of α and β continue to have almost zero bias even under the mixture-normal models, but their variances increase substantially as the nonnormality becomes pronounced. Furthermore, we note from Table 10.1 and 10.2 that the effect of nonnormality on the MLE of σ is quite drastic. Even a small departure from normality is seen to increase its bias as well as variance very drastically. On the other hand, the MML estimator of σ displays its robustness characteristic through much smaller bias and much smaller variance.

10.6 Comparison with other robust procedures for departures from normality

In this section we first briefly describe the Adichie's (1976a,b) non-parametric estimation method, Andrews's (1973,1974) M-estimation methods based on two ψ functions discussed in Andrews et al. (1972) and Huber (1981), Yale and Forsythe's (1976) Winsorized estimation method and Tan and Tabatabai (1988) Modified winsorized estimation method.

10.6.1 Classical estimators

Let us consider the simple linear regression model given in (10.1). Then, the classical estimators of α, β and σ^2 are given by

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{10.49}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \hat{x})^2} \tag{10.50}$$

and

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{(n-2)}; \tag{10.51}$$

Under the assumption of normality for the error variables ϵ_i's in (10.1), the classical estimators have several distributional and optimal properties (see Myers, 1986). By using Monte Carlo simulations, we show in this section that the classical estimators (particularly the estimator of σ) are very sensitive to departures from normality. We shall compare these classical estimators of α, β and σ to some non-parametric and robust estimators which are described below.

10.6.2 Adichie's estimators

Adichie (1967a,b) proposed a non-parametric method of estimation that is implicit in nature in which the estimator of the slope β in (10.1) is a weighted median of the set of all pairwise slopes $\{(y_j - y_i)/(x_j - x_i) for i \neq j\}$ with weights being proportional to the absolute distance $|x_i - x_j|$ between the values of the independent variable x in the linear regression model. The corresponding estimator of the intercept α is given by

$$\hat{\alpha} = \underset{i \leq j}{Med} \frac{1}{2}\{y_i + y_j - \hat{\beta}(x_i + x_j)\} \tag{10.52}$$

Draper (1988) has given these estimators in an explicit form. It needs to be mentioned that Adichie has not given an estimator of σ; however, we used the form of the estimator in (10.51) for this purpose.

10.6.3 M-Estimators

Andrews (1973,1974) discussed the M-estimation of the parameters in a linear regression model. In the M-estimation method, one needs to solve

$$min_{\underline{\beta}} \sum_{i=1}^{n} \rho(e_i/s) = min_{\underline{\beta}} \sum_{i=1}^{n} \rho[(y_i - x_i'\beta)/s] \tag{10.53}$$

where s is a robust estimator of σ. A popular choice for s is $s = median|e_i - median(e_i)|/0.6745$, where the constant 0.6745 makes s an approximately unbiased estimator of σ for large n and when the error variables are normally distributed.

To solve Eq. (2.6) for $\underline{\beta}$ one could follow the approach used by Beaton and Tukey (1974), namely, "iteratively reweighted least squares method", which is briefly described below.

1. Obtain the estimated residuals using least-squares method.

2. Calculate weights using

$$w_{i0} = \begin{cases} \frac{\psi[(y_i-\hat{\alpha}-\hat{\beta}x_i)/s]}{(y_i-\hat{\alpha}-\hat{\beta}x_i)/s} & \text{if } y_i \neq \hat{\alpha} + \hat{\beta}x_i \\ 1 & \text{if } y_i = \hat{\alpha} + \hat{\beta}x_i \end{cases} \tag{10.54}$$

3. Solve $\underline{X}'\underline{W}_0\underline{X}\underline{\beta} = \underline{X}'\underline{W}_0\underline{Y}$, weighted least square for $\hat{\beta}$, where $\underline{W}_0$ matrix is a diagonal matrix with diagonal elements as given above.

4. Use these estimators and go to step 2. Repeat this process to achieve the required convergence.

In Andrews' method $\psi()$ function is taken to be

$$\psi(z) = \begin{cases} \sin(z/a) & \text{if } |z| \leq a\pi \\ 0 & \text{if } |z| > a\pi \end{cases} \tag{10.55}$$

and, consequently, the weight function $w()$ is

$$w(z) = \begin{cases} \frac{\sin(z/a)}{(z/a)} & \text{if } |z| \leq a\pi \\ 0 & \text{if } |z| > a\pi \end{cases} \tag{10.56}$$

The choice for a was made as 1.5 in our study. In Huber's method $\psi()$ function is taken to be

$$\psi(z) = \begin{cases} z & \text{if } |z| \leq 1.5 \\ 1.5sign(z) & \text{if } |z| > 1.5 \end{cases} \tag{10.57}$$

and the corresponding weight function $w(z)$ is given by

$$w(z) = \begin{cases} 1.0 & \text{if } |z| \leq 1.5 \\ \frac{1.5}{|z|} & \text{if } |z| > 1.5 \end{cases} \tag{10.58}$$

10.6.4 Winsorized estimators

The winsorized method of estimation of parameters in a linear regression model is due to Yale and Forsythe (1976). This method requires the estimated residuals which may be obtained by one of the following two ways. The first one, referred to as the RES method, is simply to compute the least-squares estimates of α and β based on all the observations and then obtain the residuals. The other, referred to as PRESS method (Allen, 1971), is to compute the least-squares estimates of α and β without the ith observation and then use these estimates of α and β to obtain the ith predicted residual.

After determining the estimated residuals by either the RES or the PRESS method, the Winsorized estimators of α, β and σ are determined as described below.

1. Order the estimated residuals $\hat{e}_i$'s as $\hat{e}_{(1)} \leq \hat{e}_{(2)} \leq \ldots \leq \hat{e}_{(n)}$ and let $(x_{(i)}, y_{(i)})$ be the (x, y) pair associated with the ith estimated residual $\hat{e}_i$

2. The estimators of α, β and σ^2 are then given by

$$\hat{\alpha} = \hat{y}_w - \hat{x}_w\hat{\beta}, \tag{10.59}$$

$$\hat{\beta} = \frac{\sum_{i=r+1}^{n-r} w_{(i)}(x_{(i)} - \hat{x}_w)y_{(i)}}{\sum_{i=r+1}^{n-r} w_{(i)}(x_{(i)} - \hat{x}_w)^2} \tag{10.60}$$

and

$$\hat{\sigma}^2 = \frac{H(\alpha, \beta)}{(n - 2r - 2)} \tag{10.61}$$

where

$$r = [0.1n],$$

$$w_i = \begin{cases} 1 + r & \text{if } i = r + 1 \\ 1 & \text{if } r + 1 < i < n - r \\ 1 + r & \text{if } i = n - r \end{cases}, \tag{10.62}$$

$$\hat{x}_w = \frac{\sum_{i=r+1}^{n-r} w_i x_{(i)}}{\sum_{i=r+1}^{n-r} w_i}, \tag{10.63}$$

$$\hat{y}_w = \frac{\sum_{i=r+1}^{n-r} w_i y_{(i)}}{\sum_{i=r+1}^{n-r} w_i}, \tag{10.64}$$

and

$$H(\alpha,\beta) = \sum_{i=r+1}^{n-r} w_i(y_{(i)} - \alpha - \beta x_{(i)})^2 \tag{10.65}$$

10.6.5 Modified Winsorized estimators

Tan and Tabatabai (1988) modified the formulae of the Winsorized estimators given by Yale and Forsythe (1976) and defined the modified Winsorized estimators of the parameters in a linear regression model. The formulae for their estimators of α, β and σ are exactly the same as those of the Winsorized estimators in (10.59), (10.60) and (10.61), respectively, with the weights w_i defined in (10.62) replaced by

$$w_i = \begin{cases} 1 + r\delta & \text{if } i = r+1 \\ 1 & \text{if } r+1 < i < n-r \\ 1 + r\delta & \text{if } i = n-r \end{cases} \tag{10.66}$$

In the above formula, $\delta = \frac{g(h_2)-g(h_1)}{h_2-h_1}$ where $g(x) = -\frac{d}{dx}[\log(1-F(x))]$ and h_1 and h_2 is given by

$$\begin{aligned} 1 - F(h_1) &= q + [q(1-q)/n]^{1/2} \\ 1 - F(h_2) &= q - [q(1-q)/n]^{1/2} \end{aligned}$$

with $q = r/n$ and $F()$ denoting the cumulative distribution function of standard normal variate. We may note here that these estimators are almost the same as the MML estimators that we have derived in Section 2 with the difference being in the estimator of σ and also in the determination of $(x_{[i]}, y_{[i]})$.

10.6.6 Monte Carlo simulation

In this section, we consider all the methods of estimation mentioned above and study their bias and mean square error through Monte Carlo simulations. We examine the performance of these estimators by considering normal and a wide range of non-normal models for the error component in the simple linear regression model and also for various sample sizes. Finally, we consider the examples given by Daniel and Wood (1971) and Mendenhall (1983) and illustrate all the methods of estimation discussed in this study.

For assessing the robustness features of all the estimators described in Section 10.6, we followed the lines of Yale and Forsythe (1976), Tan and Tabatabai (1988), and Singh and Gupta (1989) and resorted to the following Monte Carlo simulation process. We considered the simple linear regression model in (10.1) with the choice of the parameters α, β and σ having been taken as 0, 1 and 1, respectively. The explanatory variables x_i's were sampled independently from the standard normal distribution. Further, we considered the following wide range of distributional models for the independent error variables ϵ_i's in (10.1):

1. Standard normal distribution
2. Mixture-normal model $0.9N(0,1) + 0.1N(0,4)$,
3. Mixture-normal model $0.9N(0,1) + 0.1N(0,6)$,
4. Mixture-normal model $0.8N(0,1) + 0.2N(0,4)$,
5. Single-outlier model with $(n-1)N(0,1)\&1N(0,4)$,
6. Single-outlier model with $(n-1)N(0,1)\&1N(0,6)$,
7. Two-outlier model with $(n-2)N(0,1)\&2N(0,4)$,
8. Two-outlier model with $(n-2)N(0,1)\&2N(0,6)$,
9. Double exponential distribution with mean 0 and variance 1,
10. Logistic distribution with mean 0 and variance 1,
11. Lognormal distribution with mean 0 and variance 1,
12. t_{12}-distribution with mean 0 and variance 1,
13. t_9-distribution with mean 0 and variance 1,
14. t_5-distribution with mean 0 and variance 1,
15. Uniform distribution with mean 0 and variance 1,

16. Gamma distribution ($P = 4$) with mean 0 and variance 1,

17. Gamma distribution ($P = 2$) with mean 0 and variance 1,

18. Gamma distribution ($P = 1$) with mean 0 and variance 1,

We simulated the values of the bias and mean square error of all the estimators of (α, β, σ) for all the 18 distributions listed above. The values corresponding to first eight distributions are presented in Tables 10.9-10.16 for sample sizes 10, 20 and 40. All the results given in these tables are based on 4,000 Monte Carlo runs. The necessary samples from the normal, uniform, and chi-square distributions were generated by the IMSL subroutines RNNOR, RNUN, and RNCHI, respectively.

10.6.7 Comparison and conclusions

From the simulated values of the bias and the mean square error of various estimators of α, β and σ under normal and a very wide range of nonnormal models for the error variables ϵ_i in the simple linear regression model (10.1), we observe the following points:

1. From Table 10.9 we see that the classical estimators have the smallest mean square errors which is to be expected as they are the best estimators when the underlying distribution for the error variable is normal. We also observe that the Adichie's non parametric estimators are almost as efficient as the classical estimators. The bias of all the estimators of α, β and σ decrease as the sample size increases; however, we note that the bias of the Winsorized and the modified Winsorized estimators of σ decrease only nominally as n increases and, as a result, these two estimators of σ remain quite biased even for a sample of size as large as 40. The M-estimators of α and β perform better than the corresponding MML estimators, but the M-estimators of σ have larger mean square errors than the MML estimators of σ.

2. From Table 10.10-10.16 we see that both classical and Adichie's estimators of σ develop serious bias and have very large mean square error as compared to all other estimators of σ when the underlying model for the error variable has scale contamination or contains a few outliers. In general, when all the estimators develop larger bias and larger mean square error under increasing departures from normality, the classical and Adichie's estimators of σ seem to face a devastating effect in this case. We observe that the bias of the Winsorized and modified Winsorized estimators of σ decrease marginally when the sample size increases and, hence, they remain quite biased even when the sample size is as large as 40. We also note that while the M-estimators of α and β have smaller mean square errors than the corresponding MML estimators, the MML estimator of σ performs better than the M-estimators of σ.

3. For the Models 9-18, we observed that the classical and Adichie's estimators perform very well and have smaller bias and mean square error than all other estimators. In addition, we find that the MML estimator of σ has smaller bias and mean square error than the Winsorized, modified Winsorized, and the M-estimators of σ; however, the M-estimators of α and β perform better than the corresponding MML estimators.

4. We noticed that the MML estimators of α, β and σ have smallest mean square error when the error variable has a lognormal distribution. We see in this case that the MML estimator of σ is quite biased and has a larger bias than the classical and Adichie's estimator of σ and much smaller bias than the Winsorized, modified Winsorized and the M-estimators of σ.

5. For the Models 16-18, we find that the classical and Adichie's estimators of α, β and σ perform very well, when the error variable has a gamma distribution. Among all other methods, the MML estimator of σ has the smallest bias and mean square error values.

From all the points mentioned above, we conclude that the classical and Adichie's estimators, though they perform well under normal and many non-normal

models, are extremely sensitive (particularly the estimators of σ) to presence of outliers and also to scale-mixtures. The MML estimators seem to perform better in almost all cases than the Winsorized and the modified Winsorized estimators and in this regard we observe the MML estimators based on the RES method to be superior to those based on the PRESS method in almost all cases. It should be mentioned here, however, that the M-estimators of α and β in most cases turn out to be better than the corresponding MML estimators while MML estimator of σ almost always has smaller bias and mean square error than the M-estimators of σ. In conclusion, we therefore, would recommend the usage of the M-estimation (based on either Andrews' ψ-function or Huber's ψ-function) or the modified maximum likelihood estimation based on the RES method for the estimation of the parameters α, β and σ in the simple linear regression model in (10.1) whenever non-normality is suspected in the error variables and particularly when one or more outliers are suspected to be present in the data.

10.6.8 Illustrative examples

In this section we consider two examples, one given by Daniel and Wood (1971) and the other given by Mendenhall (1983), and illustrate the various methods of estimation examined in this empirical study. These two examples will also help explain some of the points made in Section 10.6.7.

Example 10.1:

Let us consider the pilot-plant data given by Daniel and Wood (1971). In this example, the independent variable x is the acid number of a chemical determined by titration and the dependent variable y is its organic acid content determined by extraction and weighing. The data, consisting of 20 pairs of observations, is presented in the two rows of Table 10.3 This example has been utilized by Yale and Forsythe (1976), Tan and Tabatabai (1988), and Balakrishnan and Ambagaspitiya (1990) to illustrate their methods of estimation of the parameters α, β and σ.

Table 10.3: Daniel and Wood's pilot-plant data

Obs.	1	2	3	4	5	6	7	8	9	10
y	76	70	55	71	55	48	50	66	41	43
x	123	109	62	104	57	37	44	100	16	28

Obs.	11	12	13	14	15	16	17	18	19	20
y	82	68	88	58	64	88	89	88	84	88
x	138	105	159	75	88	164	169	167	149	167

Table 10.4: Estimates of parameters for Daniel and Wood's data

Estimator	α	β	σ
MML Estimators (RES)	35.096	0.325	1.294
MML Estimators (PRESS)	35.573	0.320	1.375
Winsorized Estimators (RES)	35.012	0.325	1.204
Winsorized Estimators (PRESS)	35.249	0.323	1.258
M-Winsorized Estimators (RES)	35.091	0.325	1.175
M-Winsorized Estimators (PRESS)	35.336	0.322	1.219
Andrew's Estimators	35.481	0.321	1.594
Huber' Estimators	35.458	0.322	1.594
Adichie's Estimator	35.331	0.323	1.231

For this data, Daniel and Wood (1971) have shown that the simple linear regression model in (10.1) provides an excellent fit. The classical estimates of α, β and σ are obtained to be 35.458, 0.322 and 1.230, respectively. The variance of the classical estimates of α and β are obtained to be 1.472 and 0.00030, respectively. The MML estimates, the winsorized estimates, the modified Winsorized estimates, the M-estimates, and Adichie's non-parametric estimates were computed similarly for this data and given in Table 10.4. The variance of the MML Estimates (RES) of α and β are obtained to be 0.476 and 0.000038, respectively

In order to illustrate the winsorization method, Yale and Forsythe (1976) contaminated the pilot-plant data in Table 10.3. by changing the first y observation

Table 10.5: Estimates of parameters for Daniel and Wood's data with contamination

Estimator	α	β	σ
MML Estimators (RES)	35.8708	0.3195	1.7192
MML Estimators (PRESS)	35.8708	0.3195	1.7192
Winsorized Estimators (RES)	35.8056	0.3201	1.5884
Winsorized Estimators (PRESS)	35.8056	0.3201	1.5884
M-Winsorized Estimators (RES)	35.8672	0.3195	1.5341
M-Winsorized Estimators(PRESS)	35.8672	0.3195	1.5341
Andrew's Estimator (c=1.5)	35.5462	0.3202	1.2731
Huber's Estimator (c=1.5)	35.8177	0.3198	1.2731
Adichie's Estimators	35.9132	0.31944	7.5054

from 76 to 100 and the third y observation from 55 to 75. The classical estimates of α, β and σ are obtained in this case to be 38.377, 0.315 and 7.209, respectively. The variance of the classical estimates of α and β are obtained to be 13.854 and 0.0011, respectively. We note here that the introduction of two outliers in the data has had a devastating effect on the classical estimators and particularly so on the estimator of σ. We also computed the estimates of α, β and σ by all the methods described in Section 10.6 and these values are presented in Table 10.5. The variance of MML estimates (RES) of α and β are obtained to be 0.878 and 0.00058, respectively. We observe the robustness property of all these estimators as two outliers introduced in the data has had a minimal effect on these estimates. However, the estimate of σ based on Adichie's method is seen to get adversely affected by two outliers in the data and this is understandable for reasons discussed in the last section.

Example 10.2:

Let us consider the EPA 1980 mileage rating data given by Mendenhall (1983, pp.410). In this example, the independent variable x is the cylinder volume of the car and the dependent variable y is the combined mileage per gallon of the subcompact cars. The data, consisting of 9 pairs of observations, is presented in

Table 10.6: Mendenhall's EPA 1980 mileage rating data

Car	VW Rabit	Datsun 210	Chevette	Dodge Omni	Mazda 626
y	24	29	26	24	24
x	97	85	98	105	120

Car	Oldsmobile Starfire	Mercury Capri	Toyota Celica	Datsun 810
y	22	23	23	21
x	151	140	134	146

Table 10.7: Estimates of parameters for Mendenhall's data

Estimator	α	β	σ
MML Estimators (RES)	31.3982	-0.0637	0.9074
MML Estimators(PRESS)	31.4082	-0.0640	0.8661
Winsorized Estimators (RES)	31.2392	-0.0625	0.8473
Winsorized Estimators (PRESS)	31.2648	-0.0629	0.8196
M-Winsorized Estimators (RES)	31.3698	-0.0635	0.8345
M-Winsorized Estimators (PRESS)	31.3823	-0.0638	0.8106
Andrew's Estimator (c=1.5)	32.4343	-0.0718	0.7654
Huber's Estimator (c=1.5)	33.0564	-0.0765	0.7654
Adichie's Estimators	33.0849	-0.0755	1.2791

Table 10.6. This example has already been utilized by Tan and Tabatabai (1988) to study the performance of their modified Winsorized estimators for data with such small sample size.

For this data, Mendenhall (1983) has shown that the simple linear regression model in (10.1) provides an excellent fit. The classical estimates of α, β and σ are obtained to be 34.009, -0.084 and 1.259, respectively. The MML estimates, the winsorized estimates, the modified winsorized estimates, the M-estimates, and Adichie's non-parametric estimates were computed similarly for this data and are given in Table 10.7.

In order to illustrate the modified Winsorization method, Tan and Tabatabai

Table 10.8: Estimates of parameters for Mendenhall's data with contamination

Estimator	α	β	σ
MML Estimators (RES)	34.6007	-0.0873	1.4527
MML Estimators (PRES)	36.8267	-0.1027	1.2330
Winsorized Estimators (RES)	34.3636	-0.0856	1.3585
Winsorized Estimators(PRES)	34.3636	-0.0856	1.3585
M-Winsorized Estimators(RES)	34.5583	-0.0870	1.3339
M-Winsorized Estimators(PRES)	34.5583	-0.0870	1.3339
Andrew's Estimator (c=1.5)	33.7987	-0.0822	3.1736
Huber's Estimator (c=1.5)	37.8241	-0.1103	3.1736
Adichie's Estimators	37.2500	-0.1060	8.1301

(1988) contaminated the EPA 1980 mileage rating data in Table 9.24 by changing the third y observation from 26 to 48. Then, the classical estimates of α, β and σ are obtained in this case to be 48.531, -0.185 and 7.578, respectively. We note once again that the introduction of just one outlier in the data has had a devastating effect on the classical estimators. We also computed the estimates of α, β and σ by all the methods described in Section 10.6 and these values are presented in Table 10.8. We observe here, as in Example 10.1, that the introduction of the outlier in the data has had a very minimal effect on all these estimators. However, the estimate of σ based on Adichie's method is once again seen to be considerable affected by the presence of the outlier in the data. Furthermore, we see that the modified maximum likelihood estimation method provides efficient estimators for the parameters and also reduces the effect of outlying observations even for sample size as small as 9.

10.6.9 Comparison with other robust procedures for departures from linearity

In this section we study the robustness of all the estimation methods mentioned earlier in this section for the case of departures from linearity, in particular we consider the model $y_i = \alpha + \beta x_i + G x_i^2 + e_i, \quad i = 1, \ldots, n$. We investigate the

bias and mean-square error of all the methods of estimation when there is actually a departure from linearity of the type considered above. For this we simulated the values of the bias and mean square error of all the estimators of (α, β, σ) for all the 18 distributions (listed in Section 10.6.6). The values for distributions 1-8 for sample sizes 10 and 20, and $G = -0.10, -0.05, 0.05, 0.1$ are presented in Table 10.17-9.27. From these Tables we observed the following points:

1. From Tables 10.17 and 10.18 we see that the classical and Adichie's estimators have the smallest mean square errors for different values of G. The bias of all the estimators of α, β and σ increase as G increases and decrease as the sample size increases.

2. From Table 10.19-10.26 we see that both the classical and Adichie's estimators of σ develop serious bias and have very large mean square error as compared to all other estimators of σ when underlying model for the error variable has scale contamination or contains few outliers. Also we notice that this effect has much more influence than the departure from linearity has.

3. We observe that the MML estimators derived in this chapter turn out to be the most efficient one (in terms of bias and mean square error) under departures from the normal distribution for the error variable accompanied with or without departures from the assumed linear model.

Finally, we conclude that all the methods of estimation mentioned earlier in Section 6 are less sensitive to small departures from linearity than when the underlying model for the error variable has scale contamination or contains a few outliers. In addition, the MML method of estimation developed in this chapter is observed to be more stable to departures from normality as well as from linearity.

Table 10.9: Bias and Mean square error of the estimators under normal model for the error

	Bias			MSEror		
	α	β	σ	α	β	σ
			n=10			
Classical	-.0055	.0150	-.0262	.1144	.1462	.0602
MML (RES)	-.0037	.0257	-.1047	.1366	.2823	.0950
MML (PRESS)	-.0050	.0258	-.0790	.1488	.3361	.1038
Winsor (RES)	-.0052	.0229	-.1613	.1337	.2672	.0972
Winsor (PRESS)	-.0053	.0246	-.1519	.1456	.3214	.0998
M-Winsor (RES)	-.0051	.0224	-.1804	.1327	.2600	.1001
M-Winsor (PRESS)	-.0052	.0241	-.1706	.1434	.3108	.1017
Andrews (c = 1.5)	-.0054	.0157	-.1320	.1260	.1553	.1256
Huber's (c = 1.5)	-.0055	.0162	-.1320	.1210	.1508	.1256
Adiche	-.0064	.0169	-.0147	.1265	.1616	.0613
			n=20			
Classical	-.0051	-.0032	-.0167	.0528	.0586	.0282
MML (RES)	-.0044	-.0046	-.0625	.0592	.1132	.0442
MML (PRESS)	-.0056	-.0016	-.0530	.0614	.1180	.0456
Winsor (RES)	-.0038	-.0053	-.1306	.0593	.1186	.0506
Winsor (PRESS)	-.0051	-.0050	-.1258	.0615	.1232	.0506
M-Winsor (RES)	-.0039	-.0050	-.1569	.0586	.1114	.0559
M-Winsor (PRESS)	-.0052	-.0047	-.1522	.0604	.1153	.0555
Andrews (c = 1.5)	-.0039	-.0043	-.0690	.0563	.0620	.0658
Huber's (c = 1.5)	-.0037	-.0040	-.0690	.0553	.0608	.0658
Adiche	-.0077	-.0037	-.0127	.0562	.0636	.0284
			n=40			
Classical	-.0045	.0018	-.0064	.0256	.0267	.0129
MML (RES)	-.0039	.0004	-.0414	.0279	.0617	.0208
MML (PRESS)	-.0031	-.0009	-.0397	.0284	.0628	.0212
Winsor (RES)	-.0040	.0018	-.1125	.0283	.0691	.0285
Winsor (PRESS)	-.0031	-.0034	-.1124	.0286	.0703	.0287
M-Winsor (RES)	-.0038	.0018	-.1424	.0278	.0620	.0349
M-Winsor (PRESS)	-.0030	-.0028	-.1422	.0281	.0632	.0349
Andrews (c = 1.5)	-.0041	.0015	-.0307	.0266	.0281	.0332
Huber's (c = 1.5)	-.0042	.0015	-.0307	.0265	.0279	.0332
Adiche	-.0050	.0018	-.0048	.0268	.0286	.0130

Table 10.10: Bias and Mean square error of the estimators under mixture-normal model (with $p = 0.9$ and $\sigma = 4$) for the error

	Bias			MSError		
	α	β	σ	α	β	σ
			n=10			
Classical	-.0151	-.0042	.1982	.1987	.2261	.3330
MML (RES)	-.0070	.0002	-.0370	.1639	.2940	.1215
MML (PRESS)	-.0135	-.0005	-.0075	.1758	.3637	.1391
Winsor (RES)	-.0091	.0057	-.0978	.1641	.2752	.1115
Winsor (PRESS)	-.0102	.0038	-.0896	.1722	.3373	.1170
M-Winsor (RES)	-.0091	.0055	-.1182	.1623	.2673	.1106
M-Winsor (PRESS)	-.0103	.0037	-.1094	.1697	.3257	.1151
Andrews (c = 1.5)	-.0098	.0025	-.0657	.1501	.1816	.1400
Huber's (c = 1.5)	-.0125	-.0008	-.0657	.1506	.1844	.1400
Adiche	-.0131	-.0003	.2220	.1531	.1899	.3750
			n=20			
Classical	-.0014	-.0014	.2443	.0926	.1055	.2490
MML (RES)	-.0004	-.0024	.0036	.0697	.1251	.0531
MML (PRESS)	-.0011	-.0032	.0144	.0719	.1338	.0582
Winsor (RES)	.0001	-.0015	-.0701	.0703	.1316	.0496
Winsor (PRESS)	-.0005	-.0016	-.0637	.0717	.1384	.0502
M-Winsor (RES)	-.0001	-.0014	-.0982	.0692	.1240	.0511
M-Winsor (PRESS)	-.0005	-.0015	-.0919	.0703	.1299	.0512
Andrews (c = 1.5)	-.0003	.0016	-.0009	.0641	.0739	.0764
Huber's (c = 1.5)	-.0005	.0003	-.0009	.0659	.0758	.0764
Adiche	-.0049	.0005	.2569	.0664	.0757	.2677
			n=40			
Classical	-.0045	.0002	.2819	.0459	.0504	.1883
MML (RES)	-.0028	.0074	.0180	.0339	.0705	.0251
MML (PRESS)	-.0024	.0081	.0199	.0344	.0722	.0257
Winsor (RES)	-.0042	.0050	-.0589	.0344	.0782	.0239
Winsor (PRESS)	-.0023	.0065	-.0564	.0347	.0794	.0241
M-Winsor (RES)	-.0040	.0048	-.0906	.0338	.0704	.0269
M-Winsor (PRESS)	-.0023	.0060	-.0884	.0339	.0716	.0268
Andrews (c = 1.5)	-.0026	.0031	.0198	.0303	.0321	.0376
Huber's (c = 1.5)	-.0034	.0025	.0198	.0314	.0337	.0376
Adiche	-.0042	.0025	.2888	.0317	.0341	.1964

Table 10.11: Bias and Mean square error of the estimators under mixture-normal model (with $p = 0.9$ and $\sigma = 6$) for the error

	Bias			MsError		
	α	β	σ	α	β	σ
			n=10			
Classical	-.0176	-.0070	.3915	.3107	.3521	.9170
MML (RES)	-.0078	.0029	-.0120	.1785	.3039	.1649
MML (PRESS)	-.0137	-.0018	.0185	.1899	.3999	.1984
Winsor (RES)	-.0091	.0034	-.0750	.1808	.2916	.1450
Winsor (PRESS)	-.0096	.0055	-.0651	.1867	.3562	.1530
M-Winsor (RES)	-.0091	.0032	-.0957	.1785	.2834	.1411
M-Winsor (PRESS)	-.0096	.0052	-.0851	.1835	.3440	.1480
Andrews (c = 1.5)	-.0097	.0004	-.0211	.1631	.2148	.1809
Huber's (c = 1.5)	-.0131	-.0019	-.0211	.1678	.2230	.1809
Adiche	-.0135	-.0012	.4322	.1613	.2083	1.0609
			n=20			
Classical	-.0014	-.0019	.4813	.1426	.1660	.7667
MML (RES)	-.0017	-.0039	.0209	.0729	.1271	.0649
MML (PRESS)	-.0004	-.0059	.0337	.0755	.1328	.0711
Winsor (RES)	-.0012	-.0021	-.0540	.0740	.1348	.0570
Winsor (PRESS)	-.0006	-.0044	-.0472	.0762	.1385	.0586
M-Winsor (RES)	-.0012	-.0020	-.0823	.0727	.1272	.0568
M-Winsor (PRESS)	-.0005	-.0041	-.0755	.0745	.1305	.0577
Andrews (c = 1.5)	-.0009	.0001	.0276	.0645	.0740	.0911
Huber's (c = 1.5)	-.0005	.0004	.0276	.0686	.0812	.0911
Adiche	-.0052	.0006	.5054	.0682	.0783	.8353
			n=40			
Classical	-.0052	-.0006	.5603	.0721	.0805	.6428
MML (RES)	-.0028	.0052	.0312	.0352	.0692	.0273
MML (PRESS)	-.0029	.0057	.0339	.0355	.0706	.0287
Winsor (RES)	-.0035	.0038	-.0460	.0358	.0760	.0242
Winsor (PRESS)	-.0028	.0057	-.0439	.0360	.0786	.0246
M-Winsor (RES)	-.0034	.0037	-.0782	.0350	.0690	.0262
M-Winsor (PRESS)	-.0027	.0053	-.0761	.0351	.0712	.0264
Andrews (c = 1.5)	-.0031	.0031	.0400	.0299	.0322	.0424
Huber's (c = 1.5)	-.0033	.0024	.0400	.0323	.0351	.0424
Adiche	-.0042	.0027	.5744	.0325	.0351	.6741

Table 10.12: Bias and Mean square error of the estimators under mixture-normal model (with $p = 0.8$ and $\sigma = 4$) for the error

	Bias			MSError		
	α	β	σ	α	β	σ
			n=10			
Classical	-.0205	-.0119	.4292	.2835	.3441	.6619
MML (RES)	-.0088	-.0026	.0503	.2076	.3513	.1836
MML (PRESS)	-.0120	-.0077	.0863	.2322	.4623	.2207
Winsor (RES)	-.0096	.0052	-.0159	.2085	.3472	.1531
Winsor (PRESS)	-.0087	-.0036	-.0047	.2228	.4153	.1644
M-Winsor (RES)	-.0096	.0048	-.0379	.2059	.3373	.1461
M-Winsor (PRESS)	-.0089	-.0036	-.0261	.2188	.4007	.1561
Andrews (c = 1.5)	-.0093	-.0028	.0219	.1879	.2477	.1780
Huber's (c = 1.5)	-.0133	-.0061	.0219	.1883	.2531	.1780
Adiche	-.0123	-.0042	.4642	.1873	.2505	.7442
			n=20			
Classical	.0004	-.0062	.4859	.1355	.1482	.5260
MML (RES)	.0016	-.0037	.0862	.0888	.1447	.0899
MML (PRESS)	.0011	-.0088	.0997	.0903	.1540	.1020
Winsor (RES)	.0034	.0020	.0076	.0891	.1482	.0699
Winsor (PRESS)	.0015	-.0040	.0149	.0923	.1552	.0730
M-Winsor (RES)	.0031	.0018	-.0232	.0871	.1394	.0648
M-Winsor (PRESS)	.0014	-.0038	-.0160	.0898	.1458	.0670
Andrews (c = 1.5)	.0007	.0021	.0665	.0755	.0891	.0961
Huber's (c = 1.5)	.0009	-.0010	.0665	.0791	.0924	.0961
Adiche	-.0042	.0001	.5050	.0782	.0900	.5630
			n=40			
Classical	-.0051	-.0008	.5307	.0653	.0725	.4363
MML (RES)	-.0032	.0054	.0920	.0397	.0774	.0429
MML (PRESS)	-.0030	.0058	.0955	.0408	.0795	.0448
Winsor (RES)	-.0028	.0073	.0106	.0409	.0848	.0283
Winsor (PRESS)	-.0034	.0046	.0130	.0412	.0868	.0294
M-Winsor (RES)	-.0029	.0067	-.0241	.0399	.0764	.0261
M-Winsor (PRESS)	-.0034	.0044	-.0217	.0401	.0782	.0269
Andrews (c = 1.5)	-.0019	.0020	.0816	.0343	.0370	.0499
Huber's (c = 1.5)	-.0034	.0017	.0816	.0364	.0401	.0499
Adiche	-.0043	.0023	.5413	.0366	.0400	.4531

Table 10.13: Bias and Mean square error of the estimators under the single-outlier model (with $\sigma = 4$) for the error

	Bias			MSEror		
	α	β	σ	α	β	σ
			n=10			
Classical	-.0075	.0249	.4664	.2934	.3729	.6135
MML (RES)	-.0107	.0258	.0269	.1743	.3113	.1161
MML (PRESS)	-.0098	.0275	.0608	.1894	.4072	.1444
Winsor (RES)	-.0122	.0198	-.0388	.1744	.3097	.0978
Winsor (PRESS)	-.0091	.0325	-.0273	.1851	.3594	.1041
M-Winsor (RES)	-.0120	.0199	-.0601	.1728	.3020	.0949
M-Winsor (PRESS)	-.0091	.0319	-.0479	.1822	.3479	.1001
Andrews (c = 1.5)	-.0092	.0199	.0240	.1743	.2542	.1651
Huber's (c = 1.5)	-.0080	.0230	.0240	.1781	.2604	.1651
Adiche	-.0101	.0248	.5074	.1790	.2492	.7094
			n=20			
Classical	-.0053	-.0044	.2562	.0918	.1028	.2080
MML (RES)	-.0043	.0002	.0017	.0680	.1246	.0463
MML (PRESS)	-.0037	.0034	.0120	.0694	.1334	.0492
Winsor (RES)	-.0042	.0011	-.0707	.0686	.1296	.0437
Winsor (PRESS)	-.0051	.0019	-.0660	.0702	.1385	.0443
M-Winsor (RES)	-.0042	.0008	-.0989	.0677	.1220	.0457
M-Winsor (PRESS)	-.0050	.0017	-.0942	.0690	.1301	.0458
Andrews (c = 1.5)	-.0043	-.0016	-.0072	.0628	.0707	.0703
Huber's (c = 1.5)	-.0042	-.0030	-.0072	.0643	.0731	.0703
Adiche	-.0087	-.0021	.2697	.0654	.0748	.2262
			n=40			
Classical	-.0060	.0003	.1499	.0349	.0363	.0758
MML (RES)	-.0040	.0008	-.0118	.0300	.0650	.0211
MML (PRESS)	-.0042	-.0007	-.0099	.0303	.0667	.0217
Winsor (RES)	-.0046	.0036	-.0856	.0302	.0732	.0246
Winsor (PRESS)	-.0037	.0008	-.0842	.0307	.0732	.0248
M-Winsor (RES)	-.0045	.0033	-.1165	.0297	.0659	.0294
M-Winsor (PRESS)	-.0037	.0007	-.1152	.0301	.0659	.0294
Andrews (c = 1.5)	-.0038	.0014	-.0046	.0279	.0297	.0343
Huber's (c = 1.5)	-.0046	.0011	-.0046	.0282	.0300	.0343
Adiche	-.0055	.0016	.1544	.0286	.0307	.0794

Table 10.14: Bias and Mean square error of the estimators under the single-outlier model (with $\sigma = 6$) for the error

	Bias			MSError		
	α	β	σ	α	β	σ
			n=10			
Classical	-.0089	.0315	.8933	.5313	.6783	1.8096
MML (RES)	-.0083	.0265	.0516	.1787	.3065	.1232
MML (PRESS)	-.0036	.0202	.0907	.1957	.4329	.1693
Winsor (RES)	-.0086	.0245	-.0138	.1810	.3220	.1014
Winsor (PRESS)	-.0055	.0280	.0022	.1908	.3635	.1116
M-Winsor (RES)	-.0086	.0243	-.0349	.1793	.3146	.0971
M-Winsor (PRESS)	-.0055	.0275	-.0183	.1877	.3522	.1059
Andrews (c = 1.5)	-.0097	.0265	.1112	.1900	.3175	.2452
Huber's (c = 1.5)	-.0087	.0284	.1112	.2083	.3487	.2452
Adiche	-.0112	.0261	.9742	.1931	.2849	2.1440
			n=20			
Classical	-.0054	-.0052	.5161	.1433	.1616	.6565
MML (RES)	-.0025	.0011	.0163	.0698	.1267	.0481
MML (PRESS)	-.0036	.0053	.0274	.0719	.1351	.0521
Winsor (RES)	-.0027	-.0019	-.0582	.0706	.1304	.0435
Winsor (PRESS)	-.0030	.0042	-.0509	.0722	.1387	.0443
M-Winsor (RES)	-.0028	-.0018	-.0865	.0695	.1233	.0446
M-Winsor (PRESS)	-.0030	.0039	-.0792	.0709	.1310	.0448
Andrews (c = 1.5)	-.0040	-.0024	.0206	.0625	.0708	.0787
Huber's (c = 1.5)	-.0043	-.0035	.0206	.0661	.0769	.0787
Adiche	-.0087	-.0021	.5428	.0671	.0770	.7231
			n=40			
Classical	-.0070	-.0006	.3121	.0476	.0493	.2520
MML (RES)	-.0043	-.0004	-.0066	.0305	.0654	.0217
MML (PRESS)	-.0044	-.0046	-.0040	.0307	.0680	.0223
Winsor (RES)	-.0050	.0024	-.0803	.0309	.0727	.0240
Winsor (PRESS)	-.0042	-.0028	-.0785	.0312	.0733	.0242
M-Winsor (RES)	-.0049	.0021	-.1114	.0303	.0658	.0284
M-Winsor (PRESS)	-.0042	-.0026	-.1096	.0306	.0663	.0285
Andrews (c = 1.5)	-.0037	.0019	.0038	.0279	.0296	.0351
Huber's (c = 1.5)	-.0047	.0011	.0038	.0285	.0305	.0351
Adiche	-.0055	.0016	.3208	.0289	.0311	.2655

Table 10.15: Bias and Mean square error of the estimators under the two-outlier model (with $\sigma = 4$) for the error

	Bias			MSError		
	α	β	σ	α	β	σ
			n=10			
Classical	-.0080	.0261	.8633	.4649	.5723	1.3206
MML (RES)	-.0099	.0188	.2687	.2983	.5153	.3537
MML (PRESS)	-.0054	.0264	.3226	.3288	.6826	.4529
Winsor (RES)	-.0072	.0124	.1871	.3056	.5239	.2714
Winsor (PRESS)	-.0069	.0289	.2025	.3095	.6044	.2956
M-Winsor (RES)	-.0072	.0128	.1601	.3000	.5086	.2484
M-Winsor (PRESS)	-.0068	.0285	.1763	.3025	.5824	.2708
Andrews (c = 1.5)	-.0107	.0174	.2048	.2564	.4004	.2709
Huber's (c = 1.5)	-.0099	.0203	.2048	.2646	.4000	.2709
Adiche	-.0110	.0255	.9181	.2531	.3769	1.4883
			n=20			
Classical	-.0068	-.0021	.5008	.1325	.1453	.4652
MML (RES)	-.0041	.0020	.0705	.0796	.1379	.0606
MML (PRESS)	-.0033	-.0023	.0832	.0803	.1443	.0664
Winsor (RES)	-.0040	.0004	-.0066	.0802	.1437	.0459
Winsor (PRESS)	-.0035	-.0013	-.0009	.0821	.1511	.0480
M-Winsor (RES)	-.0040	.0003	-.0367	.0788	.1350	.0437
M-Winsor (PRESS)	-.0036	-.0011	-.0308	.0804	.1417	.0451
Andrews (c = 1.5)	-.0033	-.0056	.0595	.0722	.0834	.0836
Huber's (c = 1.5)	-.0037	-.0048	.0595	.0753	.0879	.0836
Adiche	-.0090	-.0039	.5217	.0756	.0880	.5031
			n=40			
Classical	-.0069	-.0013	.2895	.0446	.0451	.1653
MML (RES)	-.0041	.0040	.0202	.0322	.0683	.0230
MML (PRESS)	-.0047	-.0024	.0225	.0326	.0702	.0235
Winsor (RES)	-.0040	.0050	-.0550	.0328	.0755	.0218
Winsor (PRESS)	-.0049	-.0020	-.0546	.0333	.0765	.0219
M-Winsor (RES)	-.0040	.0046	-.0872	.0322	.0680	.0247
M-Winsor (PRESS)	-.0048	-.0018	-.0866	.0326	.0689	.0247
Andrews (c = 1.5)	-.0030	.0020	.0223	.0297	.0315	.0365
Huber's (c = 1.5)	-.0043	.0009	.0223	.0303	.0320	.0365
Adiche	-.0051	.0013	.2965	.0308	.0328	.1728

Table 10.16: Bias and Mean square error of the estimators under the two-outlier model (with $\sigma = 6$) for the error

	Bias			MSError		
	α	β	σ	α	β	σ
			n=10			
Classical	-.0097	.0335	1.5844	.9295	1.1426	3.9300
MML (RES)	-.0056	.0194	.4699	.4506	.7625	.8489
MML (PRESS)	-.0042	.0184	.5415	.5041	1.0221	1.0791
Winsor (RES)	-.0089	.0112	.3802	.4765	.8187	.6792
Winsor (PRESS)	-.0051	.0224	.4004	.4867	.8836	.7554
M-Winsor (RES)	-.0090	.0115	.3488	.4643	.7927	.6245
M-Winsor (PRESS)	-.0052	.0222	.3697	.4716	.8502	.6953
Andrews (c = 1.5)	-.0137	.0257	.3922	.3358	.6128	.5296
Huber's (c = 1.5)	-.0125	.0255	.3922	.3714	.6428	.5296
Adiche	-.0103	.0294	1.6949	.3157	.5257	4.5054
			n=20			
Classical	-.0079	-.0013	.9638	.2382	.2615	1.5023
MML (RES)	-.0024	.0006	.1032	.0831	.1371	.0725
MML (PRESS)	-.0024	-.0082	.1167	.0848	.1470	.0798
Winsor (RES)	-.0026	.0003	.0212	.0842	.1429	.0515
Winsor (PRESS)	-.0024	-.0068	.0329	.0854	.1484	.0551
M-Winsor (RES)	-.0028	.0001	-.0091	.0826	.1349	.0470
M-Winsor (PRESS)	-.0026	-.0062	.0023	.0836	.1399	.0495
Andrews (c = 1.5)	-.0023	-.0043	.1181	.0729	.0854	.1147
Huber's (c = 1.5)	-.0033	-.0050	.1181	.0812	.0998	.1147
Adiche	-.0091	-.0038	1.0070	.0801	.0938	1.6413
			n=40			
Classical	-.0085	-.0033	.5812	.0701	.0705	.5731
MML (RES)	-.0046	-.0025	.0331	.0331	.0669	.0244
MML (PRESS)	-.0044	-.0044	.0359	.0335	.0696	.0257
Winsor (RES)	-.0043	.0024	-.0439	.0336	.0737	.0214
Winsor (PRESS)	-.0053	-.0011	-.0420	.0342	.0758	.0216
M-Winsor (RES)	-.0043	.0022	-.0764	.0329	.0668	.0236
M-Winsor (PRESS)	-.0051	-.0011	-.0744	.0334	.0686	.0235
Andrews (c = 1.5)	-.0030	.0027	.0401	.0296	.0314	.0397
Huber's (c = 1.5)	-.0044	.0011	.0401	.0312	.0331	.0397
Adiche	-.0051	.0013	.5957	.0316	.0335	.6020

Table 10.17: Bias and Mean square error of the estimators under normal model for the error ($n = 10$)

				G=-.10		
Classical	-0.0839	0.0157	-0.0203	0.1231	0.1525	0.0610
MML (RES)	-0.0684	0.0279	-0.1009	0.1444	0.2921	0.0960
MML (PRESS)	-0.0689	0.0274	-0.0723	0.1539	0.3415	0.1043
Winsor (RES)	-0.0723	0.0278	-0.1566	0.1399	0.2737	0.0981
Winsor (PRES)	-0.0681	0.0261	-0.1487	0.1506	0.3297	0.0997
M-Winsor (RES)	-0.0732	0.0274	-0.1756	0.1389	0.2665	0.1007
M-Winsor (PRESS)	-0.0691	0.0255	-0.1673	0.1487	0.3191	0.1014
Andrews (c = 1.5)	-0.0873	0.0169	-0.1240	0.1363	0.1623	0.1269
Hubers (c = 1.5)	-0.0865	0.0165	-0.1240	0.1303	0.1574	0.1269
Adiche	-0.0872	0.0182	-0.0086	0.1361	0.1682	0.0623
				G=-.05		
Classical	-0.0447	0.0153	-0.0249	0.1167	0.1476	0.0604
MML (RES)	-0.0354	0.0236	-0.1043	0.1384	0.2842	0.0956
MML (PRESS)	-0.0383	0.0267	-0.0752	0.1488	0.3324	0.1049
Winsor (RES)	-0.0389	0.0231	-0.1603	0.1349	0.2689	0.0977
Winsor (PRES)	-0.0346	0.0289	-0.1513	0.1458	0.3220	0.1000
M-Winsor (RES)	-0.0393	0.0228	-0.1793	0.1338	0.2616	0.1005
M-Winsor (PRESS)	-0.0352	0.0281	-0.1699	0.1438	0.3116	0.1019
Andrews (c = 1.5)	-0.0465	0.0162	-0.1296	0.1286	0.1569	0.1264
Hubers (c = 1.5)	-0.0459	0.0165	-0.1296	0.1233	0.1522	0.1264
Adiche	-0.0466	0.0174	-0.0133	0.1289	0.1630	0.0616
				G=0.05		
Classical	0.0338	0.0147	-0.0243	0.1165	0.1482	0.0602
MML (RES)	0.0296	0.0229	-0.1045	0.1391	0.2899	0.0958
MML (PRESS)	0.0264	0.0244	-0.0783	0.1496	0.3337	0.1030
Winsor (RES)	0.0278	0.0225	-0.1608	0.1358	0.2743	0.0973
Winsor (PRES)	0.0254	0.0243	-0.1518	0.1485	0.3251	0.1008
M-Winsor (RES)	0.0283	0.0220	-0.1798	0.1347	0.2669	0.1001
M-Winsor (PRESS)	0.0261	0.0239	-0.1703	0.1463	0.3144	0.1025
Andrews (c = 1.5)	0.0351	0.0151	-0.1309	0.1282	0.1579	0.1250
Hubers (c = 1.5)	0.0353	0.0159	-0.1309	0.1231	0.1531	0.1250
Adiche	0.0343	0.0168	-0.0128	0.1289	0.1640	0.0613
				G=0.10		
Classical	0.0730	0.0144	-0.0191	0.1227	0.1538	0.0606
MML (RES)	0.0617	0.0239	-0.1007	0.1445	0.2985	0.0963
MML (PRESS)	0.0588	0.0246	-0.0736	0.1557	0.3419	0.1033
Winsor (RES)	0.0613	0.0238	-0.1574	0.1412	0.2877	0.0969
Winsor (PRES)	0.0584	0.0253	-0.1492	0.1541	0.3354	0.1010
M-Winsor (RES)	0.0623	0.0234	-0.1765	0.1401	0.2798	0.0996
M-Winsor (PRESS)	0.0594	0.0246	-0.1677	0.1520	0.3246	0.1026
Andrews (c = 1.5)	0.0752	0.0146	-0.1252	0.1351	0.1641	0.1243
Hubers (c = 1.5)	0.0757	0.0157	-0.1252	0.1297	0.1593	0.1243
Adiche	0.0748	0.0166	-0.0078	0.1360	0.1702	0.0619

Table 10.18: Bias and Mean square error of the estimators under normal model for the error ($n = 20$)

			G=-.10			
Classical	-0.0946	-0.0013	-0.0076	0.0630	0.0626	0.0284
MML (RES)	-0.0877	-0.0051	-0.0528	0.0679	0.1184	0.0435
MML (PRESS)	-0.0891	-0.0038	-0.0454	0.0696	0.1241	0.0455
Winsor (RES)	-0.0857	-0.0047	-0.1223	0.0689	0.1262	0.0488
Winsor (PRES)	-0.0874	-0.0019	-0.1171	0.0696	0.1283	0.0493
M-Winsor (RES)	-0.0869	-0.0043	-0.1488	0.0683	0.1187	0.0537
M-Winsor (PRESS)	-0.0883	-0.0017	-0.1438	0.0688	0.1203	0.0537
Andrews (c = 1.5)	-0.0945	-0.0019	-0.0593	0.0667	0.0667	0.0651
Hubers (c = 1.5)	-0.0943	-0.0016	-0.0593	0.0656	0.0651	0.0651
Adiche	-0.0985	-0.0022	-0.0035	0.0669	0.0680	0.0286
			G=-.05			
Classical	-0.0499	-0.0023	-0.0142	0.0556	0.0596	0.0282
MML (RES)	-0.0459	-0.0031	-0.0582	0.0617	0.1141	0.0436
MML (PRESS)	-0.0477	-0.0045	-0.0513	0.0631	0.1213	0.0457
Winsor (RES)	-0.0441	-0.0039	-0.1282	0.0619	0.1214	0.0500
Winsor (PRES)	-0.0469	-0.0054	-0.1236	0.0635	0.1259	0.0506
M-Winsor (RES)	-0.0449	-0.0037	-0.1545	0.0613	0.1139	0.0552
M-Winsor (PRESS)	-0.0473	-0.0049	-0.1501	0.0625	0.1180	0.0552
Andrews (c = 1.5)	-0.0494	-0.0031	-0.0664	0.0592	0.0632	0.0651
Hubers (c = 1.5)	-0.0492	-0.0028	-0.0664	0.0582	0.0619	0.0651
Adiche	-0.0530	-0.0032	-0.0102	0.0593	0.0648	0.0284
			G=0.05			
Classical	0.0397	-0.0041	-0.0152	0.0545	0.0596	0.0283
MML (RES)	0.0373	-0.0054	-0.0607	0.0604	0.1132	0.0442
MML (PRESS)	0.0370	-0.0041	-0.0537	0.0629	0.1200	0.0456
Winsor (RES)	0.0371	-0.0048	-0.1290	0.0601	0.1182	0.0505
Winsor (PRES)	0.0368	-0.0036	-0.1256	0.0634	0.1272	0.0508
M-Winsor (RES)	0.0376	-0.0047	-0.1553	0.0594	0.1111	0.0557
M-Winsor (PRESS)	0.0372	-0.0035	-0.1518	0.0623	0.1192	0.0555
Andrews (c = 1.5)	0.0416	-0.0054	-0.0660	0.0582	0.0629	0.0660
Hubers (c = 1.5)	0.0417	-0.0052	-0.0660	0.0572	0.0620	0.0660
Adiche	0.0375	-0.0044	-0.0112	0.0578	0.0645	0.0285
			G=0.10			
Classical	0.0844	-0.0051	-0.0096	0.0608	0.0626	0.0286
MML (RES)	0.0784	-0.0045	-0.0575	0.0667	0.1182	0.0446
MML (PRESS)	0.0794	-0.0030	-0.0493	0.0685	0.1269	0.0458
Winsor (RES)	0.0766	-0.0049	-0.1247	0.0660	0.1238	0.0502
Winsor (PRES)	0.0785	-0.0042	-0.1213	0.0684	0.1302	0.0499
M-Winsor (RES)	0.0777	-0.0048	-0.1512	0.0654	0.1166	0.0551
M-Winsor (PRESS)	0.0794	-0.0040	-0.1477	0.0675	0.1222	0.0544
Andrews (c = 1.5)	0.0866	-0.0066	-0.0591	0.0648	0.0664	0.0652
Hubers (c = 1.5)	0.0867	-0.0060	-0.0591	0.0637	0.0652	0.0652
Adiche	0.0829	-0.0045	-0.0055	0.0641	0.0673	0.0288

Table 10.19: Bias and Mean square error of the estimators under mixture-normal model (with $p = 0.9$ and $\sigma = 4$) for the error ($n = 10$)

			G=-.10			
Classical	-0.0955	-0.0037	0.2039	0.2101	0.2333	0.3361
MML (RES)	-0.0749	0.0033	-0.0332	0.1738	0.3055	0.1234
MML (PRESS)	-0.0785	0.0008	-0.0040	0.1857	0.3723	0.1447
Winsor (RES)	-0.0784	0.0084	-0.0942	0.1732	0.2903	0.1126
Winsor (PRES)	-0.0738	0.0047	-0.0853	0.1787	0.3458	0.1186
M-Winsor (RES)	-0.0794	0.0082	-0.1145	0.1716	0.2823	0.1114
M-Winsor (PRESS)	-0.0749	0.0044	-0.1051	0.1764	0.3343	0.1164
Andrews (c = 1.5)	-0.0930	0.0025	-0.0608	0.1629	0.1912	0.1446
Hubers (c = 1.5)	-0.0947	-0.0004	-0.0608	0.1630	0.1925	0.1446
Adiche	-0.0940	-0.0012	0.2281	0.1659	0.1972	0.3782
			G=-.05			
Classical	-0.0553	-0.0039	0.1996	0.2022	0.2280	0.3340
MML (RES)	-0.0407	-0.0002	-0.0366	0.1665	0.2948	0.1219
MML (PRESS)	-0.0467	0.0005	-0.0063	0.1775	0.3610	0.1431
Winsor (RES)	-0.0446	0.0078	-0.0967	0.1667	0.2806	0.1124
Winsor (PRES)	-0.0412	0.0047	-0.0881	0.1733	0.3372	0.1179
M-Winsor (RES)	-0.0450	0.0076	-0.1171	0.1650	0.2726	0.1113
M-Winsor (PRESS)	-0.0418	0.0045	-0.1079	0.1709	0.3258	0.1159
Andrews (c = 1.5)	-0.0515	0.0026	-0.0643	0.1537	0.1849	0.1420
Hubers (c = 1.5)	-0.0536	-0.0006	-0.0643	0.1544	0.1866	0.1420
Adiche	-0.0529	-0.0009	0.2235	0.1573	0.1916	0.3760
			G=0.05			
Classical	0.0251	-0.0045	0.1996	0.1995	0.2276	0.3332
MML (RES)	0.0260	-0.0014	-0.0364	0.1661	0.2983	0.1217
MML (PRESS)	0.0181	-0.0004	-0.0092	0.1758	0.3624	0.1384
Winsor (RES)	0.0258	0.0041	-0.0963	0.1638	0.2761	0.1116
Winsor (PRES)	0.0208	0.0021	-0.0893	0.1736	0.3382	0.1171
M-Winsor (RES)	0.0262	0.0039	-0.1167	0.1622	0.2683	0.1105
M-Winsor (PRESS)	0.0213	0.0021	-0.1090	0.1709	0.3265	0.1151
Andrews (c = 1.5)	0.0303	0.0017	-0.0633	0.1518	0.1830	0.1396
Hubers (c = 1.5)	0.0287	-0.0010	-0.0633	0.1519	0.1859	0.1396
Adiche	0.0278	-0.0002	0.2235	0.1538	0.1910	0.3753
			G=0.10			
Classical	0.0653	-0.0048	0.2039	0.2049	0.2326	0.3345
MML (RES)	0.0589	-0.0007	-0.0327	0.1697	0.3024	0.1215
MML (PRESS)	0.0506	-0.0010	-0.0047	0.1803	0.3676	0.1401
Winsor (RES)	0.0608	0.0039	-0.0931	0.1688	0.2825	0.1117
Winsor (PRES)	0.0535	0.0023	-0.0868	0.1782	0.3479	0.1170
M-Winsor (RES)	0.0615	0.0038	-0.1135	0.1673	0.2747	0.1104
M-Winsor (PRESS)	0.0545	0.0021	-0.1065	0.1756	0.3359	0.1149
Andrews (c = 1.5)	0.0716	0.0013	-0.0571	0.1580	0.1881	0.1401
Hubers (c = 1.5)	0.0698	-0.0012	-0.0571	0.1579	0.1913	0.1401
Adiche	0.0683	-0.0012	0.2278	0.1595	0.1960	0.3770

Table 10.20: Bias and Mean square error of the estimators under mixture-normal model (with $p = 0.9$ and $\sigma = 4$) for the error ($n = 20$)

			G=-.10			
Classical	-0.0904	-0.0006	0.2507	0.1019	0.1105	0.2524
MML (RES)	-0.0837	-0.0031	0.0115	0.0787	0.1323	0.0547
MML (PRESS)	-0.0852	-0.0042	0.0201	0.0803	0.1380	0.0595
Winsor (RES)	-0.0806	-0.0055	-0.0618	0.0795	0.1398	0.0489
Winsor (PRES)	-0.0825	-0.0047	-0.0568	0.0802	0.1440	0.0503
M-Winsor (RES)	-0.0819	-0.0052	-0.0902	0.0785	0.1320	0.0500
M-Winsor (PRESS)	-0.0834	-0.0043	-0.0853	0.0790	0.1355	0.0508
Andrews (c = 1.5)	-0.0895	0.0017	0.0052	0.0738	0.0786	0.0756
Hubers (c = 1.5)	-0.0901	0.0010	0.0052	0.0754	0.0806	0.0756
Adiche	-0.0949	0.0008	0.2631	0.0769	0.0814	0.2710
			G=-.05			
Classical	-0.0459	-0.0010	0.2458	0.0950	0.1070	0.2500
MML (RES)	-0.0422	-0.0025	0.0059	0.0725	0.1278	0.0536
MML (PRESS)	-0.0431	-0.0025	0.0150	0.0740	0.1340	0.0579
Winsor (RES)	-0.0400	-0.0051	-0.0673	0.0725	0.1328	0.0490
Winsor (PRES)	-0.0430	-0.0027	-0.0612	0.0742	0.1387	0.0511
M-Winsor (RES)	-0.0408	-0.0047	-0.0956	0.0715	0.1252	0.0504
M-Winsor (PRESS)	-0.0434	-0.0025	-0.0896	0.0729	0.1304	0.0518
Andrews (c = 1.5)	-0.0450	0.0017	0.0001	0.0666	0.0749	0.0754
Hubers (c = 1.5)	-0.0454	0.0007	0.0001	0.0683	0.0771	0.0754
Adiche	-0.0502	0.0006	0.2583	0.0695	0.0774	0.2686
			G=0.05			
Classical	0.0431	-0.0018	0.2463	0.0947	0.1062	0.2494
MML (RES)	0.0415	-0.0022	0.0043	0.0714	0.1278	0.0535
MML (PRESS)	0.0425	-0.0007	0.0152	0.0738	0.1356	0.0587
Winsor (RES)	0.0411	0.0014	-0.0685	0.0722	0.1340	0.0493
Winsor (PRES)	0.0415	-0.0015	-0.0635	0.0741	0.1414	0.0504
M-Winsor (RES)	0.0415	0.0012	-0.0966	0.0712	0.1261	0.0507
M-Winsor (PRESS)	0.0419	-0.0013	-0.0915	0.0728	0.1326	0.0513
Andrews (c = 1.5)	0.0443	0.0011	0.0027	0.0667	0.0748	0.0775
Hubers (c = 1.5)	0.0444	-0.0003	0.0027	0.0681	0.0768	0.0775
Adiche	0.0397	-0.0002	0.2589	0.0681	0.0772	0.2680
			G=0.10			
Classical	0.0876	-0.0022	0.2517	0.1014	0.1090	0.2512
MML (RES)	0.0820	-0.0020	0.0090	0.0770	0.1303	0.0540
MML (PRESS)	0.0858	-0.0026	0.0201	0.0804	0.1413	0.0593
Winsor (RES)	0.0802	-0.0007	-0.0637	0.0780	0.1414	0.0496
Winsor (PRES)	0.0831	-0.0010	-0.0579	0.0799	0.1435	0.0508
M-Winsor (RES)	0.0813	-0.0007	-0.0920	0.0771	0.1332	0.0507
M-Winsor (PRESS)	0.0839	-0.0011	-0.0861	0.0787	0.1348	0.0513
Andrews (c = 1.5)	0.0887	-0.0001	0.0095	0.0734	0.0785	0.0789
Hubers (c = 1.5)	0.0890	-0.0010	0.0095	0.0748	0.0803	0.0789
Adiche	0.0845	-0.0013	0.2641	0.0746	0.0805	0.2698

Table 10.21: Bias and Mean square error of the estimators under mixture-normal model (with $p = 0.8$ and $\sigma = 6$) for the error ($n = 10$)

				G=-.10		
Classical	-0.1071	-0.0193	0.8124	0.5204	0.6366	1.9324
MML (RES)	-0.0738	0.0046	0.1323	0.2668	0.4381	0.3563
MML (PRESS)	-0.0783	-0.0023	0.1762	0.3077	0.6035	0.4373
Winsor (RES)	-0.0782	0.0080	0.0611	0.2750	0.4602	0.2949
Winsor (PRES)	-0.0708	0.0017	0.0798	0.2865	0.5103	0.3221
M-Winsor (RES)	-0.0791	0.0075	0.0377	0.2707	0.4467	0.2764
M-Winsor (PRESS)	-0.0719	0.0013	0.0566	0.2802	0.4921	0.3012
Andrews (c = 1.5)	-0.0901	-0.0107	0.1228	0.2365	0.3509	0.3011
Hubers (c = 1.5)	-0.0955	-0.0113	0.1228	0.2472	0.3736	0.3011
Adiche	-0.0925	-0.0083	0.8782	0.2299	0.3218	2.2163
				G=-.05		
Classical	-0.0669	-0.0196	0.8092	0.5120	0.6313	1.9288
MML (RES)	-0.0401	0.0048	0.1290	0.2593	0.4254	0.3543
MML (PRESS)	-0.0476	-0.0063	0.1740	0.2996	0.5990	0.4373
Winsor (RES)	-0.0443	0.0099	0.0579	0.2690	0.4486	0.2953
Winsor (PRES)	-0.0377	0.0026	0.0746	0.2790	0.4828	0.3158
M-Winsor (RES)	-0.0446	0.0092	0.0345	0.2646	0.4356	0.2769
M-Winsor (PRESS)	-0.0385	0.0022	0.0516	0.2727	0.4659	0.2954
Andrews (c = 1.5)	-0.0492	-0.0108	0.1203	0.2276	0.3440	0.2998
Hubers (c = 1.5)	-0.0546	-0.0116	0.1203	0.2387	0.3672	0.2998
Adiche	-0.0527	-0.0082	0.8753	0.2205	0.3156	2.2131
				G=0.05		
Classical	0.0135	-0.0201	0.8095	0.5086	0.6310	1.9288
MML (RES)	0.0263	-0.0012	0.1296	0.2597	0.4309	0.3571
MML (PRESS)	0.0198	-0.0124	0.1742	0.3107	0.6154	0.4422
Winsor (RES)	0.0254	0.0083	0.0577	0.2673	0.4428	0.2939
Winsor (PRES)	0.0243	0.0004	0.0724	0.2816	0.4949	0.3154
M-Winsor (RES)	0.0258	0.0078	0.0342	0.2628	0.4299	0.2756
M-Winsor (PRESS)	0.0244	0.0002	0.0495	0.2749	0.4772	0.2952
Andrews (c = 1.5)	0.0334	-0.0093	0.1226	0.2267	0.3409	0.2990
Hubers (c = 1.5)	0.0268	-0.0121	0.1226	0.2365	0.3664	0.2990
Adiche	0.0264	-0.0066	0.8754	0.2164	0.3153	2.2120
				G=0.10		
Classical	0.0538	-0.0204	0.8130	0.5136	0.6360	1.9324
MML (RES)	0.0580	-0.0046	0.1350	0.2665	0.4357	0.3589
MML (PRESS)	0.0514	-0.0106	0.1748	0.3170	0.6154	0.4394
Winsor (RES)	0.0595	0.0085	0.0618	0.2735	0.4486	0.2928
Winsor (PRES)	0.0564	-0.0043	0.0751	0.2844	0.4998	0.3154
M-Winsor (RES)	0.0603	0.0079	0.0384	0.2690	0.4357	0.2745
M-Winsor (PRESS)	0.0570	-0.0045	0.0522	0.2779	0.4824	0.2951
Andrews (c = 1.5)	0.0736	-0.0091	0.1287	0.2324	0.3451	0.2999
Hubers (c = 1.5)	0.0676	-0.0123	0.1287	0.2425	0.3718	0.2999
Adiche	0.0664	-0.0072	0.8787	0.2220	0.3215	2.2148

Table 10.22: Bias and Mean square error of the estimators under mixture-normal model (with $p = 0.8$ and $\sigma = 6$) for the error ($n = 20$)

				G=-.10		
Classical	-0.0873	-0.0092	0.9280	0.2503	0.2708	1.6632
MML (RES)	-0.0833	-0.0010	0.1497	0.1152	0.1562	0.1681
MML (PRESS)	-0.0846	-0.0032	0.1651	0.1183	0.1694	0.1858
Winsor (RES)	-0.0820	-0.0008	0.0680	0.1190	0.1678	0.1275
Winsor (PRES)	-0.0843	-0.0027	0.0760	0.1216	0.1752	0.1360
M-Winsor (RES)	-0.0829	-0.0008	0.0354	0.1153	0.1580	0.1138
M-Winsor (PRESS)	-0.0849	-0.0028	0.0434	0.1174	0.1648	0.1208
Andrews (c = 1.5)	-0.0893	0.0022	0.1314	0.0859	0.0974	0.1374
Hubers (c = 1.5)	-0.0878	0.0000	0.1314	0.0965	0.1115	0.1374
Adiche	-0.0930	0.0007	0.9668	0.0943	0.1040	1.7982
				G=-.05		
Classical	-0.0428	-0.0096	0.9245	0.2439	0.2671	1.6581
MML (RES)	-0.0428	-0.0037	0.1433	0.1089	0.1529	0.1668
MML (PRESS)	-0.0418	-0.0028	0.1589	0.1124	0.1657	0.1846
Winsor (RES)	-0.0407	-0.0018	0.0629	0.1129	0.1621	0.1269
Winsor (PRES)	-0.0435	-0.0070	0.0713	0.1161	0.1688	0.1361
M-Winsor (RES)	-0.0411	-0.0018	0.0305	0.1091	0.1525	0.1135
M-Winsor (PRESS)	-0.0436	-0.0065	0.0388	0.1117	0.1585	0.1212
Andrews (c = 1.5)	-0.0449	0.0024	0.1276	0.0789	0.0937	0.1366
Hubers (c = 1.5)	-0.0434	-0.0004	0.1276	0.0897	0.1081	0.1366
Adiche	-0.0490	0.0009	0.9636	0.0867	0.1000	1.7934
				G=0.05		
Classical	0.0462	-0.0104	0.9250	0.2447	0.2660	1.6577
MML (RES)	0.0397	-0.0012	0.1463	0.1095	0.1541	0.1675
MML (PRESS)	0.0414	-0.0013	0.1633	0.1126	0.1661	0.1907
Winsor (RES)	0.0414	-0.0033	0.0632	0.1129	0.1636	0.1289
Winsor (PRES)	0.0409	-0.0012	0.0721	0.1147	0.1688	0.1344
M-Winsor (RES)	0.0420	-0.0031	0.0307	0.1091	0.1538	0.1152
M-Winsor (PRESS)	0.0416	-0.0013	0.0397	0.1105	0.1581	0.1197
Andrews (c = 1.5)	0.0443	0.0023	0.1271	0.0789	0.0934	0.1381
Hubers (c = 1.5)	0.0457	-0.0011	0.1271	0.0897	0.1081	0.1381
Adiche	0.0400	0.0002	0.9643	0.0857	0.0989	1.7937
				G=0.10		
Classical	0.0907	-0.0108	0.9290	0.2518	0.2686	1.6624
MML (RES)	0.0829	-0.0004	0.1529	0.1153	0.1593	0.1719
MML (PRESS)	0.0842	-0.0012	0.1683	0.1199	0.1763	0.1932
Winsor (RES)	0.0823	0.0020	0.0673	0.1186	0.1735	0.1304
Winsor (PRES)	0.0815	-0.0011	0.0793	0.1217	0.1777	0.1370
M-Winsor (RES)	0.0834	0.0018	0.0349	0.1150	0.1630	0.1163
M-Winsor (PRESS)	0.0826	-0.0012	0.0466	0.1176	0.1666	0.1216
Andrews (c = 1.5)	0.0887	0.0016	0.1316	0.0856	0.0971	0.1400
Hubers (c = 1.5)	0.0900	-0.0014	0.1316	0.0967	0.1118	0.1400
Adiche	0.0845	-0.0006	0.9681	0.0921	0.1029	1.7980

Table 10.23: Bias and Mean square error of the estimators under the single-outlier model (with $\sigma = 4$) for the error ($n = 10$)

				G=-.10		
Classical	-0.0860	0.0255	0.4703	0.3024	0.3792	0.6178
MML (RES)	-0.0738	0.0279	0.0324	0.1823	0.3216	0.1177
MML (PRESS)	-0.0718	0.0291	0.0655	0.1962	0.4228	0.1480
Winsor (RES)	-0.0768	0.0249	-0.0354	0.1829	0.3155	0.0999
Winsor (PRES)	-0.0721	0.0297	-0.0246	0.1925	0.3719	0.1065
M-Winsor (RES)	-0.0775	0.0249	-0.0566	0.1813	0.3078	0.0966
M-Winsor (PRESS)	-0.0730	0.0295	-0.0452	0.1897	0.3601	0.1023
Andrews (c = 1.5)	-0.0896	0.0210	0.0320	0.1837	0.2595	0.1665
Hubers (c = 1.5)	-0.0883	0.0241	0.0320	0.1875	0.2670	0.1665
Adiche	-0.0908	0.0262	0.5116	0.1879	0.2565	0.7146
				G=-.05		
Classical	-0.0467	0.0252	0.4672	0.2957	0.3743	0.6146
MML (RES)	-0.0427	0.0244	0.0276	0.1763	0.3097	0.1164
MML (PRESS)	-0.0403	0.0289	0.0604	0.1914	0.4098	0.1453
Winsor (RES)	-0.0453	0.0236	-0.0377	0.1774	0.3120	0.0989
Winsor (PRES)	-0.0406	0.0344	-0.0281	0.1873	0.3621	0.1049
M-Winsor (RES)	-0.0455	0.0236	-0.0591	0.1757	0.3043	0.0958
M-Winsor (PRESS)	-0.0411	0.0338	-0.0487	0.1845	0.3507	0.1009
Andrews (c = 1.5)	-0.0493	0.0207	0.0264	0.1774	0.2550	0.1653
Hubers (c = 1.5)	-0.0482	0.0235	0.0264	0.1807	0.2618	0.1653
Adiche	-0.0507	0.0257	0.5083	0.1812	0.2504	0.7109
				G=0.05		
Classical	0.0317	0.0246	0.4679	0.2952	0.3749	0.6144
MML (RES)	0.0213	0.0280	0.0283	0.1765	0.3171	0.1160
MML (PRESS)	0.0247	0.0215	0.0629	0.1918	0.4106	0.1414
Winsor (RES)	0.0224	0.0221	-0.0382	0.1761	0.3145	0.0981
Winsor (PRES)	0.0215	0.0311	-0.0254	0.1879	0.3655	0.1048
M-Winsor (RES)	0.0231	0.0221	-0.0594	0.1745	0.3067	0.0951
M-Winsor (PRESS)	0.0221	0.0306	-0.0461	0.1850	0.3539	0.1007
Andrews (c = 1.5)	0.0307	0.0188	0.0267	0.1753	0.2559	0.1646
Hubers (c = 1.5)	0.0321	0.0224	0.0267	0.1798	0.2629	0.1646
Adiche	0.0299	0.0240	0.5089	0.1808	0.2523	0.7104
				G=0.10		
Classical	0.0710	0.0243	0.4718	0.3014	0.3805	0.6175
MML (RES)	0.0546	0.0291	0.0319	0.1825	0.3274	0.1162
MML (PRESS)	0.0552	0.0192	0.0674	0.1963	0.4190	0.1436
Winsor (RES)	0.0558	0.0231	-0.0345	0.1819	0.3287	0.0987
Winsor (PRES)	0.0533	0.0298	-0.0227	0.1921	0.3756	0.1056
M-Winsor (RES)	0.0569	0.0230	-0.0558	0.1804	0.3206	0.0955
M-Winsor (PRESS)	0.0544	0.0294	-0.0434	0.1894	0.3638	0.1013
Andrews (c = 1.5)	0.0705	0.0173	0.0324	0.1809	0.2623	0.1647
Hubers (c = 1.5)	0.0717	0.0217	0.0324	0.1862	0.2694	0.1647
Adiche	0.0702	0.0231	0.5125	0.1874	0.2597	0.7132

Table 10.24: Bias and Mean square error of the estimators under the single-outlier model (with $\sigma = 4$) for the error ($n = 20$)

				G=-.10		
Classical	-0.0948	-0.0025	0.2640	0.1020	0.1069	0.2118
MML (RES)	-0.0865	-0.0026	0.0090	0.0766	0.1307	0.0473
MML (PRESS)	-0.0886	0.0021	0.0185	0.0789	0.1383	0.0506
Winsor (RES)	-0.0864	0.0022	-0.0654	0.0771	0.1370	0.0435
Winsor (PRES)	-0.0861	0.0016	-0.0585	0.0784	0.1445	0.0443
M-Winsor (RES)	-0.0873	0.0021	-0.0935	0.0764	0.1292	0.0451
M-Winsor (PRESS)	-0.0871	0.0014	-0.0868	0.0775	0.1359	0.0453
Andrews (c = 1.5)	-0.0946	0.0005	-0.0006	0.0735	0.0753	0.0703
Hubers (c = 1.5)	-0.0944	-0.0008	-0.0006	0.0748	0.0776	0.0703
Adiche	-0.0984	-0.0011	0.2774	0.0765	0.0792	0.2299
				G=-.05		
Classical	-0.0500	-0.0035	0.2585	0.0946	0.1038	0.2091
MML (RES)	-0.0455	0.0007	0.0033	0.0706	0.1281	0.0468
MML (PRESS)	-0.0459	-0.0001	0.0133	0.0722	0.1345	0.0497
Winsor (RES)	-0.0448	-0.0015	-0.0691	0.0710	0.1318	0.0437
Winsor (PRES)	-0.0459	0.0011	-0.0643	0.0724	0.1400	0.0446
M-Winsor (RES)	-0.0453	-0.0015	-0.0974	0.0701	0.1242	0.0456
M-Winsor (PRESS)	-0.0463	0.0010	-0.0925	0.0714	0.1315	0.0459
Andrews (c = 1.5)	-0.0498	-0.0005	-0.0055	0.0658	0.0717	0.0702
Hubers (c = 1.5)	-0.0495	-0.0019	-0.0055	0.0673	0.0742	0.0702
Adiche	-0.0534	-0.0017	0.2719	0.0685	0.0760	0.2273
				G=0.05		
Classical	0.0395	-0.0053	0.2573	0.0935	0.1037	0.2085
MML (RES)	0.0366	0.0012	0.0028	0.0698	0.1257	0.0457
MML (PRESS)	0.0381	0.0024	0.0118	0.0717	0.1347	0.0495
Winsor (RES)	0.0365	0.0011	-0.0698	0.0698	0.1298	0.0435
Winsor (PRES)	0.0374	0.0039	-0.0639	0.0714	0.1387	0.0446
M-Winsor (RES)	0.0370	0.0009	-0.0981	0.0689	0.1223	0.0455
M-Winsor (PRESS)	0.0379	0.0032	-0.0922	0.0702	0.1304	0.0459
Andrews (c = 1.5)	0.0411	-0.0026	-0.0052	0.0647	0.0717	0.0706
Hubers (c = 1.5)	0.0411	-0.0040	-0.0052	0.0660	0.0741	0.0706
Adiche	0.0369	-0.0032	0.2708	0.0667	0.0758	0.2268
				G=0.10		
Classical	0.0843	-0.0063	0.2617	0.0997	0.1066	0.2107
MML (RES)	0.0790	0.0016	0.0079	0.0757	0.1293	0.0465
MML (PRESS)	0.0787	0.0031	0.0162	0.0769	0.1378	0.0498
Winsor (RES)	0.0766	0.0056	-0.0657	0.0754	0.1341	0.0434
Winsor (PRES)	0.0781	0.0031	-0.0597	0.0777	0.1430	0.0442
M-Winsor (RES)	0.0778	0.0048	-0.0941	0.0746	0.1266	0.0451
M-Winsor (PRESS)	0.0792	0.0024	-0.0880	0.0766	0.1344	0.0454
Andrews (c = 1.5)	0.0860	-0.0037	0.0005	0.0712	0.0751	0.0713
Hubers (c = 1.5)	0.0861	-0.0049	0.0005	0.0725	0.0775	0.0713
Adiche	0.0814	-0.0038	0.2751	0.0727	0.0791	0.2289

Table 10.25: Bias and Mean square error of the estimators under the single-outlier model (with $\sigma = 6$) for the error($n = 10$)

			G=-.10			
Classical	-0.0882	0.0341	1.5864	0.9388	1.1496	3.9359
MML (RES)	-0.0723	0.0251	0.4738	0.4622	0.7593	0.8607
MML (PRESS)	-0.0711	0.0264	0.5462	0.5140	1.0442	1.0814
Winsor (RES)	-0.0762	0.0130	0.3845	0.4801	0.8208	0.6872
Winsor (PRES)	-0.0674	0.0286	0.4026	0.4904	0.9056	0.7419
M-Winsor (RES)	-0.0769	0.0132	0.3531	0.4683	0.7951	0.6320
M-Winsor (PRESS)	-0.0683	0.0283	0.3719	0.4760	0.8725	0.6829
Andrews (c = 1.5)	-0.0932	0.0279	0.3974	0.3474	0.6240	0.5372
Hubers (c = 1.5)	-0.0924	0.0269	0.3974	0.3827	0.6502	0.5372
Adiche	-0.0891	0.0290	1.6967	0.3255	0.5310	4.5089
			G=-.05			
Classical	-0.0490	0.0338	1.5847	0.9320	1.1444	3.9310
MML (RES)	-0.0396	0.0208	0.4707	0.4545	0.7542	0.8532
MML (PRESS)	-0.0369	0.0243	0.5425	0.5059	1.0477	1.0780
Winsor (RES)	-0.0430	0.0140	0.3825	0.4769	0.8104	0.6818
Winsor (PRES)	-0.0375	0.0256	0.3994	0.4845	0.8938	0.7416
M-Winsor (RES)	-0.0434	0.0141	0.3510	0.4649	0.7851	0.6269
M-Winsor (PRESS)	-0.0379	0.0253	0.3688	0.4699	0.8610	0.6826
Andrews (c = 1.5)	-0.0524	0.0274	0.3939	0.3389	0.6153	0.5327
Hubers (c = 1.5)	-0.0526	0.0262	0.3939	0.3746	0.5446	0.5327
Adiche	-0.0499	0.0297	1.6952	0.3187	0.5267	4.5058
			G=0.05			
Classical	0.0295	0.0332	1.5855	0.9312	1.1444	3.9330
MML (RES)	0.0267	0.0166	0.4715	0.4540	0.7648	0.8471
MML (PRESS)	0.0289	0.0181	0.5436	0.5066	1.0189	1.0780
Winsor (RES)	0.0252	0.0126	0.3813	0.4781	0.8150	0.6792
Winsor (PRES)	0.0249	0.0192	0.4023	0.4872	0.8927	0.7554
M-Winsor (RES)	0.0255	0.0128	0.3500	0.4659	0.7894	0.6244
M-Winsor (PRESS)	0.0252	0.0191	0.3716	0.4724	0.8592	0.6952
Andrews (c = 1.5)	0.0273	0.0242	0.3919	0.3377	0.6144	0.5279
Hubers (c = 1.5)	0.0276	0.0248	0.3919	0.3727	0.6445	0.5279
Adiche	0.0288	0.0281	1.6956	0.3182	0.5286	4.5076
			G=0.10			
Classical	0.0687	0.0329	1.5880	0.9371	1.1496	3.9399
MML (RES)	0.0583	0.0168	0.4764	0.4628	0.7736	0.8496
MML (PRESS)	0.0618	0.0111	0.5496	0.5099	1.0291	1.0878
Winsor (RES)	0.0595	0.0119	0.3839	0.4841	0.8194	0.6785
Winsor (PRES)	0.0602	0.0192	0.4057	0.4945	0.9027	0.7534
M-Winsor (RES)	0.0600	0.0121	0.3526	0.4720	0.7939	0.6238
M-Winsor (PRESS)	0.0606	0.0190	0.3750	0.4795	0.8691	0.6935
Andrews (c = 1.5)	0.0671	0.0244	0.3944	0.3446	0.6212	0.5278
Hubers (c = 1.5)	0.0678	0.0243	0.3944	0.3788	0.6504	0.5278
Adiche	0.0668	0.0267	1.6977	0.3251	0.5356	4.5125

Table 10.26: Bias and Mean square error of the estimators under the single-outlier model (with $\sigma = 6$) for the error($n = 20$)

				G=-.10		
Classical	-0.0974	0.0005	0.9695	0.2492	0.2658	1.5112
MML (RES)	-0.0840	-0.0069	0.1120	0.0922	0.1449	0.0750
MML (PRESS)	-0.0861	-0.0074	0.1215	0.0945	0.1556	0.0828
Winsor (RES)	-0.0832	-0.0033	0.0301	0.0929	0.1515	0.0521
Winsor (PRES)	-0.0830	-0.0097	0.0392	0.0954	0.1635	0.0556
M-Winsor (RES)	-0.0845	-0.0031	-0.0005	0.0916	0.1432	0.0471
M-Winsor (PRESS)	-0.0843	-0.0090	0.0087	0.0937	0.1538	0.0497
Andrews (c = 1.5)	-0.0916	-0.0024	0.1276	0.0831	0.0911	0.1169
Hubers (c = 1.5)	-0.0927	-0.0027	0.1276	0.0917	0.1046	0.1169
Adiche	-0.0990	-0.0021	1.0125	0.0918	0.0985	1.6502
				G=-.05		
Classical	-0.0526	-0.0004	0.9655	0.2414	0.2626	1.5049
MML (RES)	-0.0431	-0.0030	0.1061	0.0852	0.1382	0.0730
MML (PRESS)	-0.0438	-0.0085	0.1170	0.0870	0.1500	0.0800
Winsor (RES)	-0.0419	-0.0046	0.0251	0.0868	0.1441	0.0518
Winsor (PRES)	-0.0437	-0.0067	0.0349	0.0885	0.1535	0.0551
M-Winsor (RES)	-0.0428	-0.0044	-0.0054	0.0853	0.1362	0.0471
M-Winsor (PRESS)	-0.0444	-0.0061	0.0043	0.0866	0.1445	0.0494
Andrews (c = 1.5)	-0.0471	-0.0033	0.1215	0.0756	0.0870	0.1152
Hubers (c = 1.5)	-0.0479	-0.0040	0.1215	0.0841	0.1010	0.1152
Adiche	-0.0543	-0.0027	1.0087	0.0836	0.0950	1.6440
				G=0.05		
Classical	0.0369	-0.0023	0.9644	0.2396	0.2623	1.5034
MML (RES)	0.0378	0.0007	0.1035	0.0849	0.1366	0.0728
MML (PRESS)	0.0390	-0.0066	0.1160	0.0861	0.1450	0.0798
Winsor (RES)	0.0372	0.0013	0.0239	0.0858	0.1431	0.0520
Winsor (PRES)	0.0379	-0.0035	0.0332	0.0870	0.1505	0.0544
M-Winsor (RES)	0.0377	0.0010	-0.0066	0.0842	0.1353	0.0474
M-Winsor (PRESS)	0.0384	-0.0033	0.0027	0.0852	0.1418	0.0489
Andrews (c = 1.5)	0.0427	-0.0055	0.1187	0.0746	0.0861	0.1162
Hubers (c = 1.5)	0.0413	-0.0059	0.1187	0.0830	0.1008	0.1162
Adiche	0.0362	-0.0053	1.0075	0.0814	0.0953	1.6420
				G=0.10		
Classical	0.0817	-0.0032	0.9672	0.2455	0.2652	1.5082
MML (RES)	0.0793	0.0000	0.1099	0.0904	0.1433	0.0747
MML (PRESS)	0.0779	-0.0034	0.1196	0.0922	0.1486	0.0819
Winsor (RES)	0.0777	-0.0009	0.0293	0.0914	0.1465	0.0530
Winsor (PRES)	0.0783	-0.0007	0.0369	0.0933	0.1565	0.0555
M-Winsor (RES)	0.0787	-0.0009	-0.0014	0.0900	0.1386	0.0480
M-Winsor (PRESS)	0.0793	-0.0007	0.0063	0.0915	0.1476	0.0497
Andrews (c = 1.5)	0.0870	-0.0064	0.1234	0.0809	0.0890	0.1182
Hubers (c = 1.5)	0.0858	-0.0066	0.1234	0.0894	0.1041	0.1182
Adiche	0.0803	-0.0065	1.0102	0.0874	0.0986	1.6466

Chapter 11

MML Estimation for Multiple Linear Regression Model

11.1 Introduction

In the last chapter, we discussed the robustness features of various methods of estimation for a simple linear regression model and also derived the MML estimators of parameters in this case. In this chapter, we extend the MML method of estimation developed in chapter 10 to the multiple linear regression model. We also conduct a comparative study of various methods by considering the linear regression model based on two explanatory variables and demonstrate the efficiency and overall superiority of the MML estimation method developed in this chapter.

To this end, let us consider the multiple linear regression model given by

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{11.1}$$

where $\epsilon \sim N(0, \sigma^2\mathbf{I})$; $\mathbf{Y}_{n\times 1}$ is the response vector $\mathbf{X}_{n\times p}$ is the design matrix and $\beta_{p\times 1}$ is the parameter vector. Then the classical estimators of β and σ are given by

$$\begin{aligned} \hat{\beta}^* &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ \hat{\sigma}^2 &= \frac{(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)}{n-p} \end{aligned}$$

In this chapter we directly adopt the modified maximum likelihood approach of Tiku (1967) for the likelihood function based on the Type II symmetrically censored sample of normalized residuals and derive the MMLEs of the parameters in a multiple linear regression model. This is an extension of the method of estimation for a simple linear regression model described in the last chapter. In Section 2 we derive the MML estimators for the parameters in a multiple linear regression based on symmetrically censored samples. In Section 3 we derive approximate expressions for the variances and covariances of these estimators via the information matrix and establish the asymptotic distribution of these estimators. In Section 4 we examine the efficiency of the MMLE's under normal and a wide range of non-normal models for the error component and compare them with the Classical estimators, Andrews' (1973,1974) M-estimators, Huber's (1981) M-estimators, Yale and Forsythe's (1976) Winsorized estimators, and Tan and Tabatabai's (1988) Modified Winsorized estimators. Finally, in Section 5 we consider an example given by Belsley et al. (1980) and illustrate the methods of estimation mentioned above.

11.2 MML estimation Method

Write

$$Z_i = \frac{Y_i - \mathbf{X}_i\beta}{\sigma}, \quad (i = 1, 2, \ldots, n) \tag{11.2}$$

and denote their order statistics by

$$Z_{(1)} \leq Z_{(2)} \leq Z_{(2)} \leq \cdots \leq Z_{(n)} \tag{11.3}$$

Further, let us denote the $(\mathbf{X}, Y)$ pair corresponding to the $i'th$ order statistic $Z_{(i)}$ by $(\mathbf{X}_{[i]}, Y_{[i]})$ for $i = 1, 2, \ldots, n$. Let us consider the Type-II symmetrically censored sample

$$Z_{(r+1)} \leq Z_{(r+2)} \leq \cdots \leq Z_{(n-r)} \tag{11.4}$$

where r is usually chosen to be $[0.5 + 0.1n]$ based on robustness considerations. Under the assumption of normality, the likelihood function based on the symmetrically

censored sample (11.4) is

$$L = \frac{n!}{r!^2}\sigma^{-(n-2r)}\left\{\Phi(Z_{(r+1)})\right\}^r\left\{1-\Phi(Z_{(n-r)})\right\}^r\prod_{i=r+1}^{n-r}\phi(z_{(i)}) \tag{11.5}$$

where $\phi(z)$ and $\Phi(z)$ are the density function and cumulative distribution function of a standard normal variable, respectively. Thus

$$\ln L = Const - A\ln\sigma + r\ln\Phi(Z_{(r+1)}) + r\ln\left\{1-\Phi(Z_{(n-r)})\right\} - \frac{1}{2}\sum_{i=r+1}^{n-r} Z_{(i)}^2 \tag{11.6}$$

where $A = n - 2r$. Upon differentiating Eq. (11.6) with respect to β and σ we obtain the following system of likelihood equations

$$\begin{aligned}\frac{\partial \ln L}{\partial\beta} &= \frac{1}{\sigma}\left[-r\mathbf{x}_{[r+1]}^T\frac{\phi(Z_{(r+1)})}{\Phi(Z_{(r+1)})} + r\mathbf{x}_{[n-r]}^T\frac{\phi(Z_{(n-r)})}{1-\Phi(Z_{(n-r)})} + \sum_{i=r+1}^{n-r}\mathbf{x}_{[i]}^T Z_{(i)}\right] \\ &= 0\end{aligned} \tag{11.7}$$

and

$$\begin{aligned}\frac{\partial \ln L}{\partial\sigma} &= \frac{1}{\sigma}\left[-rZ_{[r+1]}\frac{\phi(Z_{(r+1)})}{\Phi(Z_{(r+1)})} + rZ_{[r+1]}\frac{\phi(Z_{(n-r)})}{1-\Phi(Z_{(n-r)})} + \sum_{i=r+1}^{n-r} Z_{(i)}^2\right] \\ &= 0\end{aligned} \tag{11.8}$$

Since Eqs. (11.7) and (11.8) do not provide explicit estimators for β and σ we use the same approximation as we did in the simple linear regression case, and the simplification is as follows:

$$\begin{aligned}\frac{\partial \ln L^*}{\partial\beta} &= \frac{1}{\sigma}\left[-r\mathbf{X}_{[r+1]}^T(\gamma+\delta Z_{(r+1)}) + r\mathbf{X}_{[n-r]}^T(\gamma+\delta Z_{(r+1)})\sum_{i=r+1}^{n-r}\mathbf{X}_{[i]}^T Z_{(i)}\right] \\ &= \frac{1}{\sigma}\left[r\gamma(\mathbf{X}_{[n-r]}^T - \mathbf{X}_{[r+1]}^T)\sum_{i=r+1}^{n-r}\mathbf{X}_{[i]}^T w_i Z_{(i)}\right]\end{aligned}$$

since

$$Z_{(i)} = \frac{Y_{[i]} - \mathbf{X}_{[i]}}{\sigma} \tag{11.9}$$

we have

$$\frac{\partial \ln L^*}{\partial\beta} = \frac{1}{\sigma^2}\left[r\gamma\sigma(\mathbf{X}_{[n-r]}^T - \mathbf{X}_{[r+1]}^T) + \sum_{i=r+1}^{n-r}\mathbf{X}_{[i]}^T w_i y_{[i]} - \sum_{i=r+1}^{n-r}\mathbf{x}_{[i]}^T w_i \mathbf{x}_{[i]}\beta\right]$$

which simplifies to

$$\begin{aligned}\frac{\partial \ln L^*}{\partial \beta} &= \frac{1}{\sigma^2}\left\{-\mathbf{X}^T\mathbf{W}\mathbf{X}\beta + \mathbf{X}^T\mathbf{W}\mathbf{Y} + r\sigma\gamma(\mathbf{X}_{n-r} - \mathbf{X}_{r+1})^T\right) \\ &= 0 \end{aligned} \tag{11.10}$$

where $\mathbf{X}$ is a $(n-2r)\times p$ matrix with i^{th} row being $\mathbf{X}_{[i]}$, for $i = r+1,\ldots,n-r$; $\mathbf{Y}$ is a column vector whose i^{th} element is $Y_{[i]}$;

$$\mathbf{W} = \begin{bmatrix} 1+r\delta & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1+r\delta \end{bmatrix}$$

Also

$$\begin{aligned}\frac{\partial \ln L}{\partial \sigma} \approx \frac{\partial \ln L^*}{\partial \sigma} &= \frac{1}{\sigma}\left[-A - rZ_{(r+1)}\left\{\gamma - \delta Z_{(r+1)}\right\} + rZ_{(n-r)}\left\{\gamma + \delta Z_{(n-r)}\right\} + \sum_{i=r+1}^{n-r} Z_{(i)}^2\right] \\ &= \frac{1}{\sigma}\left[-A + r\gamma(Z_{(n-r)} - Z_{(r+1)}) + \sum_{i=r+1}^{n-r} w_i Z_{(i)}^2\right] \\ &= \frac{1}{\sigma^3}\left[-A\sigma^2 + r\gamma\sigma(Y_{[n-r]} - Y_{[r+1]}) - \right. \\ &\quad \left. r\gamma\sigma(\mathbf{X}_{[n-r]} - \mathbf{X}_{[r+1]})\beta + \sum_{i=r+1}^{n-r} w_i(Y_{[i]} - \beta\mathbf{X}_{[i]})^2\right] \\ &= \frac{1}{\sigma^3}\left[-A\sigma^2 + r\gamma\sigma(Y_{[n-r]} - Y_{[r+1]})r\gamma\sigma(\mathbf{X}_{[n-r]} - \mathbf{X}_{[r+1]})\beta + \right. \\ &\quad \left. (\mathbf{Y} - \beta\mathbf{X})^T\mathbf{W}(\mathbf{Y} - \beta\mathbf{X})\right] \\ &= \frac{1}{\sigma^3}\left[-A\sigma^2 + r\gamma\sigma(Y_{[n-r]} - Y_{[r+1]})r\gamma\sigma(\mathbf{X}_{[n-r]} - \mathbf{X}_{[r+1]})\beta + \right. \\ &\quad \left. (\mathbf{Y}^T\mathbf{W}\mathbf{Y} + \beta\mathbf{X}^T\mathbf{W}\mathbf{X}\beta - 2\beta\mathbf{X}^T\mathbf{W}\mathbf{Y})\right] \\ &= 0 \end{aligned} \tag{11.11}$$

From Eq. (11.10) we derive the MML estimator of β as

$$\hat{\beta} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y} + r\gamma\hat{\sigma}\left\{(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}_{[n-r]} - \mathbf{X}_{[r+1]})^T\right\}. \tag{11.12}$$

Upon substituting the expression of $\hat{\beta}$ in (11.12) in Eq. (11.11) we obtain

$$\frac{\partial \ln L^*}{\partial \sigma} = \frac{1}{\sigma^3}\left[-A\sigma^2 + r\gamma\sigma(y_{(n-r)} - y_{(r+1)}) + r\gamma\sigma(\mathbf{x}_{(n-r)} - \mathbf{x}_{(r+1)})\hat{\beta} + (\mathbf{Y}^T\mathbf{W}\mathbf{Y} + \right.$$

$$\beta^T \mathbf{X}^T(\mathbf{W}\mathbf{Y} + r\gamma\sigma(\mathbf{x}_{[n-r]}^T - x_{[r+1]}^T)) - 2\beta\mathbf{X}^T\mathbf{W}\mathbf{Y}\Big]$$
$$= 0. \tag{11.13}$$

Upon simplifying the above equation we obtain

$$A\hat{\sigma}^2 + B\sigma - C = 0 \tag{11.14}$$

where

$$\begin{aligned} A &= n - 2r \\ B &= r\gamma\mathbf{Y}^T\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}_{[n-r]} - \mathbf{X}_{[r+1]}) - r\gamma(Y_{[n-r]} - Y_{[r+1]}) \end{aligned} \tag{11.15}$$

and

$$C = \mathbf{Y}^T(\mathbf{W} - \mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W})\mathbf{Y} \tag{11.16}$$

Let $\mathbf{Y}^* = \mathbf{W}^{\frac{1}{2}}\mathbf{Y}$ and $\mathbf{X}^* = \mathbf{W}^{\frac{1}{2}}\mathbf{X}$ then $C = \mathbf{Y}^{*T}(\mathbf{I} - \mathbf{X}^*(\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{X}^{*T})\mathbf{Y}^{*T}$, where $\mathbf{W}^{\frac{1}{2}}$ is a diagonal matrix whose elements are square root of $\mathbf{W}$. Since the matrix $(\mathbf{I} - \mathbf{X}^*(\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{X}^{*T})$ is idempotent, it has eigen values 1 or 0; this implies the matrix $(\mathbf{I} - \mathbf{X}^*(\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{X}^{*T})$ is positive semidefinite, therefore $C \geq 0$. Hence, the quadratic equation in (11.14) admits only one positive root thus yielding the MML estimator of σ to be

$$\hat{\sigma} = \frac{-B + \sqrt{B^2 + 4AC}}{2A} \tag{11.17}$$

Since $(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}r\gamma(\mathbf{X}_{[n-r]} - \mathbf{X}_{[r+1]})$ will usually be small, (11.12) essentially gives the simplified estimator

$$\hat{\beta} \approx (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y} \tag{11.18}$$

It is of interest to note here that the simplified estimator $\hat{\beta}$ is exactly of the same form as the classical weighted least squares estimator of β. It is this simplified estimator which will be studied in the rest of this chapter. Furthermore, following the suggestion of Tiku and Stewart (1977) and Tiku, Tan and Balakrishnan (1986), we may correct the estimator $\hat{\sigma}$ in (11.17) for its bias by defining

$$\hat{\sigma} = \frac{-B + \sqrt{B^2 + 4AC}}{2\{A(A-p)\}^{1/2}} \tag{11.19}$$

We may observe here that when $r = 0$ the MML estimators $\hat{\beta}$ and $\hat{\sigma}$ given in (11.18) and (11.19) simply reduce to the corresponding classical estimators.

In order to determine the $\mathbf{X}_{[i]}$ and $Y_{[i]}$ necessary for the computation of the MML estimators of $\hat{\beta}$ and $\hat{\sigma}$ we proceed as follows (with the same reasoning as in the simple linear regression case):

1. From the given data, we first obtain the least square estimates of β;
2. Next we compute the normalized estimated residuals for all the n pairs of observations and then arrange them in increasing order of magnitude; and
3. $(\mathbf{x}_{[i]}, y_{[i]})$ is taken to be the (x, y) pair corresponding to the $i'th$ largest normalized estimated residual for $i = 1, 2, \ldots, n$.

11.3 Asymptotic properties

Let us denote $\mu_{i:n}$ and $\mu_{i:n}^{(2)}$ $(1 \geq i \geq n)$ for the first and second moment, respectively, of the $i'th$ order statistic in a sample of size n from the standard normal distribution. These quantities have been quite extensively tabulated by Harter (1961) and Tietjen, Kahaner and Beckman (1971).

Then, we obtain from Equs. (11.10) and (11.13):

$$\begin{aligned}
\frac{\partial^2 \ln L^*}{\partial \beta^2} &= -\frac{\mathbf{X}^T\mathbf{W}\mathbf{X}}{\sigma^2} \\
\frac{\partial^2 \ln L^*}{\partial \beta \partial \sigma} &= -2\frac{(\mathbf{X}^T\mathbf{W}\mathbf{X})}{\sigma^3} + 2\frac{\mathbf{X}^T\mathbf{W}\mathbf{Y}}{\sigma^3} + r\gamma\frac{\mathbf{x}_{[n-r]}^T - \mathbf{x}_{[r+1]}^T}{\sigma^2} \\
\frac{\partial^2 \ln L^*}{\partial \sigma^2} &= -\frac{A}{\sigma^2} + 2\frac{r\gamma}{\sigma^2}\left[Z_{(n-r)} - Z_{(r+1)}\right] + 3\frac{r\beta}{\sigma^2}\sum_{i=r+1}^{n-r} w_i Z_{(i)}^2
\end{aligned}$$

From these formulae and the fact that

$$\begin{aligned}
E(\mathbf{Y}) &= E(\mathbf{Z}\sigma + \mathbf{X}\beta) \\
&= \mathbf{M}\sigma + \mathbf{X}\beta
\end{aligned}$$

where

$$\mathbf{M} = \begin{pmatrix} \mu_{r+1:n} \\ \mu_{r+2:n} \\ \vdots \\ \mu_{n-r:n} \end{pmatrix} \tag{11.20}$$

and w_i are the diagonal elements of W, we have

$$E\left[-\frac{\partial^2 \ln L}{\partial \beta^2}\right] \approx E\left[-\frac{\partial^2 \ln L^*}{\partial \beta}\right] = \frac{(\mathbf{X}^T\mathbf{W}\mathbf{X})}{\sigma^2} \tag{11.21}$$

$$E\left[-\frac{\partial^2 \ln L}{\partial \beta \sigma}\right] \approx E\left[-\frac{\partial^2 \ln L^*}{\partial \beta \sigma}\right] = \frac{2\mathbf{X}^T\mathbf{W}\mathbf{M} + r\gamma(\mathbf{x}_{[n-r]} - \mathbf{x}_{[r+1]})^T}{\sigma^2} \tag{11.22}$$

and

$$E\left[-\frac{\partial^2 \ln L}{\partial \sigma^2}\right] \approx E\left[-\frac{\partial^2 \ln L^*}{\partial \sigma^2}\right] = \frac{-A + 2r\gamma(\mu_{n-r:n} - \mu_{r+1:n}) + 3\sum_{i=r+1}^{n-r} w_i \mu_{i:n}^{(2)}}{\sigma^2} \tag{11.23}$$

Since the linear approximations in this method are very close approximates L is approximated very closely by L^* so that inference procedures for (β, σ) based on L^* would be almost as efficient as those based on L. We may note further that asymptotically L^* is identical to L, which establishes the asymptotic optimality of the MML estimators. Hence, by using standard results of maximum likelihood estimation, we have the following theorem which gives the asymptotic distribution of the MML estimators $(\hat{\beta}, \hat{\sigma})$ of (β, σ). For a proof of this theorem, reference may be made to Kendall and Stuart (1973) or Rao (1975).

Theorem 21 *Asymptotically, that is, for large $A = n - 2r$ (hence large n), $(\hat{\beta}, \hat{\sigma})$ has a $p+1$-variate normal distribution with mean vector $(E(\hat{\beta}) = \beta, E(\hat{\sigma}) = \sigma)$ variance-covariance matrix*

$$V = \begin{bmatrix} E\left[-\frac{\partial^2 \ln L^*}{\partial \beta^2}\right] & E\left[-\frac{\partial^2 \ln L^*}{\partial \beta \partial \sigma}\right] \\ & E\left[-\frac{\partial^2 \ln L^*}{\partial \sigma^2}\right] \end{bmatrix}^{-1} \tag{11.24}$$

$$= \sigma^2 \begin{bmatrix} (\mathbf{X}^T\mathbf{W}\mathbf{X}) & 2\mathbf{X}^T\mathbf{W}\mathbf{M} + r\gamma(\mathbf{X}_{[n-r]} - \mathbf{X}_{[r+1]})^T \\ & -A + 2r\gamma(\mu_{n-r:n} - \mu_{r+1:n}) + 3\sum_{i=r+1}^{n-r} w_i \mu_{i:n}^{(2)} \end{bmatrix}^{-1} \tag{11.25}$$

11.4 Comparison with the classical and other estimators

In this section, we make a comparative study of the MML estimation method developed in this chapter with other methods described in the last chapter. For this

purpose, we simulated the bias and mean square error of these MMLE's along with the classical estimators, Andrews' (1973,1974) M-estimators, Huber's (1981) M-estimators, Yale and Forsythe's (1976) Winsorized estimators and Tan and Tabatabai's (1988) modified Winsorized estimators for a wide range of alternatives to the normal distribution for the error component.

For assessing the robustness features of all the estimators described above, we followed the lines of Yale and Forsythe (1976), Tan and Tabatabai (1988), and Singh and Gupta (1989) and resorted to the following Monte Carlo simulation process. For this comparative study, we consider the multiple linear regression model with two explanatory variables (that is, the model in (11.1) with $p = 3$):

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

and with the choice of the parameters β_0, β_1, β_2 and σ having been taken as 0, 1, 1 and 1, respectively. The explanatory variables x_i's were sampled independently from the standard bivariate normal distribution with correlation coefficient $\rho = 0.5$. Further, we considered the following wide range of distributional models for the independent error variables ϵ_i's in (11.1):

1. Standard normal distribution
2. Mixture-normal model $0.9N(0,1) + 0.1N(0,4)$,
3. Mixture-normal model $0.9N(0,1) + 0.1N(0,6)$,
4. Mixture-normal model $0.8N(0,1) + 0.2N(0,4)$,
5. Single-outlier model with $(n-1)N(0,1)\&1N(0,4)$,
6. Single-outlier model with $(n-1)N(0,1)\&1N(0,6)$,
7. Two-outlier model with $(n-2)N(0,1)\&2N(0,4)$,
8. Two-outlier model with $(n-2)N(0,1)\&2N(0,6)$.

We simulated the values of the bias and mean square error of all the estimators of $(\beta_0, \beta_1, \beta_2, \sigma)$ for all 8 distributional models given above. These values are presented in Table 11.5-11.11 for sample sizes 10, 20 and 40. In these tables, bias and mean square errors of estimators are given in the following order:

1. Classical estimators
2. MML estimators (RES)
3. MML estimators (PRESS)
4. Winsorized estimators (RES)
5. Winsorized estimators (PRESS)
6. Modified Winsorized estimators (RES)
7. Modified Winsorized estimators (PRESS)
8. Andrews' M-estimators (c=1.5)
9. Huber's M-estimators (c=1.5)

All the results given in these tables are based on 4,000 Monte Carlo runs.

11.4.1 Comparison and conclusions

From the simulated values of the bias and mean square error of various estimators of β_0, β_1, β_2 and σ under normal and a very wide range of non-normal models for the error variables ϵ_i in the multiple linear regression model (11.1) presented in Tables 11.4-11.11, we observe the, following points:

1. From Table 11.4 we see that the classical estimators have the smallest mean square errors which is to be expected as they are the best estimators when the underlying distribution for the error variable is normal. The bias of all the estimators of β_0, β_1, β_2 and σ decrease as the sample size increases;

however, we note that the bias of the Winsorized and the modified Winsorized estimators of σ decrease only nominally as n increases and, as a result, these two estimators of σ remain quite biased even for a sample of size as large as 40. The M-estimators of β_0, β_1 and β_2 perform better than the corresponding MML estimators, but the M-estimators of σ have larger mean square errors than the MML estimators of σ.

2. From Table 11.5-11.11 we see that the classical estimator of σ develops serious bias and has very large mean square error as compared to all other estimators of σ when the underlying model for the error variable has scale contamination or contains a few outliers. In general, when all the estimators develop larger bias and larger mean square error under increasing departures from normality, the classical estimator of σ seems to face a devastating effect in this case. We observe that the bias of the Winsorized and Modified Winsorized estimators of σ decrease marginally when the sample size increases and, hence, they remain quite biased even when the sample size is as large as 40. We also note that while the M-estimators of α and β have smaller mean square errors than the corresponding MML estimators, the MML estimator of σ performs better than the M-estimators of σ.

From these points, we conclude that the classical estimators, though they perform well under normal and some non-normal models, are extremely sensitive (particularly the estimator of σ) to presence of outliers and also to scale-mixtures. The MML estimators seem to perform better in almost all cases than the Winsorized and the modified Winsorized estimators and in this regard we observe the MML estimators based on the RES method to be superior to those based on the PRESS method in almost all cases. It should be mentioned that the M-estimators of β_0, β_1 and β_2 in most cases turn out to be better than the corresponding MML estimators while MML estimator of σ almost always has smaller bias and mean square error than the M-estimators of σ. In conclusion, we therefore, would recommend the usage of the M-estimation (based on either Andrews' ψ-function or Huber's ψ-function) with a better estimator for σ or the modified maximum likelihood estimation based

on the RES method for the estimation of the parameters β_0, β_1, β_2 and σ in the multiple linear regression model in (11.1) whenever non-normality is suspected in the error variables and particularly when one or more outliers are suspected to be present in the data. It is of interest to note that the observations made here are exactly the same as the observations made in Chapter 10 while discussing a simple linear regression model.

11.4.2 Illustrative example

In this section we consider an example given by Belsley et al. (1980), and illustrate the various methods of estimation examined in this chapter.

Example:

Let us consider the data of cross-sectional study of the life-cycle savings in 50 countries given by Belsley et al. (1980). In this model the saving ratio (aggregate personal saving divided by disposable income) is explained by per-capita disposable income, the percentage rate of change in per-capita disposable income, and two demographic variables: the percentage of population less than 15 years old and the percentage of population over 75 years old. Belsley et al. (1980) have averaged over the decade 1960-1970 to remove the business cycle or other short-term fluctuations. These data are presented in Table 11.1.

Table 11.1: Intercountry life-cycle saving data

Index	Country	SR	POP15	POP75	DPI	ΔDPI
1	Australia	11.43	29.35	2.87	2329.68	2.87
2	Austria	12.07	23.32	4.41	1507.99	3.93
3	Belgium	13.17	23.80	4.43	2108.47	3.82
4	Bolivia	5.75	41.89	1.67	189.13	0.22
5	Brazil	12.88	42.19	0.83	728.47	4.56
6	Canada	8.79	31.72	2.85	2982.88	2.43
7	Chile	0.60	39.74	1.34	662.86	2.67
8	China(Taiwan)	11.90	44.75	0.67	289.52	6.51
9	Colombia	4.98	46.64	1.06	276.65	3.08
10	Costa Rica	10.78	47.64	1.14	471.24	2.80
11	Denmark	16.85	24.42	3.93	2496.53	3.99
12	Ecuador	3.59	46.31	1.19	287.77	2.19
13	Finland	11.24	27.84	2.37	1681.25	4.32
14	France	12.64	25.06	4.70	2213.82	4.52
15	Germany (F.R.)	12.55	23.31	3.35	2457.12	3.44
16	Greece	10.67	25.62	3.10	870.85	6.28
17	Guatemala	3.01	46.05	0.87	289.71	1.48
18	Honduras	7.70	47.32	0.58	232.44	3.19
19	Iceland	1.27	34.03	3.08	1900.10	1.12
20	India	9.00	41.31	0.96	88.94	1.54
21	Ireland	11.34	31.16	4.19	1139.95	2.99
22	Italy	14.28	24.52	3.48	1390.00	3.54
23	Japan	21.10	27.01	1.91	1257.28	8.21
24	Korea	3.98	41.74	0.91	207.68	5.81
25	Luxembourg	10.35	21.80	3.73	2449.29	1.57

Table 11.2: Continued

Index	Country	SR	POP15	POP75	DPI	ΔDPI
26	Malta	15.48	32.54	2.47	601.05	8.12
27	Norway	10.25	25.95	3.67	2231.03	3.62
28	Netherlands	14.65	24.71	3.25	1740.70	7.66
29	New Zealand	10.67	32.61	3.17	1487.52	1.76
30	Nicaragua	7.30	45.04	1.21	325.54	2.48
31	Panama	4.44	43.56	1.20	568.56	3.61
32	Paraguay	2.02	41.18	1.05	220.56	1.03
33	Peru	12.70	44.19	1.28	400.06	0.67
34	Philippines	12.78	46.26	1.12	152.01	2.00
35	Portugal	12.49	28.96	2.85	579.51	7.48
36	South Africa	11.14	31.94	2.28	651.11	2.19
37	South Rhodesia	13.30	31.92	1.52	250.96	2.00
38	Spain	11.77	27.74	2.87	768.79	4.35
39	Sweden	6.86	21.44	4.54	3299.49	3.01
40	Switzerland	14.13	23.49	3.73	2630.96	2.70
41	Turkey	5.13	43.42	1.08	389.66	2.96
42	Tunisia	2.81	46.12	1.21	249.87	1.13
43	United Kingdom	7.81	23.27	4.46	1813.93	2.01
44	United States	7.56	29.81	3.43	4001.89	2.45
45	Venezuela	9.22	46.40	0.90	813.39	0.53
46	Zambia	18.56	45.25	0.56	138.33	5.14
47	Jamaica	7.72	41.12	1.73	380.47	10.23
48	Uruguay	9.24	28.13	2.72	766.54	1.88
49	Libya	8.89	43.69	2.07	123.58	16.71
50	Malaysia	4.71	47.20	0.66	243.69	5.08

Table 11.3: Estimates of parameters for Belsley et al. (1980) data.

Est.	β_0	β_1	β_2	β_3	β_4	σ
(1)	28.5663	-0.4612	-1.6914	-0.0003	0.4097	3.8027
(2)	21.9358	-0.3021	-0.7667	-0.0002	0.4143	2.6679
(3)	20.8038	-0.2831	-0.7070	-0.0010	0.4171	2.8443
(4)	18.6685	-0.2688	-0.3706	0.0000	0.3458	3.2573
(5)	20.1942	-0.2670	-0.6488	-0.0010	0.4175	3.1919
(6)	19.7575	-0.2897	-0.4929	0.0000	0.3359	3.1405
(7)	20.7917	-0.2828	-0.7058	-0.0010	0.4171	3.1215
(8)	28.6802	-0.4675	-1.6242	-0.0004	0.3865	3.7832
(9)	28.7444	-0.4688	-1.6531	-0.0003	0.3947	3.7832

For this data, Belsley et al. (1980) have fitted a multiple linear regression model. The classical estimates, the MML estimates, the Winsorized estimates, the Modified Winsorized estimates and the M-estimates were computed for this data and are presented in Table 11.2

In order to illustrate the robustness of MML estimates we contaminated Belsley et al. (1980) data in Table 11.9 by changing the fourth y (SR) from 5.75 to 15.00, the sixth y (SR) from 8.79 to 15.87, thirty first y (SR) from 4.44 to 12.78 and forty fourth y (SR) from 7.56 to 12.56. Then we fit the same model and the estimates of parameters computed in this case are presented in Table 11.3. We observe here that introduction of outliers had some effect on all the estimators, the classical estimates being the most affected ones and MML estimates being the least affected ones.

Table 11.4: estimation of parameters for Belsley et al. (1980) data with contamination

Est.	β_0	β_1	β_2	β_3	β_4	σ
(1)	25.2465	-0.3695	-1.7652	0.0007	0.3456	4.0931
(2)	22.2940	-0.3507	-1.7645	0.0023	0.4391	2.7462
(3)	24.5553	-0.3478	-2.3948	0.0014	0.4689	3.3091
(4)	22.2715	-0.3538	-1.7981	0.0024	0.4556	3.0075
(5)	22.8640	-0.3305	-1.8454	0.0022	0.2644	2.7512
(6)	22.2935	-0.3508	-1.7652	0.0023	0.4394	2.9619
(7)	22.9059	-0.3318	-1.8164	0.0021	0.2660	2.7372
(8)	25.0649	-0.3677	-1.7309	0.0008	0.3282	3.6800
(9)	25.3736	-0.3731	-1.7916	0.0008	0.3337	3.6800

Table 11.5: Bias and Mean square error of the estimators under the normal model for the error

	Bias				MSError			
	β_0	β_1	β_2	σ	β_0	β_1	β_2	σ
	$n = 10$							
(1)	0.0002	0.0139	-0.0113	-0.0414	0.1301	0.2112	0.2127	0.0696
(2)	0.0022	0.0122	-0.0098	-0.1875	0.1742	0.3956	0.4008	0.1282
(3)	0.0080	0.0114	0.0048	-0.1336	0.2036	0.5180	0.5071	0.1323
(4)	0.0026	0.0087	-0.0054	-0.2285	0.1662	0.3715	0.3598	0.1312
(5)	0.0096	0.0176	-0.0028	-0.2136	0.1954	0.4696	0.4775	0.1327
(6)	0.0026	0.0088	-0.0055	-0.2436	0.1644	0.3634	0.3521	0.1344
(7)	0.0094	0.0173	-0.0027	-0.2279	0.1920	0.4580	0.4657	0.1347
(8)	0.0013	0.0121	-0.0085	-0.2087	0.1468	0.2313	0.2327	0.1438
(9)	0.0012	0.0121	-0.0098	-0.2087	0.1380	0.2184	0.2195	0.1438
	$n = 20$							
(1)	-0.0050	0.0003	0.0028	-0.0162	0.0562	0.0841	0.0825	0.0290
(2)	-0.0033	-0.0001	-0.0013	-0.1052	0.0667	0.1540	0.1583	0.0531
(3)	-0.0026	0.0026	-0.0020	-0.0891	0.0692	0.1682	0.1699	0.0551
(4)	-0.0022	-0.0030	-0.0016	-0.1652	0.0668	0.1609	0.1601	0.0627
(5)	-0.0031	-0.0011	-0.0028	-0.1588	0.0697	0.1704	0.1738	0.0625
(6)	-0.0024	-0.0029	-0.0013	-0.1874	0.0656	0.1528	0.1522	0.0681
(7)	-0.0032	-0.0008	-0.0027	-0.1809	0.0682	0.1615	0.1648	0.0673
(8)	-0.0024	-0.0003	0.0017	-0.0962	0.0601	0.0886	0.0875	0.0679
(9)	-0.0033	-0.0006	0.0026	-0.0962	0.0588	0.0863	0.0856	0.0679
	$n = 40$							
(1)	-0.0015	-0.0008	0.0031	-0.0081	0.0272	0.0373	0.0365	0.0135
(2)	-0.0022	-0.0036	0.0009	-0.0715	0.0312	0.0809	0.0808	0.0260
(3)	-0.0019	-0.0069	0.0084	-0.0673	0.0315	0.0829	0.0826	0.0269
(4)	-0.0024	-0.0024	0.0009	-0.1387	0.0316	0.0900	0.0885	0.0367
(5)	-0.0030	-0.0041	0.0032	-0.1364	0.0320	0.0898	0.0900	0.0365
(6)	-0.0023	-0.0023	0.0009	-0.1643	0.0309	0.0821	0.0807	0.0430
(7)	-0.0027	-0.0038	0.0032	-0.1621	0.0312	0.0818	0.0819	0.0426
(8)	-0.0011	-0.0014	0.0036	-0.0438	0.0283	0.0389	0.0378	0.0347
(9)	-0.0011	-0.0011	0.0033	-0.0438	0.0282	0.0387	0.0376	0.0347

Table 11.6: Bias and Mean square error of the estimators under the mixture normal model (with $p = 0.9$ and $\sigma = 4$) for the error

	Bias				MSError			
	β_0	β_1	β_2	σ	β_0	β_1	β_2	σ
	$n = 10$							
(1)	0.0064	-0.0211	0.0047	0.4206	0.3263	0.5322	0.5239	0.7049
(2)	-0.0030	-0.0085	0.0010	-0.0238	0.2732	0.5996	0.5897	0.1925
(3)	-0.0101	-0.0081	-0.0030	0.0525	0.3269	0.7677	0.7661	0.2821
(4)	-0.0042	-0.0171	0.0083	-0.0628	0.2673	0.5568	0.5583	0.1809
(5)	-0.0054	-0.0120	0.0010	-0.0416	0.3003	0.6688	0.6805	0.2073
(6)	-0.0039	-0.0171	0.0083	-0.0805	0.2639	0.5459	0.5468	0.1753
(7)	-0.0052	-0.0124	0.0011	-0.0582	0.2948	0.6519	0.6614	0.1994
(8)	-0.0009	-0.0137	0.0037	-0.0012	0.2483	0.4276	0.4323	0.2277
(9)	0.0028	-0.0183	0.0040	-0.0012	0.2416	0.4208	0.4270	0.2277
	$n = 20$							
(1)	-0.0077	0.0087	0.0015	0.4853	0.1422	0.2161	0.2039	0.5229
(2)	-0.0046	-0.0012	0.0088	0.0479	0.0936	0.1971	0.2024	0.0858
(3)	-0.0060	-0.0017	0.0084	0.0640	0.1005	0.2151	0.2209	0.0969
(4)	-0.0051	0.0015	0.0055	-0.0211	0.0948	0.2021	0.2033	0.0715
(5)	-0.0060	-0.0008	0.0104	-0.0093	0.1009	0.2192	0.2275	0.0764
(6)	-0.0051	0.0011	0.0057	-0.0471	0.0927	0.1925	0.1935	0.0680
(7)	-0.0059	-0.0007	0.0101	-0.0352	0.0981	0.2083	0.2152	0.0717
(8)	-0.0054	0.0012	0.0062	0.0678	0.0792	0.1274	0.1306	0.1054
(9)	-0.0065	0.0026	0.0062	0.0678	0.0833	0.1357	0.1345	0.1054
	$n = 10$							
(1)	-0.0002	0.0001	0.0071	0.5250	0.0681	0.0906	0.0949	0.4316
(2)	-0.0009	-0.0012	0.0062	0.0634	0.0446	0.0990	0.1008	0.0410
(3)	0.0001	-0.0004	0.0068	0.0680	0.0455	0.1021	0.1039	0.0447
(4)	-0.0028	-0.0014	0.0082	-0.0131	0.0453	0.1062	0.1082	0.0302
(5)	0.0000	0.0018	0.0036	-0.0099	0.0466	0.1119	0.1112	0.0321
(6)	-0.0025	-0.0013	0.0078	-0.0427	0.0440	0.0968	0.0988	0.0291
(7)	0.0001	0.0016	0.0036	-0.0395	0.0451	0.1018	0.1015	0.0304
(8)	-0.0013	0.0026	0.0037	0.0863	0.0371	0.0515	0.0528	0.0540
(9)	-0.0009	0.0014	0.0046	0.0863	0.0393	0.0536	0.0550	0.0540

Table 11.7: Bias and Mean square error of the estimators under the mixture normal model (with $p = 0.9$ and $\sigma = 6$) for the error

	Bias				MSError			
	β_0	β_1	β_2	σ	β_0	β_1	β_2	σ
	$n = 10$							
(1)	0.0092	-0.0277	0.0070	0.7975	0.5903	0.9672	0.9472	2.0121
(2)	-0.0061	-0.0171	0.0075	0.0515	0.3539	0.7639	0.7511	0.3386
(3)	-0.0080	-0.0062	-0.0075	0.1417	0.4335	0.9970	0.9956	0.5453
(4)	-0.0047	-0.0153	-0.0030	0.0157	0.3560	0.7409	0.7686	0.3212
(5)	-0.0074	-0.0145	0.0071	0.0374	0.3824	0.8495	0.8546	0.3698
(6)	-0.0044	-0.0155	-0.0028	-0.0030	0.3508	0.7268	0.7529	0.3060
(7)	-0.0072	-0.0150	0.0071	0.0200	0.3752	0.8281	0.8299	0.3521
(8)	-0.0010	-0.0175	0.0049	0.1414	0.3254	0.6046	0.6346	0.4526
(9)	0.0031	-0.0223	0.0040	0.1414	0.3360	0.6306	0.6481	0.4526
	$n = 20$							
(1)	-0.0100	0.0124	-0.0003	0.9190	0.2555	0.3898	0.3642	1.6475
(2)	-0.0059	-0.0036	0.0051	0.1006	0.1146	0.2268	0.2327	0.1466
(3)	-0.0074	-0.0008	0.0045	0.1222	0.1221	0.2484	0.2509	0.1713
(4)	-0.0071	0.0017	0.0091	0.0348	0.1167	0.2303	0.2366	0.1214
(5)	-0.0091	-0.0023	0.0057	0.0487	0.1241	0.2503	0.2605	0.1355
(6)	-0.0071	0.0011	0.0091	0.0078	0.1130	0.2197	0.2244	0.1110
(7)	-0.0090	-0.0022	0.0057	0.0219	0.1196	0.2375	0.2459	0.1230
(8)	-0.0075	0.0018	0.0064	0.1661	0.0839	0.1395	0.1429	0.1928
(9)	-0.0079	0.0040	0.0057	0.1661	0.0961	0.1665	0.1643	0.1928
	$n = 40$							
(1)	-0.0006	-0.0024	0.0107	1.0061	0.1210	0.1628	0.1717	1.4556
(2)	-0.0010	0.0000	0.0048	0.0998	0.0504	0.1080	0.1103	0.0645
(3)	-0.0001	0.0021	0.0056	0.1062	0.0511	0.1120	0.1118	0.0686
(4)	-0.0016	-0.0023	0.0076	0.0260	0.0524	0.1172	0.1173	0.0466
(5)	0.0000	0.0014	0.0053	0.0274	0.0535	0.1217	0.1195	0.0473
(6)	-0.0013	-0.0020	0.0072	-0.0045	0.0504	0.1066	0.1074	0.0414
(7)	0.0001	0.0013	0.0050	-0.0026	0.0513	0.1107	0.1094	0.0418
(8)	-0.0009	0.0039	0.0034	0.1482	0.0374	0.0520	0.0537	0.0844
(9)	-0.0012	0.0009	0.0055	0.1482	0.0428	0.0595	0.0615	0.0844

Table 11.8: Bias and Mean square error of the estimators under the mixture normal model (with $p = 0.8$ and $\sigma = 4$) for the error

	Bias				MSError			
	β_0	β_1	β_2	σ	β_0	β_1	β_2	σ
	$n = 10$							
(1)	0.0200	-0.0296	0.0032	0.7914	0.5214	0.8582	0.8727	1.3724
(2)	0.0039	-0.0150	0.0112	0.1819	0.4520	0.9577	0.9771	0.3962
(3)	-0.0025	-0.0299	0.0040	0.2851	0.5214	1.2987	1.3071	0.5804
(4)	0.0038	-0.0136	0.0108	0.1427	0.4265	0.8476	0.8884	0.3612
(5)	-0.0017	-0.0209	0.0158	0.1742	0.4910	1.1193	1.1587	0.4126
(6)	0.0043	-0.0139	0.0105	0.1210	0.4207	0.8300	0.8702	0.3394
(7)	-0.0014	-0.0213	0.0157	0.1539	0.4810	1.0869	1.1247	0.3883
(8)	0.0062	-0.0154	0.0008	0.2142	0.3947	0.7076	0.7393	0.4241
(9)	0.0099	-0.0242	0.0035	0.2142	0.3892	0.6918	0.7176	0.4241
	$n = 20$							
(1)	-0.0101	0.0119	-0.0021	0.8926	0.2254	0.3383	0.3172	1.1947
(2)	-0.0109	0.0025	0.0010	0.2660	0.1561	0.2903	0.2854	0.2533
(3)	-0.0099	0.0049	0.0022	0.2899	0.1644	0.3135	0.3017	0.2911
(4)	-0.0091	-0.0011	0.0041	0.1867	0.1564	0.3062	0.3007	0.1907
(5)	-0.0109	0.0037	0.0028	0.2013	0.1684	0.3276	0.3155	0.2103
(6)	-0.0091	-0.0007	0.0042	0.1544	0.1514	0.2896	0.2834	0.1677
(7)	-0.0106	0.0036	0.0029	0.1692	0.1616	0.3089	0.2960	0.1848
(8)	-0.0076	0.0009	0.0104	0.2538	0.1241	0.2092	0.2054	0.2210
(9)	-0.0084	0.0029	0.0062	0.2538	0.1253	0.2105	0.2025	0.2210
	$n = 40$							
(1)	-0.0031	-0.0021	0.0039	0.9341	0.1108	0.1453	0.1466	1.0719
(2)	-0.0051	0.0026	0.0009	0.2573	0.0714	0.1424	0.1425	0.1456
(3)	-0.0046	0.0002	0.0040	0.2633	0.0725	0.1421	0.1457	0.1543
(4)	-0.0073	0.0003	0.0032	0.1682	0.0732	0.1547	0.1566	0.0944
(5)	-0.0059	0.0041	0.0008	0.1729	0.0751	0.1550	0.1571	0.0989
(6)	-0.0066	0.0006	0.0031	0.1316	0.0702	0.1397	0.1414	0.0769
(7)	-0.0055	0.0039	0.0011	0.1363	0.0717	0.1400	0.1418	0.0805
(8)	-0.0034	0.0020	0.0038	0.2369	0.0532	0.0764	0.0764	0.1221
(9)	-0.0033	0.0011	0.0036	0.2369	0.0567	0.0790	0.0793	0.1221

Table 11.9: Bias and Mean square error of the estimators under the single outlier-normal model (with $\sigma = 4$) for the error

	Bias				MSError			
	β_0	β_1	β_2	σ	β_0	β_1	β_2	σ
	$n = 10$							
(1)	0.0037	0.0282	-0.0134	0.4257	0.3326	0.5062	0.5376	0.5592
(2)	-0.0010	0.0107	-0.0053	-0.0581	0.2217	0.4671	0.4701	0.1307
(3)	0.0103	0.0247	-0.0062	0.0094	0.2795	0.7235	0.7075	0.1762
(4)	0.0031	0.0192	-0.0077	-0.1000	0.2280	0.4740	0.4849	0.1207
(5)	0.0056	0.0197	-0.0106	-0.0817	0.2491	0.5888	0.5750	0.1289
(6)	0.0031	0.0191	-0.0076	-0.1166	0.2257	0.4647	0.4735	0.1190
(7)	0.0056	0.0195	-0.0100	-0.0969	0.2448	0.5740	0.5610	0.1259
(8)	0.0072	0.0099	-0.0046	-0.0100	0.2174	0.3724	0.4055	0.1754
(9)	0.0061	0.0165	-0.0066	-0.0100	0.2243	0.3788	0.4108	0.1754
	$n = 20$							
(1)	-0.0033	0.0029	-0.0015	0.2634	0.0988	0.1428	0.1476	0.2189
(2)	-0.0033	0.0025	-0.0009	-0.0364	0.0779	0.1692	0.1693	0.0524
(3)	-0.0021	0.0023	-0.0057	-0.0187	0.0798	0.1806	0.1808	0.0570
(4)	-0.0020	-0.0021	0.0048	-0.1015	0.0774	0.1728	0.1722	0.0525
(5)	-0.0034	0.0029	-0.0062	-0.0920	0.0806	0.1845	0.1804	0.0537
(6)	-0.0022	-0.0020	0.0046	-0.1254	0.0760	0.1645	0.1644	0.0548
(7)	-0.0034	0.0029	-0.0058	-0.1157	0.0787	0.1753	0.1717	0.0551
(8)	-0.0023	0.0016	0.0010	-0.0206	0.0675	0.1026	0.1017	0.0713
(9)	-0.0026	0.0011	0.0013	-0.0206	0.0690	0.1046	0.1047	0.0713
	$n = 40$							
(1)	-0.0009	-0.0011	0.0025	0.1396	0.0365	0.0499	0.0499	0.0700
(2)	-0.0020	-0.0006	0.0030	-0.0408	0.0336	0.0860	0.0845	0.0246
(3)	-0.0012	-0.0041	0.0079	-0.0337	0.0338	0.0868	0.0871	0.0253
(4)	-0.0018	0.0043	0.0009	-0.1109	0.0342	0.0944	0.0922	0.0315
(5)	-0.0011	0.0004	0.0039	-0.1059	0.0348	0.0939	0.0951	0.0304
(6)	-0.0017	0.0039	0.0008	-0.1372	0.0334	0.0862	0.0842	0.0363
(7)	-0.0011	0.0004	0.0036	-0.1326	0.0338	0.0856	0.0867	0.0351
(8)	-0.0013	-0.0011	0.0031	-0.0150	0.0300	0.0414	0.0403	0.0347
(9)	-0.0009	-0.0010	0.0029	-0.0150	0.0303	0.0417	0.0407	0.0347

Table 11.10: Bias and Mean square error of the estimators under the single outlier-normal model (with $\sigma = 6$) for the error

	Bias				MSError			
	β_0	β_1	β_2	σ	β_0	β_1	β_2	σ
	$n = 10$							
(1)	0.0061	0.0378	-0.0148	0.8330	0.5990	0.9094	0.9705	1.6458
(2)	0.0003	0.0137	-0.0120	-0.0299	0.2343	0.4863	0.4782	0.1343
(3)	0.0066	0.0322	-0.0091	0.0444	0.3254	0.8886	0.8517	0.2255
(4)	0.0045	0.0229	-0.0119	-0.0666	0.2479	0.5185	0.5291	0.1231
(5)	0.0021	0.0170	-0.0074	-0.0450	0.2629	0.6153	0.6043	0.1416
(6)	0.0046	0.0227	-0.0115	-0.0829	0.2455	0.5093	0.5202	0.1200
(7)	0.0023	0.0168	-0.0072	-0.0601	0.2585	0.5996	0.5900	0.1369
(8)	0.0061	0.0125	-0.0063	0.1359	0.2568	0.4642	0.5127	0.3006
(9)	0.0082	0.0221	-0.0090	0.1359	0.2907	0.5187	0.5751	0.3006
	$n = 20$							
(1)	-0.0022	0.0045	-0.0044	0.5290	0.1558	0.2220	0.2345	0.6859
(2)	-0.0027	-0.0002	-0.0042	-0.0190	0.0800	0.1712	0.1687	0.0543
(3)	-0.0046	-0.0008	0.0008	-0.0039	0.0825	0.1843	0.1830	0.0586
(4)	-0.0014	-0.0029	0.0023	-0.0853	0.0803	0.1711	0.1676	0.0510
(5)	-0.0035	0.0035	-0.0057	-0.0731	0.0833	0.1885	0.1840	0.0525
(6)	-0.0017	-0.0029	0.0019	-0.1089	0.0788	0.1637	0.1608	0.0524
(7)	-0.0035	0.0033	-0.0057	-0.0967	0.0814	0.1795	0.1759	0.0529
(8)	-0.0028	0.0020	0.0010	0.0256	0.0673	0.1031	0.1026	0.0897
(9)	-0.0025	0.0013	0.0016	0.0256	0.0719	0.1120	0.1128	0.0897
	$n = 40$							
(1)	-0.0004	-0.0014	0.0021	0.2936	0.0489	0.0673	0.0678	0.2307
(2)	-0.0011	-0.0006	0.0027	-0.0323	0.0343	0.0848	0.0835	0.0245
(3)	-0.0016	-0.0035	0.0036	-0.0272	0.0343	0.0879	0.0886	0.0256
(4)	-0.0018	-0.0027	0.0014	-0.1030	0.0348	0.0935	0.0920	0.0301
(5)	-0.0005	-0.0021	0.0050	-0.0976	0.0350	0.0935	0.0923	0.0294
(6)	-0.0017	-0.0024	0.0013	-0.1296	0.0340	0.0858	0.0843	0.0345
(7)	-0.0004	-0.0020	0.0047	-0.1244	0.0341	0.0857	0.0846	0.0335
(8)	-0.0016	-0.0006	0.0031	-0.0030	0.0300	0.0413	0.0401	0.0354
(9)	-0.0009	-0.0007	0.0028	-0.0030	0.0307	0.0423	0.0414	0.0354

Table 11.11: Bias and Mean square error of the estimators under the two outlier-normal model (with $\sigma = 4$) for the error

	Bias				MSError			
	β_0	β_1	β_2	σ	β_0	β_1	β_2	σ
	$n = 10$							
(1)	0.0086	0.0281	-0.0213	0.8337	0.5449	0.8637	0.8489	1.2816
(2)	0.0081	0.0031	0.0067	0.1497	0.3735	0.7736	0.7658	0.2806
(3)	0.0129	0.0056	0.0031	0.2522	0.4810	1.1739	1.1047	0.4421
(4)	0.0040	0.0060	0.0060	0.1062	0.3719	0.7559	0.7784	0.2367
(5)	0.0084	0.0156	0.0044	0.1386	0.4254	0.9686	0.9268	0.2809
(6)	0.0042	0.0063	0.0054	0.0857	0.3670	0.7408	0.7621	0.2214
(7)	0.0084	0.0153	0.0045	0.1192	0.4160	0.9420	0.9013	0.2627
(8)	0.0140	0.0074	-0.0063	0.2087	0.3502	0.6398	0.6443	0.3101
(9)	0.0103	0.0158	-0.0127	0.2087	0.3544	0.6396	0.6491	0.3101
	$n = 20$							
(1)	-0.0042	0.0025	-0.0040	0.5132	0.1404	0.2045	0.2054	0.4935
(2)	-0.0016	0.0014	-0.0017	0.0333	0.0887	0.1838	0.1901	0.0614
(3)	0.0006	0.0029	-0.0034	0.0535	0.0912	0.1973	0.1954	0.0709
(4)	-0.0003	-0.0010	0.0037	-0.0350	0.0882	0.1928	0.1941	0.0515
(5)	0.0011	0.0043	-0.0084	-0.0233	0.0920	0.1963	0.1981	0.0541
(6)	-0.0004	-0.0008	0.0037	-0.0604	0.0867	0.1838	0.1853	0.0501
(7)	0.0007	0.0042	-0.0080	-0.0486	0.0900	0.1869	0.1885	0.0516
(8)	-0.0001	-0.0019	0.0027	0.0649	0.0782	0.1219	0.1205	0.0897
(9)	-0.0008	-0.0009	0.0014	0.0649	0.0812	0.1265	0.1258	0.0897
	$n = 40$							
(1)	-0.0018	-0.0019	0.0036	0.2843	0.0473	0.0631	0.0634	0.1619
(2)	-0.0017	-0.0010	-0.0010	-0.0098	0.0358	0.0912	0.0885	0.0249
(3)	-0.0012	-0.0001	0.0027	-0.0036	0.0361	0.0923	0.0909	0.0269
(4)	-0.0017	0.0016	0.0014	-0.0808	0.0364	0.1008	0.0970	0.0270
(5)	-0.0009	-0.0035	0.0054	-0.0762	0.0373	0.1019	0.0998	0.0273
(6)	-0.0016	0.0014	0.0012	-0.1082	0.0356	0.0920	0.0886	0.0304
(7)	-0.0009	-0.0031	0.0051	-0.1038	0.0362	0.0928	0.0911	0.0303
(8)	-0.0012	-0.0014	0.0023	0.0153	0.0314	0.0438	0.0424	0.0358
(9)	-0.0012	-0.0014	0.0027	0.0153	0.0324	0.0450	0.0436	0.0358

Table 11.12: Bias and Mean square error of the estimators under the two outlier-normal model (with $\sigma = 6$) for the error

	Bias				MSError			
	β_0	β_1	β_2	σ	β_0	β_1	β_2	σ
	$n = 10$							
(1)	0.0142	0.0375	-0.0280	1.5435	1.0961	1.7395	1.6948	3.8186
(2)	0.0102	-0.0054	0.0107	0.3273	0.5552	1.1468	1.1121	0.6145
(3)	0.0201	0.0092	0.0115	0.4845	0.7487	1.9332	1.7415	1.0457
(4)	0.0055	0.0037	0.0014	0.2962	0.5909	1.2315	1.2337	0.5590
(5)	0.0133	0.0093	0.0133	0.3437	0.6586	1.5391	1.4194	0.6757
(6)	0.0057	0.0044	0.0010	0.2728	0.5810	1.2031	1.2057	0.5215
(7)	0.0133	0.0095	0.0130	0.3212	0.6411	1.4923	1.3783	0.6320
(8)	0.0134	0.0059	-0.0017	0.4931	0.5179	1.0289	1.0279	0.7364
(9)	0.0145	0.0198	-0.0162	0.4931	0.5611	1.0852	1.0900	0.7364
	$n = 20$							
(1)	-0.0037	0.0039	-0.0085	0.9846	0.2535	0.3669	0.3709	1.5806
(2)	-0.0018	0.0025	-0.0016	0.0643	0.0937	0.1914	0.1859	0.0736
(3)	-0.0019	0.0042	0.0015	0.0823	0.0988	0.2111	0.2075	0.0824
(4)	-0.0015	-0.0009	0.0018	-0.0001	0.0933	0.1925	0.1944	0.0578
(5)	-0.0016	0.0035	0.0004	0.0128	0.0985	0.2080	0.2009	0.0627
(6)	-0.0015	-0.0009	0.0020	-0.0254	0.0916	0.1840	0.1863	0.0540
(7)	-0.0016	0.0036	0.0002	-0.0121	0.0961	0.1977	0.1917	0.0576
(8)	-0.0017	0.0001	0.0009	0.1605	0.0803	0.1271	0.1253	0.1502
(9)	-0.0005	-0.0012	0.0003	0.1605	0.0903	0.1474	0.1469	0.1502
	$n = 40$							
(1)	-0.0020	-0.0026	0.0039	0.5729	0.0737	0.0980	0.0997	0.5624
(2)	-0.0019	-0.0039	0.0005	0.0036	0.0371	0.0912	0.0906	0.0272
(3)	-0.0004	-0.0077	0.0038	0.0104	0.0376	0.0938	0.0923	0.0295
(4)	-0.0024	-0.0049	-0.0004	-0.0673	0.0378	0.0996	0.0992	0.0264
(5)	-0.0018	-0.0082	0.0036	-0.0641	0.0384	0.1011	0.0997	0.0274
(6)	-0.0023	-0.0044	-0.0003	-0.0947	0.0369	0.0915	0.0910	0.0288
(7)	-0.0017	-0.0075	0.0035	-0.0915	0.0373	0.0929	0.0916	0.0294
(8)	-0.0013	-0.0008	0.0023	0.0426	0.0312	0.0435	0.0423	0.0404
(9)	-0.0012	-0.0014	0.0026	0.0426	0.0333	0.0465	0.0451	0.0404

Chapter 12

Contributions, Conclusions and Further Research

12.1 Contributions and Conclusions

In this thesis, we have presented some robust classification procedures and robust regression procedures. The existing techniques for these problems are either non robust to departures from normality or too complicated to use. The assumption that observations come from a normal distribution is unrealistic in practice and the classical methods are too sensitive to departures from normality. Since the extreme observations or outliers are censored under the MML approach, one could expect procedures based on MML estimators or estimators obtained by using the MML approach to be robust. In this thesis, we have explored these ideas for problems of statistical classification and linear regression. Summary of the contributions and conclusions made in this thesis are listed below:

1. We have used the multivariate MML estimators to develop a robust classification procedure for two way classification problem. By applying this procedure along with the classical procedure and some nonparametric procedures to Majumdar's (1941) anthropometric data we have shown that the robust procedure gives smaller errors of misclassification than the classical procedure.

2. We have studied the robustness of Balakrishnan and Tiku's (1988) classification procedure based on MML estimators for data consisting of a dichotomous and an associated univariate normal variable. From this study we have found that Balakrishnan and Tiku's (1988) procedure has a smaller average error of misclassification than Chang and Afifi's (1974) classical procedure under departures from normality.

3. We have derived a classification procedure for data consisting of a dichotomous and an associated multivariate normal variable using the multivariate MML estimators. We have performed a robustness study of this procedure and the corresponding classical procedure and have found that the errors of misclassification of the robust procedure to be smaller under departures from normality.

4. We have derived the distribution function, density function, and expected values of errors of misclassification for the classical classification procedure for data consisting of $k-$variate dichotomous and an associated univariate normal variable when both errors of misclassifications are allowed to float. Then we have carried out a similar study assuming that the underlying distributions are outlier-normal and mixture normal. Using these results, for the case $k = 2$, we have illustrated the non-robust characteristic of the classical procedure for departures from normality.

5. We have derived a robust classification procedure based on MML estimators for samples with $k-$variate dichotomous and an associated univariate normal variables. We have carried out a comparative study of this procedure with the corresponding classical procedure for the case $k = 2$ and have shown that the procedure proposed here is robust to departures from normality.

6. We have considered the classification procedures considered in Chapters 3, 4, 5, and 6 in a very general setup, namely, when the data contains a multivariate dichotomous and an associated multivariate normal variables. We have proposed a robust classification procedure based on MML estimators and have

studied the asymptotic properties of this procedure.

7. We have studied the robustness of the classical linear univariate classification procedure under heterogeneity of variances. We have derived the distribution function, density function and expected values of errors of misclassification when both errors are allowed to float. We have carried out a similar study assuming that the underlying distribution contains some outliers. From this study we have found that the errors of misclassification tend to become large as variances of the two populations get apart. In addition to the heterogeneity of variances, if the samples contain some outliers, the errors of misclassification tend to become very large.

8. We have derived robust classification procedures based on MML estimators for data consisting of a univariate dichotomous and an associated normal variable with unequal conditional variances for the two populations. We have derived two classification procedures based on MML estimators, one being the quadratic classification procedure and the other being transformed linear classification procedure. From our robustness study we have found that the procedures based on MML estimators perform better than the corresponding classical procedures under departures from normality and the transformed linear classification procedure based on MML estimators being the best.

9. We have directly adopted the MML approach for the likelihood function based on the Type II symmetrically censored sample of normalized residuals and have derived the MMLEs of the parameters in a simple linear and multiple linear regression models. We have derived approximate expressions for the variances and covariances of these estimators via the information matrix and have established the asymptotic distribution of these estimators. Also we have compared MMLE procedure with some other prominent robust estimation procedures for departures from normality and for departures from linearity and have found that the MMLE procedure is the best.

12.2 Possible further research

Several problems of interest in the areas of classification and regression have been studied in this thesis as explained already in the last section. This study is by no means complete since the developments in this thesis have brought out some more problems that are worth considering for future research. Of special interest among these problems are the problems that we list below:

1. The quadratic and transformed linear classification procedures developed and discussed in detail in chapter 9 need to be extended to the most general situation where the classification problem is based on data consisting of a k-variate dichotomous variable and an associated $p-$variate continuous variable with the assumption that the conditional variance-covariance matrices of the two populations to be unequal. These classification procedures need to be proposed for the case when one of the two errors of misclassification is at a pre-fixed level and also for the case when both the errors are allowed to float.

2. All the discussions regarding classification in this thesis have been about two-way classification; that is, the new observation has to belong to one of the two populations. However, it is quite common to have more than two groups in the classification problem. For example, Fisher's iris data consist of four measurements on three types of iris flowers- setosa, versicolor, and virginica. Similarly, the anthropological example considered in Chapter 2 consist of several populations (23 to be specific). It, therefore, will be of interest to extend the classical and the robust classification procedures to the multi-way classification situation.

3. In the classification problem based on data consisting of a k-variate dichotomous variable and an associated $p-$variate normal variable (see Chapter 7), in the case when all the parameters in the model are unknown one has to estimate 2^{k+1} cell probabilities (i.e. $p_{\mathbf{x}}^{(1)}, p_{\mathbf{x}}^{(2)}$ for 2^k cells), and 2^{k+1} mean vectors and 2^k variance-covariance matrices (with the assumption of the equality of

conditional variance-covariance matrices of the two populations). For example, when $k = 3$ one has to estimate 16 cell probabilities, 16 mean vectors of dimension $p \times 1$, and 8 variance-covariance matrices of dimension $p \times p$. It is very probable that in small sample sizes the number of observations in some cells will be either very small or zero. Hence, the corresponding estimates will at best be poor, and at worst be unobtainable. For the classification rule to be of some practical use, it is therefore necessary to seek an approximation which will yield parameter estimates for all possible cells. Krzanowski (1975) has used contingency tables to represent the dichotomous variables and the log-linear model to estimate the parameters. Krzanowski (1975) has also used the multiple linear regression models to estimate the parameters corresponding to the continuous variables. Both these estimation methods, however, have been developed by assuming the underlying distribution to be normal and hence may lack robustness features. On the other hand, the classification procedures and the method of estimation for the linear regression model developed in this thesis based on the MML estimators are both efficient and robust to departures from normality. As a result, it will be of interest to examine the applicability of these techniques in estimating the parameters in log-linear models since we already proposed in this thesis MML method of estimation for parameters in a multiple linear regression model. It will also be of interest to see whether these two methods could be combined in order to come up with a robust classification procedure analogous to the one proposed by Krzanowski (1975).

4. It is readily seen that the MML estimation method discussed in Chapters 10 and 11 for the simple and the multiple linear regression models, respectively, can be adopted to the case when the data consist of multiple measurements at each level of the explanatory variables. It will be of interest then to compare this estimation procedure to the trimmed and Winsorized estimation methods given by Leone and Moussa-Hamouda (1973) and Moussa-Hamouda and Leone (1974, 1977a, 1977b, 1977c) and a somewhat different MML estimation method given by Tiku (1980b) and Tiku, Tan and Balakrishnan (1986).

Appendix A

Some Results for Normal Integrals

In this appendix we prove some results concerning univariate and bivariate normal distributions. These results will be used repeatedly in this thesis. For this purpose, we exclusively use the following formulae given by Owen (1980):

$$\int_{-\infty}^{\infty} \phi(x)\Phi[a+bx]dx = \Phi\left[\frac{a}{\sqrt{1+b^2}}\right], \tag{1.1}$$

$$\int_{-\infty}^{\infty} x\phi(x)\Phi[a+bx]dx = \frac{b}{\sqrt{1+b^2}}\phi\left[\frac{a}{\sqrt{1+b^2}}\right], \tag{1.2}$$

$$\int_{-\infty}^{y} \phi(x)\Phi[a+bx]dx = H\left[\frac{a}{\sqrt{1+b^2}}, y; -\frac{b}{\sqrt{1+b^2}}\right]. \tag{1.3}$$

In the above formulae, $\Phi(a)$ denotes the cumulative distribution function of a standard normal variable and $H(a,b;\rho)$ denotes the bivariate cumulative distribution function of a standard bivariate normal variable with correlation coefficient ρ.

Result 1

$$\int_{-\infty}^{\infty} x\phi(a+bx)\Phi(c+dx)dx = \frac{d}{b^2\sqrt{b^2+d^2}}\phi\left(\frac{bc-da}{\sqrt{b^2+d^2}}\right) - \frac{a}{b^2}\Phi\left(\frac{bc-da}{\sqrt{b^2+d^2}}\right) \tag{1.4}$$

Proof: Let $u = a + bx$, then

$$\begin{aligned}\int_{-\infty}^{\infty} x\phi(a+bx)\Phi(c+dx)dx &= \int_{-\infty}^{\infty} \frac{(u-a)}{b}\phi(u)\Phi(c-\frac{da}{b}+\frac{d}{b}u)\frac{du}{b} \\ &= \int_{-\infty}^{\infty} u\phi(u)\Phi(c-\frac{da}{b}+\frac{d}{b}u)\frac{du}{b^2} -\end{aligned}$$

$$\int_{-\infty}^{\infty} \phi(u)\Phi(c - \frac{da}{b} + \frac{d}{b}u)a\frac{du}{b^2}.$$

By using (1.1) and (1.2) we obtain the expression on the right-hand side of (1.4).

Result 2

$$\frac{d}{dz}H[k(z), a; \rho] = k'(z)\phi(k(z))\Phi\left[\frac{a - \rho k(z)}{(1-\rho^2)^{\frac{1}{2}}}\right]$$

Proof: Since

$$H[k(z), a; \rho] = \frac{1}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{k(z)}\int_{-\infty}^{a} \exp\left[-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right] dydx$$

we have

$$\begin{aligned}\frac{d}{dz}H[k(z), a; \rho] &= k'(z)\frac{1}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{a} \exp\left[-\frac{k(z)^2 - 2\rho k(z)y + y^2}{2(1-\rho^2)}\right] dy \\ &= k'(z)\frac{1}{2\pi\sqrt{1-\rho^2}}\phi(k(z))\int_{-\infty}^{a} \exp\left[-\frac{1}{2}\left(\frac{y - \rho k(z)}{\sqrt{1-\rho^2}}\right)^2\right] dy\end{aligned}$$

Hence the result.

Result 3 If the joint distribution of (U, V) is $BVN[\mu_u, \mu_v; \sigma_u^2, \sigma_v^2; \rho]$, then

$$\Pr\{U < 2X, V < 0 | X \sim N(\mu, \sigma^2)\} = H\left[\frac{2\mu - \mu_u}{(4\sigma^2 + \sigma_u^2)^{1/2}}, -\frac{\mu_v}{\sigma_v}; -\frac{\rho\sigma_u}{(4\sigma^2 + \sigma_u^2)^{1/2}}\right] \tag{1.5}$$

and

$$\Pr\{U \geq 2X, V < 0 | X \sim N(\mu, \sigma^2)\} = H\left[-\frac{2\mu - \mu_u}{(4\sigma^2 + \sigma_u^2)^{1/2}}, -\frac{\mu_v}{\sigma_v}; -\frac{\rho\sigma_u}{(4\sigma^2 + \sigma_u^2)^{1/2}}\right] \tag{1.6}$$

Proof: Using the total probability rule, we can write

$$\begin{aligned}&\Pr\{U < 2X, V < 0 | X \sim N(\mu, \sigma^2)\} \\ &= \int_{-\infty}^{\infty} Pr\left\{\frac{U-\mu_u}{\sigma_u} < \frac{2x-\mu_u}{\sigma_u}, \frac{V-\mu_v}{\sigma_v} < -\frac{\mu_v}{\sigma_v}\right\}\phi(x; \mu, \sigma^2)dx,\end{aligned}$$

where $\phi(x; \mu, \sigma^2)$ is the density function of a $N(\mu, \sigma^2)$ variable. So, we get

$$Pr\{U < 2X, V > 0 | X \sim N(\mu, \sigma^2)\} = \int_{-\infty}^{\infty} H\left[\frac{2x - \mu_u}{\sigma_u}, -\frac{\mu_v}{\sigma_v}; \rho\right]\phi(x, \mu, \sigma^2)dx \tag{1.7}$$

Now, in order to use (1.3), let us set

$$\begin{aligned} \frac{2x-\mu_u}{\sigma_u} &= \frac{a}{\sqrt{1+b^2}}, \\ -\frac{\mu_v}{\sigma_v} &= y, \\ \text{and} \quad \rho &= -\frac{b}{\sqrt{1+b^2}}; \end{aligned}$$

solving for a and b, we get

$$\begin{aligned} a &= \frac{2x-\mu_u}{\sigma_u\sqrt{1-\rho^2}} \\ \text{and} \quad b &= \frac{\rho}{\sqrt{1-\rho^2}}. \end{aligned}$$

So, from (1.3) we immediately have

$$H\left[\frac{2x-\mu_u}{\sigma_u}, -\frac{\mu_v}{\sigma_v}; \rho\right] = \int_{-\infty}^{\frac{-\mu_x}{\sigma_v}} \phi(t)\Phi\left[\frac{2x-\mu_u+\rho t\sigma_u}{\sigma_u\sqrt{1-\rho^2}}\right] dt.$$

By substituting this expression in Eq. (1.7), we get

$$\begin{aligned} &Pr\{U < 2X, V < 0 | X \sim N(\mu, \sigma^2)\} \\ &= \int_{-\infty}^{\infty} \phi(x, \mu, \sigma^2) \int_{-\infty}^{-\frac{\mu_x}{\sigma_v}} \Phi\left(\frac{2x-\mu_u+\rho t\sigma_u}{\sigma_u\sqrt{1-\rho^2}}\right) \phi(t) dx dt, \end{aligned}$$

which upon interchanging the order of integration gives

$$\begin{aligned} &Pr\{U < 2X, V > 0 | X \sim N(\mu, \sigma^2)\} \\ &= \int_{-\infty}^{-\frac{\mu_x}{\sigma_v}} \phi(t) \left\{\int_{-\infty}^{\infty} \phi(x; \mu, \sigma^2)\Phi\left(\frac{2x-\mu_u+\rho t\sigma_u}{\sigma_u\sqrt{1-\rho^2}}\right) dx\right\} dt. \end{aligned}$$

Setting $y = \frac{x-\mu_x}{\sigma}$, we get

$$\begin{aligned} &Pr\{U < 2X, V < 0 | X \sim N(\mu, \sigma^2)\} \\ &= \int_{-\infty}^{-\frac{\mu_x}{\sigma_v}} \phi(t) \left\{\int_{-\infty}^{\infty} \phi(y)\Phi\left(\frac{2y\sigma+2\mu-\mu_u+\rho t\sigma_u}{\sigma_u\sqrt{1-\rho^2}}\right) dy\right\} dt. \end{aligned} \tag{1.8}$$

Now, applying (1.1) with

$$\begin{aligned} a &= \frac{2\mu-\mu_u+\rho t\sigma_u}{\sigma_u\sqrt{1-\rho^2}} \\ &\text{and} \\ b &= \frac{2\sigma}{\sigma_u\sqrt{1-\rho^2}}, \end{aligned}$$

we get

$$\int_{-\infty}^{\infty} \phi(y)\Phi\left(\frac{2y\sigma + 2\mu - \mu_u + \rho t\sigma_u}{\sigma_u\sqrt{1-\rho^2}}\right) dz = \Phi\left(\frac{2\mu - \mu_u + \rho t\sigma_u}{\sqrt{4\sigma^2 + \sigma_u^2(1-\rho^2)}}\right),$$

which when substituted in Eq. (1.8) gives

$$\begin{aligned} & Pr\{U < 2X, V < 0 | X \sim N(\mu, \sigma^2)\} \\ = & \int_{-\infty}^{-\frac{\mu_x}{\sigma_x}} \phi(t)\Phi\left(\frac{2\mu_x - \mu_u + \rho t\sigma_u}{\sqrt{4\sigma^2 + \sigma_u^2(1-\rho^2)}}\right) dt. \end{aligned}$$

Upon using (1.3) with

$$\begin{aligned} a &= \frac{2\mu_x - \mu_u}{\sqrt{4\sigma^2 + \sigma_u^2(1-\rho^2)}} \\ \text{and } b &= \frac{\rho\sigma_u}{\sqrt{4\sigma^2 + \sigma_u^2(1-\rho^2)}} \end{aligned}$$

and realizing that

$$\begin{aligned} \frac{b}{\sqrt{1+b^2}} &= \frac{\rho\sigma_u}{\sqrt{4\sigma^2 + \sigma_u^2}} \\ \frac{a}{\sqrt{1+b^2}} &= \frac{2\mu_x - \mu_u}{\sqrt{4\sigma^2 + \sigma_u^2}} \end{aligned}$$

we derive

$$Pr\{U < 2X, V < 0 | X \sim N(\mu, \sigma^2)\} = H\left[\frac{2\mu - \mu_u}{(4\sigma^2 + \sigma_u^2)^{1/2}}, -\frac{\mu_v}{\sigma_v}; \frac{\rho\sigma_u}{(4\sigma^2 + \sigma_u^2)^{1/2}}\right]$$

which is the results presented in (1.5). To prove the second formula of **Result 3**, let us set

$$\begin{aligned} U^* &= -U \\ X^* &= -X \end{aligned}$$

then (U^*, V) is $BVN[-\mu_u, \mu_v; \sigma_u, \sigma_v; -\rho]$. Then, we can write

$$\Pr[U \geq 2X, V < 0 | X \sim N(\mu, \sigma^2)] = \Pr[U^* \leq 2X^*, V < 0 | X^* \sim N(-\mu, \sigma^2)]$$

From (1.5), the right hand-side is equal to

$$H\left[\frac{-2\mu+\mu_u}{(4\sigma^2+\sigma_u^2)^{1/2}}, -\frac{\mu_v}{\sigma_v}; \frac{\rho\sigma_u}{(4\sigma^2+\sigma_u^2)^{1/2}}\right],$$

which gives the result presented in (1.6).

Result 4 Let the joint distribution of (U,V) be a mixture of bivariate normal i.e. the pdf of (U,V) is

$$h(u,v)=\sum_r^{n_1}\sum_s^{n_2} f(r,s)BVN[u,v;\mu_u,\mu_v;\sigma_u^2,\sigma_v^2;\rho] \tag{1.9}$$

where $f(r,s)$ is a function of r,s which satisfies

$$\sum_r^{n_1}\sum_s^{n_2} f(r,s)=1.$$

Let X be a univariate mixture normal random variable independent of (U,V), i.e. the pdf of X is

$$g(x)=\sum_{i=1}^{n} p_i\phi(x;\mu_x,\sigma_x^2) \tag{1.10}$$

with $\sum_{i=1}^n p_i=1$. Then

$$P[U<2X,V<0|X\sim g(x)]=\sum_r^{n_1}\sum_s^{n_2}\sum_i^{n} f(r,s)p_iH\left[\frac{2\mu_x-\mu_u}{(4\sigma_x^2+\sigma_u^2)^{1/2}}, -\frac{\mu_v}{\sigma_v}; \frac{\rho\sigma_u}{(4\sigma_x^2+\sigma_u^2)^{1/2}}\right] \tag{1.11}$$

We may note here that $\mu_u,\mu_v,\sigma_u^2,\sigma_v^2$ and ρ may depend on r and s, and similarly μ_x and σ_x^2 may depend on i.

Proof: From the total probability rule, we can write

$$\begin{aligned}&P[U<2X,V<0|X\sim g(x)]\\ &=\sum_r^{n_1}\sum_s^{n_2}\sum_i^{n} f(r,s)p_i\int_{-\infty}^{\infty} BVN[2x,0;\mu_u,\mu_v,\sigma_u^2,\sigma_v^2;\rho]\phi(x;\mu_x,\sigma_c^2x)dx\end{aligned}$$

Note that all the summations are finite and so the summation and integration are interchangeable. Result then follows by applying **Result 3**.

Bibliography

[1] Adichie, J.N. (1967a). Asymptotic efficiency of a class of non-parametric tests for regression parameters, **Ann. Math. Statist. 38**, 884-893.

[2] Adichie, J.N. (1967b). Estimates of regression parameters based on rank tests, **Ann. Math. Statist. 38**, 894-903.

[3] Aitkin, M. and Tunnicliffe Wilson, G. (1980). Mixture models, outliers, and the EM algorithm. **Technometrics 22**, 325-331.

[4] Amoh, R.K. (1983). Classification procedures associated with the Inverse Gaussian Distribution, **Ph.D. Thesis**, Department of Statistics, University of Manitoba, Canada.

[5] Anderson, J.A. (1972). Separate sample logistic discrimination, **Biometrika 59**, 19-36.

[6] Anderson, T.W. (1951). Classification by multivariate analysis, **Psychometrika 16**, 31-50.

[7] Anderson, T.W. (1984). **Introduction to Multivariate Statistical Analysis**, Second Edition, John Wiley & Sons, New York.

[8] Anderson, T.W. and Bahadur, R. (1962). Classification into two multivariate normal distributions with different covariance matrices, **Ann. Math. Statist. 33**, 420-431.

[9] Anderson, T.W. (1964). On Bayes procedures for a problem with choice of observations, **Ann. Math. Statist. 35**, 1128-1135.

[10] Anderson, T.W. (1966). Some nonparametric multivariate procedures based on statistically equivalent blocks, **Proc. of an International Symposium on Multivariate Analysis**, Academic Press, New York.

[11] Anderson, T.W. (1972a). An asymptotic expansion of the distribution of the "Studentized" classification statistic W, Stanford Univ., Department of Statistics **Tech. Rep. 9**.

[12] Anderson, T.W. (1972b). Asymptotic evaluation of the probabilities of misclassification by linear discriminant functions, Stanford Univ., Department of Statistics **Tech. Rep. 10**.

[13] Anderson, T.W. (1958). **Introduction to Multivariate Statistical Analysis**, John Wiley & Sons, New York.

[14] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W. (1972). **Robust Estimates of Location: Survey and Advances**, Princeton University Press, Princeton, N.J.

[15] Andrews, D.F. (1973) Robust estimation for multiple linear regression models, **Bull. Int. Statist. Inst. 45**, 105-111.

[16] Andrews, D.F. (1974) A robust method for multiple linear regression, **Technometrics 16**, 523-531.

[17] Aoyama, H. (1950). A note on the classification data, **Ann. Inst. Statist. Math. 2**, 17-20.

[18] Arnold, B.C. and Balakrishnan, N. (1989). **Relations, Bounds and Approximations For Order Statistics**, Lecture Notes in Statistics No. 53, Springer-Verlag, New York.

[19] Bahadur, R.R. (1961). On classification based on responses to N dichotomous items, **Ibid**, 169-176.

[20] Balakrishnan, N. (1988). Bayesian insight into a robust two-sample $t-$test based on asymmetric censored samples, **Biometrical Journal 30**, 425-440.

[21] Balakrishnan, N. and Kocherlakota, S. (1985). Robustness to nonnormality of the linear discriminant function: Mixtures of normal distributions, **Commun. Statist.** - Theor. Meth. **14**(2). 465-478.

[22] Balakrishnan, N., Kocherlakota, S., and Kocherlakota, K. (1986). On the errors of misclassification based on dichotomous and normal variables, **Ann. Inst. Statist. Math. 38**, 529-538.

[23] Balakrishnan, N., Kocherlakota, S., and Kocherlakota, K. (1988). Robustness of the double discriminant function in nonormal situations, **S. Afr. Statist. J. 22**, 15-43.

[24] Balakrishnan, N. and Tan, W.Y. (1988). Bayesian insight into a robust test for linear contrast based on asymmetric censored samples, **Biometrical Journal 30**, 517-532.

[25] Balakrishnan, N. and Tiku, M.L. (1988a). Robust classification procedures, In **Classification and Related Methods of Data Analysis** (Ed., H.H. Bock). pp. 269-276, North-Holland, Amsterdam.

[26] Balakrishnan, N. and Tiku, M.L. (1988b). Robust classification procedures based on dichotomous and continuous variables, **Journal of Classification 5**, 53-80.

[27] Balakrishnan, N., Tiku, M.L., and El Shaarawi, A.H. (1985). Robust univariate two-way classification, **Biometrical Journal 27**, 123-138.

[28] Balakrishnan, N. and Cohen, A.C. (1990). Order Statistics and Inference: Estmation Methods, Academic Press, Boston.

[29] Banerjee, K.S. and Marcus, L. F. (1965). Bounds in a minimax classification procedure, **Biometrika 52**, 653-654.

[30] Barnard, M.M. (1935). The secular variations of skull characters in four series of Egyptian skulls, **Ann. Eug. 6**, 352-371.

[31] Barnard, E. (1991). Performance and generalization of the classification figure of merit criterion function, **IEEE Transaction on neural networks 2(2)**, 322-325.

[32] Barnett, V., and Lewis, T. (1978). **Outliers in Statistical Data**, John Wiley & Sons, Chichester, England.

[33] Barrodale, I. (1968). L_1 approximations and the analysis of data, **Appl. Statist. 17**, 51-57.

[34] Barrodale, I. and F.D.K. Roberts (1973). An improved algorithm for discrete L_1 linear approximation, **SIAM J. Numer. Anal., 10**, 839-848.

[35] Bartlett, M.S. and Please, N.W. (1963). Discrimination in the case of zero mean differences, **Biometrika 50**, 17-21.

[36] Beaton, A.E. and J.W. Tukey (1974). The fitting of power series, meaning polynomials, illustrated on band spectroscopic data, **Technometrics, 16**, 147-185.

[37] Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). **Regression Diagnostics**, John Wiley & Sons, New York.

[38] Bhattacharya, D.P. and Narayan, R.D. (1941). Moments of the D^2- statistic for populations with unequal dispersions, **Sankhya 5**, 401- 412.

[39] Bhattacharya, P.K. and Das Gupta, S. (1964). Classification into exponential populations, **Sankhya A 26**, 17-24.

[40] Bhattacharyya, P.K. (1985). The asymptotics of maximum likelihood and related estimators based on Type II censored data, **J. Amer. Statist. Assoc. 80**, 398-404.

[41] Book, D., J. Booker, H. O. Hartley, and R.L. Sielken, Jr. (1980). Unbiased L_1 Estimators and Their Covariances, bf ONR THEMIS Technical Report No. 64, Institute of Statistics, Texas A&M University.

[42] Bose, R.C. (1936a). On the exact distribution and moment coefficients of the D^2-statistic, **Sankhya 2**, 143-154.

[43] Bose, R. C. (1936b). A note on the distribution of differences in mean values of two samples drawn from two multivariate normally distributed populations, and the definition of the D^2-statistic, **Sankhya 2**, 379- 384.

[44] Bose, S.N. (1936c). On the complete moment coefficients of the D^2-statistic, **Sankhya 2**, 385-396.

[45] Bose, R.C. and Roy, S.N. (1938). The distribution of the studentized D^2-statistic, **Sankhya 4**, 19-38.

[46] Bowker, A.H. (1960). A representation of Hotelling's T^2 and Anderson's classification statistic, **Contrib. Probability and Statistics** (Hotelling Vol.). 142-149.

[47] Bowker, A.H. (1960). A representation of Hotteling's T^2 and Anderson's classification statistic W in terms of simple statistics, In **Contributions to Probability and Statistics: Essays in Honour of H. Hotelling**, Stanford University Press, Stanford, California.

[48] Bowker, A.H. and Sitgreaves, R. (1961). An asymptotic expansion for the distribution function of the $W-$classification statistic, **Ibid**, 293- 310.

[49] Bowker, A.H. (1961). A representation of Hotelling's T^2 and Anderson's classification statistic W in terms of simple statistics, **Ibid**, 285-292.

[50] Broffit, J.D. (1969). Estimating the probability of misclassification based on discriminant function techniques, Ph.D. dissertation, Colorado State Univ., Fort Collins.

[51] Brown, G.W. (1947). Discriminant functions, **Ann. Math. Statist. 18**, 514-528.

[52] Bunke, O. (1964). Uber optimale verfahern der diskrinazanalyse, **Abl. Deutsch. Akad. Wiss. Klasse Math. Phys. Tech. 4**, 35-41 (MR 32-6624).

[53] Bunke, O. (1966). Nichparametrishe Klassifikations verfahren fur qualitative und quantitative Beobachtungen, **Wiss Z. Humboldt Univ. Berlin Math. Naturwiss. Reihe 15**, 15-18 (MR 36-1031).

[54] Bunke, O. (1967). **Z. Wahr Theor. und Verwundte Gebiete** (MR 35-6267)

[55] Burnaby, T.P. (1966). Growth invariant discriminant functions and generalized distances, **Biometrics 22**, 96-110.

[56] Cacoullos, T. (1964). Estimation of a multivariate density, **Technical Report No. 40**, Department of Statistics, University of Minnesota, Minneapolis.

[57] Cacoullos, T. (1965a). Comparing Mahalanobis distances, I: Comparing distances between k normal populations and another unknown, **Sankhya A 27**, 1-22.

[58] Cacoullos, T. (1965b). Comparing Mahalanobis distances, II: Bayes procedures when the mean vectors are unknown, **Sankhya A 27**, 23-32.

[59] Cacoullos, T. (1966). On a class of admissible partitions, **Ann. Math. Statist. 37**, 189-195.

[60] Cavalli, L. L. (1945). Alcuni problemi della analisi biometrica di popolazioni naturali, **Mem. Ist. Ital. Idrobiol. 2** 301-323.

[61] Chadda, R.L. and Marcus, L.F. (1968). An empirical comparison of distance statistics for populations with unequal covariance matrices, **Biometrics 24**, 683-.

[62] Chan, L.S. (1970). The treatment of missing values in discriminant analysis, **Doctoral Dissertation**, University of California, Los Angeles.

[63] Chan, L.S. and Dunn, O.J. (1972). The treatment of missing values in discriminant analysis I, **J. Amer.. Statist. Assoc. 67**, 433-477.

[64] Chang, P.C., and Afifi, A.A. (1974). Classification based on dichotomous and continuous variables, **J. Amer. Statist. Assoc. 69**, 336-339.

[65] Chernoff, H. (1956). A classification problem, Stanford University, Department of Statistics **Technical Report 33**.

[66] Ching'anda, E.F. (1976). Misclassification errors and their distributions, **M.Sc. Thesis**, University of Manitoba, Winnipeg.

[67] Ching'anda, E.F. and Subrahmaniam, K. (1979). Robustness of the linear discriminant function to nonnormality: Johnson's system, **J. Statist. Plann. Inf. 3**, 69-77.

[68] Clunies-Ross, C.W. and Riffenburgh, R.H. (1960a). Linear discriminant analysis, **Pacif. Sci. 14**, 251-256.

[69] Cochran, W.G. (1961). On the performance of the linear discriminant function, **Bull. Intl. Statist. Inst. 34**, 436-446.

[70] Cochran, W.G. (1964). Comparison of two methods of handling covariates in discriminatory analysis, **Ann. Inst. Statist. Math. 16**, 43- 53.

[71] Cochran, W.G. (1966). Analyse des classifications d'ordre, **J. Roy. Statist. Appl. 14**, 5-17.

[72] Cochran, W.G. and Hopkins, C. (1961). Some classification methods with multivariate qualitative data, **Biometrics 17**, 10-32.

[73] Cohen, A.C. (1959) Simplified estimators for the normal distribution when samples are singly censored or truncaed, **Technometric 1**, 217-237.

[74] Cohen, A.C. (1961) Tables for maximum likelihood estimates: singly truncated and singly censored samples, **Technometric 3**, 535-541.

[75] Cooper, P.W. (1962a). The hyperplane in pattern recognition, **Cybernetica 5**, 215-238.

[76] Cooper, P.W. (1962b). The hypersphere in pattern recognition, **Inform. Control 5**, 324-346.

[77] Cooper, P.W. (1963). Statistical classification with quadratic forms, **Biometrika 50**, 439-448.

[78] Cooper, P.W. (1963). Statistical classification with quadratic forms, **Biometrika 50**, 439-448.

[79] Cooper, P.W. (1965). Quadratic discriminant functions in pattern recognition, **IEEE Trans. Inform. Theory IT-11**, 313-315.

[80] Crooke, W. (1896). **Castes and Tribes of North Western Provinces and Oudh**, Vols. 1-6.

[81] Daniel, C. and Wood, F.S. (1971). **Fitting Equations to Data**, Second edition, John Wiley & Sons, New York.

[82] Das Gupta, S. (1962). On the optimum properties of some classification rules, **Inst. Statist. Mimeo,**, No. 333. Univ. of N.C., Chapel Hill.

[83] Das Gupta, S. (1964). Nonparametric classification rules, **Sankhya**, Ser. A **26**, 25-30.

[84] Das Gupta, S. (1965). Optimum classification rules for classification into two multivariate normal populations, **Ann. Math. Statist. 36**, 1174-1184.

[85] Das Gupta, S. (1972). Probability inequalities and error in classification, Univ. of Minnesota, School of Statistics **Tech. Rep. 190**.

[86] Das Gupta, S. (1973). Theories and Methods in Classification: A review **Proc. of a NATO Advanced Study Institute on Discriminant Analysis and Applications, Athens, Greece, June 8-20, 1972** (Ed. T. Cacoullos,). Academic Press, New Yoark and London.

[87] David, F.N. and Johnson, N.L. (1954). Statistical treatment of censored data. I. Fundamental formulae, **Biometrika 41**, 228-240.

[88] David, H.A. (1981). **Order Statistics**, Second edition, John Wiley & Sons, New York.

[89] Day, B.B. and Sandomire, M. (1942). Use of the discriminant function for more than two groups, **Jour. Amer. Statist. Assoc. 37**, 461- 472.

[90] Day, N.E. and Kerridge, D.F. (1967). A general maximum likelihood discriminant, **Biometrics 23**, 313-323.

[91] Day, N.E. and Kerridge, D.F. (1967). A general maximum likelihood discriminant, **Biometrics 23**, 313-323.

[92] Dixon, W.J. (1960). Simplified estimation from censored normal samples, **Ann. Math. Statist. 31**, 385-391.

[93] Dixon, W.J. and Tukey, J.W. (1968). Approximate behavior of the distribution of Winsorized t (trimming/Winsorization 2). **Technometrics 10**, 83-98.

[94] Duda, R.O. and Hart, P.E. (1973). **Pattern Classification and Scene Analysis**, Wiley, New York.

[95] Dunn, O.J. and Varady, P.D. (1966). Probabilities of correct classification in discriminant analysis, **Biometrics 22**, 908-924.

[96] Dunn, O.J. (1971). Some expected values for probabilities of correct classification in discriminant analysis, **Technometrics 13**, 345- 353.

[97] Elashoff, J. D., Elashoff, R.M. and Goldman, G.E. (1967). On the choice of variables in classification problems with dichotomous variables, **Biometrika 54**, 668-670.

[98] Elfving, G. (1961). An expansion principle for distribution functions with applications to Student's t-statistic and the one-dimensional classification statistic, **Ibid**, 276-284.

[99] Ellison, B.E. (1962). A classification problem in which information about alternative distributions is based on samples, **Ann. Math. Statist. 33**, 213-223.

[100] Ellison, B.E. (1965). Multivariate normal classification with covariance known, **Ann. Math. Statist. 36**, 1787-1793.

[101] Everitt, B.S. (1978). **Graphical Techniques in Multivariate Analysis**, Heinemann Educational Books, London.

[102] Fu, K.S. (1968). Rates of convergence for nearest neighbour procedures, **Proc. Hawaii Internat. Conf. On System Sciences**, 413-415.

[103] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, **Ann. Eug. 7**, 179-188.

[104] Fisher, R. A. (1938). The statistical utilization of multiple measurements, **Ann. Eug. 8**, 376-386.

[105] Fix, E. and Hodges, J.L., Jr. (1951). Discriminating analysis; Nonparametric discrimination; Consistency properties, **USAF School of Aviation Medicine**, Report No. 4.

[106] Friedman, H.D. (1965). On the expected error in the probability of misclassification, **Proc. IEEE 53**, 658-659.

[107] Forsythe, A.B. (1972) Robust estimation of straight line regression coefficients by minimizing $p-$th power deviations, **Technometrics 14**, 159-166.

[108] Gaul, W. and Schader M. (Eds) (1986). Classification as a tool of research, **Amsterdam: North Holland.**

[109] Geisser, S. (1967). Estimation associated with linear discriminant, **Ann. Math. Statist. 38**, 807-817.

[110] Geisser, S. (1964). Posterior odds for multivariate normal classification, **J. Roy. Statist. Soc. Ser. B 26**, 69-76.

[111] Geisser, S. (1966). Predictive discrimination, **Proc. Internat. Symp. Multiv. Analysis** (P.R. Krishnaiah, Ed.). Academic Press: New York.

[112] Gentle, J.M., W. J. Kennedy, and V.A. Sposito (1977). On least absolute deviations estimators, **Commun. Statist., A6**, 839-845.

[113] Gessaman, M.P. (1970). A consistent nonparametric multivariate density estimator based on statistically equivalent blocks, **Ann. Math. Statist. 41**, 1344-1346.

[114] Gessaman, M.P. and Gessaman, P.H. (1972). A comparison of some multivariate discrimination procedures, **J. Amer. Statist. Assoc. 67**, 468-472.

[115] Geva, S. and Sitte, J. (1991). Adaptive nearest neighbor pattern classification, **IEEE Transactions on neural networks 2(2)**, 318-322.

[116] Gilbert, E. (1968). On discrimination using qualitative variables, **J. Amer. Statist. Assoc. 63**, 1399-1412.

[117] Gilbert, E.S. (1969). The effect of unequal variance-covariance matrices on Fisher's linear discriminant function, **Biometrics 25**, 505-515.

[118] Glick, N. (1969). Estimating unconditional probabilities of correct classification, Stanford Univ. Department of Statistics, **Tech. Report 3.**

[119] Gnanadesikan, R. (1977). **Methods for Statistical Data Analysis of Multivariate Observations**, John Wiley & Sons, New York.

[120] Goodwin, C.N. and Morant, G.M. (1940). The human remains of Iron Age and other periods from Maiden Castle, Dorset, **Biometrika 31**, 295-

[121] Govindarajulu, Z. and Gupta, A.K. (1972). Some classification rules for c univariate normal populations, Univ. of Michigan, Department of Statistics **Tech. Rep. 14**.

[122] Grayson, D.A. (1988). Two-group classification in latent trait theory: scores with monotone likelihood ratio, **Psychometrika 53(3)**, 383-392.

[123] Grenander, U. (1972). Asymptotic distribution of quadratic forms: large deviations (unpublished)

[124] Gupta, A.K. and Kim, B.K. (1978). On a distribution-free discriminant analysis, **Biometrical Journal 20**, 729-736.

[125] Hampel, F. R. (1977). Rejection rules and robust estimates of location: An anlysis of some Monte Carlo results. **Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes**, 1974 European Meeting of Statisticians, Vol. A, pp. 187-194.

[126] Han, Chien Pai (1970). Distribution of discriminant function in circular models, **Ann. Inst. Statist. Math. 22**, 117-125.

[127] Han, Chien Pai (1968). A note on discrimination in the case of unequal covariance matrices, **Biometrika 55**, 586-587.

[128] Han, Chien Pai (1969). Distribution of discriminant function when covariance matrices are propositional, **Ann. Math. Statist. 40**, 979-985.

[129] Harter, H. L. (1951). On the distribution of Wald's classification statistic, **Ann. Math. Statist. 22**, 58-67.

[130] Harter, H.L. and Moore, A.H. (1966) Iterative maximum likelihood estimation of the parameters of normal populations from singly and doubly censored samples, **Biometrika 53**, 205-213.

[131] Hills, M. (1966). Allocation rules and their error rates **J. Roy. Statist. Soc. Ser. B 28**, 1-31.

[132] Hills, M. (1967). Discrimination and allocation with discrete data, **Applied Statist. 16**, 237-250.

[133] Hodges, J.L. (1950)). Survey of discriminatory analysis, **USAF School of Aviation Medicine, Rep. No. 1**, Randolph Field, Texas.

[134] Hogg, R.V. and R.H. Randles (1975). Adaptive distribution-free regression methods and their applications, **Technometrics, 17**, 399-407.

[135] Hotelling, H. (1931). The generalization of Student's ratio, **Ann. Math. Statist. 2**, 360-378.

[136] Huber, P.J. (1964). Robust estimation of a location parameter, **Ann. Math. Statist. 35**, 73-101.

[137] Huber, P.J. (1981). **Robust Statistics**, John Wiley & Sons, New York.

[138] Hudimoto, H. (1964). On a distribution-free two-way classification, **Ann. Inst. Statist. Math. 16**, 247-253.

[139] Jackson, E.C. (1968). Missing values in linear multiple discriminant analysis, **Biometrics 24**, 835-844.

[140] Jaeckel, L.A. (1972). Estimating regression coefficients by minimizing the dispersions of the residuals, **Ann. Math. Statist., 43**, 1449-1458

[141] Jain A.K. (1987) Advances in Statistical Pattern Recognition. **Pattern Recognition Theory and Applications**, Springer-Verlag, Berlin, 1-20.

[142] John, S. (1959). The distribution of Wald's classification statistic when the dispersion matrix is known, **Sankhya 21**, 371-376.

[143] John, S. (1960). On some classification statistics I, II, **Sankhya 22**, 301-308, 309-316.

[144] Johnson, P.O. (1950). The quantification of qualitative data in discriminant analysis, **J. Amer. Statist. Assoc. 45**, 65-76.

[145] Johnn, S. (1961). Errors in discrimination, **Ann. Math. Statist. 32**, 1125-1144.

[146] John, S. (1960a). On some classification problems I, **Sankhya 22**, 301-308.

[147] John, S. (1960b). On some classification statistics, **Sankhya 22**, 309-316.

[148] John, S. (1963). On Classification by the statistics R and Z, **Ann. Inst. Statist. Math. 14**, 237-246. (Correction: **Ann. Inst. Statist. Math. 17** (1965). 113)

[149] Jureckova, J. (1977). Asymptotic relations of M-estimates and R-estimates in linear regression models, **Ann. Statist., 5**, 464-472.

[150] Johnson, N.L., Nixon, E., Amos, D.E., and Pearson, E.S. (1963). Tables of percentage points of Pearson curves, **Biometrika 50**, 459-498.

[151] Kabe, D.G. (1963). Some results on the distribution of two random matrices used in classification procedures, **Ann. Math. Statist. 34**, 181-185.

[152] Kocherlakota, S., Balakrishnan, N., and Kocherlakota, K. (1987). The linear discriminant function: Sampling from the truncated normal distribution, **Biometrical Journal 29**, 131-139.

[153] Kendall, M. G. (1966). Discrimination and classification, **Proc. Internat. Symp. Multiv. Anal.** (P.R. Krishnaiah). 165-185, Academic Press: New York.

[154] Kiefer, J. and Schwartz, R. (1965). Admissible Bayes character of T^2-, R^2- and other full invariant test for classical multivariate normal problems, **Ann. Math Statist. 36**, 747-770.

[155] Kinderman, A. (1972). On some properties of classification: classifiability, asymptotic relative efficiency, and a complete class theorem, Univ. of Minnesota, Department of Statistics ,**Tech. Rep. 178**.

[156] Kocherlakota, S., Kocherlakota, K., and Balakrishnan, N. (1987). Asymptotic expansions for errors of misclassification: Nonnormal situations, In **Advances in Multivariate Statistical Analysis** (Ed., A.K.Gupta). pp. 191-211, D. Reidel Publishing Company.

[157] Krzanowski, W.J. (1975). Discrimination and classification using both binary and continuous variables, **J. Amer. Statist. Assoc. 70**, 782-790.

[158] Kudo, A. (1960). The classification problem viewed as a two- decision problem II, **Mem. Fac. Sci. Kyushu Univ. Ser. A. 14**, 63-83.

[159] Kudo, A. (1959). The classification problem viewed as a two- decision problem I, **Mem. Fac. Sci. Kyushu Univ. Ser. A. 13**, 96-125.

[160] Kullback, S. (1952). An application of information theory to multivariate analysis, **Ann. Math. Statis. 23**, 88-102.

[161] Kullback, S. (1959). **Information Theory and Statistics**, Wiley: New York (1968, Dover: New york).

[162] Lachenbruch, P.A. (1966). Discriminant analysis when the initial samples are misclassified, **Technometrics 8**, 657-662.

[163] Lachenbruch, P.A. and Mickey, M.R. (1968). estimation of error rates in discriminant analysis, **Technometrics 10**, 1-11.

[164] Lachenbruch, P.A. (1967). An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis, **Biometrics 23**, 639-645.

[165] Lachenbruch, P.A. and Mickey, M.R. (1968). Estimation of error rates in discriminant analysis, **Technometrics 10**, 1-11.

[166] Lachenbruch, P.A., Sneeringer, C., and Revo, L.T. (1973). Robustness of the linear and quadratic discriminant functions to certain types of nonnormality, **Commun. Statist. 1**, 39-57.

[167] Lbov, G.S. (1964). Errors in the classification of patterns for unequal covariance matrices, **Akad. Nauk. SSSR Sibirisk Otdel Inst. Mat. Vyc. Sistemy 14**, 31-38 (MR 31-5724).

[168] Linhart, H. (1959). Techniques for discriminant analysis with discrete variables, **Metrika 2**, 138-149 (MR 21-60067).

[169] Linhart, H. (1961). Fur wahl von variablen in der Tremanalsye, **Metrika 4**, 126-139.

[170] Loftsgaarden, D.O. and Quesenberry, C.P. (1965). A non-parametric estimate of a multivariate density function, **Ann. Math. Statist. 36**, 1049-1051.

[171] Mahalanobis, P.C. (1936). On the generalized distance in statistics, **Proc. Nat. Inst. Sci. India 2**, 49-55.

[172] Mahalanobis, P.C. (1948). Anthropological observations. Supplement: Ethnological notes, **Sankhya 9**, 181-236.

[173] Mahalanobis, P.C. and Rao, C.R. (1948). Statistical analysis, **Sankhya 9**, 111-180.

[174] Mahalanobis, P.C. (1930). On tests an measurements of group divergence, **Jour. and Proc. Asiatic Soc. Bengal 26**, 541-588.

[175] Mahalnobis, P.C. (1927). Analysis of race mixture in Bengal, **Jour. and Proc. Asiatic Soc. of Bengal 23**, No. 3.

[176] Majumdar, D.N. (1941). **Collection of U.P. Anthropometric Data in the 1941 Survey.**

[177] Majumdar, D.N. (1944). **The Fortunes of Primitive Tribes**, Lucknow, India.

[178] Marshall, A.W. and Olkin, I. (1968). A general approach to some screening and classification problems, **J. Roy. Statist. Soc. B 30**, 407- 435.

[179] Martin, D.C. and Bradley, R.A. (1972). Probability models, estimation and classification for multivariate dichotomous populations, **Biometrics 28**, 203-222.

[180] Martin, E. S. (1936). A study of the Egyptian series of mandibles with special reference to mathematical methods of sexing, **Biometrika 28**, 149-178.

[181] Maronna, R.A. (1976). Robust M-estimators of multivariate location and scatter, **Ann. Statist. 4**, 51-67.

[182] Matusita, K. (1956). Decision rule, based on the distance, for the classification problem, **Ann. Inst. Statist. Math. 8**, 67-77.

[183] Matusita, K. (1967). Classification based on distance in multivariate Gaussian case, **Proc. 5th Berk. Symp. Math. Stat. Prob. 1**, 299-304, Univ. of California Press, Berkeley.

[184] Mclachan, G.J. (1972). Asymptotic results for discriminant analysis when the initial samples are misclassified, **Technometrics 14**, 415-422.

[185] Memon, A.Z. and Okamoto, M. (1970). The classification statistics W^* in covariate discriminant analysis, **Ann. Math. Statist. 41**, 1491- 1499.

[186] Memon, A.Z. and Okamoto, M. (1971). Asymptotic expansion of the distribution of the $Z-$statistic in discriminant analysis, **J. Multiv. Anal. 1**, 294-307.

[187] Mendenhall, W. (1983). **Introduction To Probability and Statistics**, PWS publication, Boston, MA.

[188] Morant, G.M. (1939). The use of statistical methods in the investigation of problems of classification in anthropology, **Biometrika 31**, 72-98.

[189] Morant, G.M. (1926). A first study of crainology of England and Scotland from neolithic to early historic times, with special reference to Anglo- Saxon skulls in London museums, **Biometrika 18**, 56-.

[190] Morant, G.M. (1928). A preliminary classification of European races based on cranial measurements, **Biometrika 20(B)**, 301-588.

[191] Moussa-Hamouda, E. (1988) Inference in regression problems based on order statistics, **Commun. Statist-Theory Meth. 17(7)**, 2343-2367.

[192] Moussa-Hamouda, E. and Leone, F.C. (1974) The O-Blue Estimators for Complete and Censored Samples in LInear Regression, **Technometrics 16(3)**, 441-446.

[193] Moussa-Hamouda, E. and Leone, F.C. (1977) The Robustness of Efficiency of Adjusted Trimmed Estimators in Linear Regression, **Technometrics 19(1)** 19-34

[194] Moussa-Hamouda, E. and Leone, F.C. (1977) The Effect of Trimmingg and Normality Assumptions on the Efficiency of the O-BLUE Estimators, **Technometric 31(4)**, 158-160.

[195] Moussa-Hamouda, E. and Leone, F.C. (1977) Efficiency of Ordinary Least Squares Estimatoors from Trimmed and Winsorized Samples in Linear Regression, **Technometric 19(3)**, 265-273.

[196] Moussa-Hamouda, E. and Leone, F.C. (1973) Relative Efficiencies of 'O-BLUE' Esttimators in Simple Linear Regression, **Journal of the American Statistical Associatio 68(344)** 953-959.

[197] Mullen, K. and Ennis, D.M. (1987). Mathematical formulation of multivariate euclidean models for discriminant methods, **Psychometrika 52(2)**, 235-249.

[198] Myers, R.H. (1986). **Classical and Modern Regression with Applications**, PWS publication, Boston, MA.

[199] Neyman, J. and Pearson, E.S. (1933a). On the problem of the most efficient tests of statistical hypotheses, **Phil. Trans. Roy. Soc. A, 231**, 281-

[200] Neyman, J. and Pearson, E.S. (1933b). On the testing of statistical hypotheses in relation to probability a priori, **Proc. Camb. Phil. Soc. 9**, 492-.

[201] Neyman, J. and Pearson, E.S. (1936). Contribution to the theory of testing statistical hypotheses, I, **Stat. Res. Memoir 1**, 1-37.

[202] Nishida, N. (1971). A note on the admissible tests and classification in multivariate analysis, **Hiroshima Math. J. 1**, 427-434.

[203] Okamoto, M. (1961). Discrimination for variance matrices, **Osaka Math. Jour. 13**, 1-39.

[204] Okamoto, M. (1963). An asymptotic expansion for the distribution of linear discriminant function, **Ann. Math. Statist. 34**, 1286-1301. Correction **39**, 1358-1359.

[205] Olkin, I. and Tate, R.F. (1961). Multivariate correlation models with mixed discrete and continuous variables, **Ann. Math. Statist. 32**, 448-465.

[206] Owen, D.B. (1980). A table of normal integrals, **Commun. Statist.** Simul. Comput. **B9(4)**, 389-419.

[207] Parzen, E. (1962). On estimation of a probability density and mode, **Ann. Math. Statist. 33**, 1065-1076.

[208] Patrick, E.A. (1972). **Fundamentals of pattern recognition**, Prentice-Hall:Englewood Cliffs, New Jersey.

[209] Pearson, K. (1926). On the coefficient of racial likeness, **Biometrika 18**, 105-117.

[210] Pearson, K. (1898). Mathematical contributions to the theory of evolution. V. On the reconstruction of stature of prehistoric races, **Phil. Transac. A 192**, 169-244.

[211] Penrose, L.S. (1947). Some notes on discrimination, **Ann. Eug., 13**, 228-237.

[212] Persson, T. and Rootzen, H. (1977). Simple and highly efficient estimators for a Type I censored normal sample, **Biometrika 63**, 123-128.

[213] Raiffa, Howard (1961). Statistical decision theory approach to item selection for dichotomous test and criterion variables. **Stud. Item Anal. Predict. (Solomon)**, 187-200.

[214] Rao, C.R. (1946). Tests with discriminant functions in multivariate analysis, **Sankhya 7**, 407-413.

[215] Rao, C.R. (1947a). The problem of classification and distance between two populations, **Nature 159**, 30-31.

[216] Rao, C.R. (1947b). A statistical criterion to determine the group to which an individual belongs, **Nature 160**, 835-836.

[217] Rao, C.R. (1948a). A statistical criterion to determine the group to which an individual belongs, **Nature 160**, 835-836.

[218] Rao, C.R. (1948b). The utilization of multiple measurements in problems of biological classification (with discussion). **J. Roy. Statist. Soc.**, Ser. B, **10**, 159-203.

[219] Rao, C.R. (1949a). On the distance between two populations, **Sankhya 9**, 246-248.

[220] Rao, C.R. (1949b). On some problems arising out of discrimination with multiple characters, **Sankhya 9**, 343-366.

[221] Rao, C.R. (1950). Statistical inference applied to classificatory problems, **Sankhya 10**, 229-256.

[222] Rao, C.R. (1952). **Advanced Statistical Methods in Biometric Research**, Wiley: New York.

[223] Rao, C.R. (1954). A general theory of discrimination when the information on alternative hypotheses is based on samples, **Ann. Math. Statist. 25**, 651-670.

[224] Rao, C.R. (1966). Discriminant function between composite hypothesis and related problems, **Biometrika 53**, 339-345.

[225] Risley, H.H. (1891). **Tribes and Castes of Bengal, Ethnographic Glossary**, Calcutta, India.

[226] Roy, S.C. (1915). **The Oraons of Chota-Nagpur**, Calcutta, India.

[227] Sedranks, N. and Okomoto, M. (1971). Estimation of the probabilities of misclassification for a linear discriminant function in the univariate normal case, **Ann. Inst. Statist. Math. 23**, 419-436.

[228] Shoemaker, P.A. (1991). A note on least-squares learning procedures and classification by neural network models, **IEEE Transactions on neural networks 2(1)**, 158-160.

[229] Sinha, B.K. and Giri, N. (1975). On the distribution of a random matrix, **Commun. Statist. 4**, 1057-1063.

[230] Sitgreaves, R. (1952). On the distribution of two random matrices used in classification procedure, **Ann. Math. Statist. 23**, 263-270.

[231] Sitgreaves, R. (1961). Some results on the distribution of the $W-$classification statistic, **Stud. Itm. Anal. Pred.** (H. Solomon, Ed.). Stanford Univ. Press, Stanford, California, 241-251.

[232] Smith, C.A.B. (1947). Some examples of discrimination, **Annals of Eugenics 13**, 272-282.

[233] Srivastava, M.S. and Khatri, C.G. (1979). **An Introduction to Multivariate Statistics**, North Holland, Amsterdam.

[234] Solomon, H. (1960). Classification procedures based on dichotomous response vectors, **Contrib. Probability and Statistics** (Hotelling Vol.). 414-423.

[235] Solomon, H. (1961). Classification procedures based on dichotomous response vectors, **Stud, Item Anal. Pred.** (H. Solomon, Ed.). Stanford Univ. Press, Stanford, California, 177-186.

[236] Sorum, M.J. (1971). Estimating the conditional probability of misclassification, **Technometrics 13**, 333-343.

[237] Sorum, M. J. (1972b). Estimating the expected and the optimal probabilities of misclassification, **Technometrics 14**, 935-943.

[238] Sorum, M.J. (1972a). Three probabilities of misclassification, **Technometrics 14**, 309-316.

[239] Srivastava, J.N. and Zaatar, M.K. (1972). On the maximum likelihood classification rule for incomplete multivariate samples and its admissibility, **J. Multiv. Anal. 2** 115-126.

[240] Srivastava, M.S. (1967). Classification into multivariate normal populations when the population means are restricted, **Ann. Inst. Statist. Math. 19**, 473-478.

[241] Srivastava, M.S. (1964). Optimum procedures for classification and related topics, Stanford University Department of Statistics **Tech. Report No. 11.**

[242] Srivastava, M.S. and Lee, G.C. (1984). On the distribution of the correlation coefficient when sampling from a mixture of two bivariate normal densities: robustness and outliers, **Canad. J. Statist. 2**, 119-113.

[243] Stoller, D.C. (1954). Univariate two population distribution-free discrimination, **J. Amer. Statist. Assoc. 49**, 770-775.

[244] Subrahmaniam, K. and Ching'anda, E.F. (1978). Robustness of the linear discriminant function to non-normality: Edgeworth series distribution, **J. Statist. Plann. Inf. 2**, 79-91.

[245] Takakura, S. (1962). Some statistical methods of classification by the theory of quantification, **Proc. Inst. Statist. Math. Tokyo 9**, 81-105 (MR 27-3063).

[246] Takane, Y. (1987). Analysis of contingency tables by ideal point discriminant analysis, **Psychometrika 52(4)**, 493-513.

[247] Tan, W.Y. (1985). On Tiku's robust procedure- a Bayesian insight, **J. Statist. Plann. Inf. 11**, 329-340.

[248] Tan, W.Y. and Balakrishnan, N. (1986). Bayesian insight into Tiku's robust procedures based on asymmetric censored samples, **J. Statist. Comput. Simul. 24**, 17-31.

[249] Tan, W.Y. and Balakrishnan, N. (1988). Bayesian approach to robust comparison of two means based on asymmetric Type-II censored samples, **J. Statist. Comput. Simul. 30**, 81-102.

[250] Tan, W.Y. and Tabatabai, M.A. (1988). A modified Winsorized regression procedure for linear models, **J. Statist. Comput. Simul. 30**, 299-313.

[251] Teichroew, D. and Sitgreaves, R. (1961). Computation of an empirical sampling distribution for the W-classification statistic, **Ibid**, 252-275.

[252] Tiku, M.L. (1967). Estimating the mean and standard deviation from a censored normal sample, **Biometrika 54**, 155-165.

[253] Tiku, M.L. (1968). Estimating the parameters of lognormal distribution from censored samples, **J. Amer. Statist. Assoc. 63**, 134-140.

[254] Tiku, M.L., Balakrishnan, N., and Ambagaspitiya, R.S. (1989). Error rates of a robust classification procedure based on dichotomous and continuous random variables, **Commun. Statist.** - Simul. Comput. **18**(2). 571-588.

[255] Tiku, M.L., Gill, P.S., and Balakrishnan, N. (1989). A robust procedure for testing the equality of mean vectors of two bivariate populations with unequal covariance matrices, **Commun. Statist.** - Theor. Meth. **18**(9). 3249-3265.

[256] Tiku, M.L., Tan, W.Y., and Balakrishnan, N. (1986). **Robust Inference**, Marcel Dekker, New York.

[257] Tiku, M.L. (1970). Monte Carlo study of some simple estimators in censored normal samples, **Biometrika 57**, 207-211.

[258] Tiku, M.L. (1980). Robustness of MML estimators based on censored samples and robust test statistics, **J. Statist. Plann. Inf. 4**, 123-143.

[259] Tiku, M.L. (1982). Testing linear contrasts of means in experimental design without assuming normality and homogeneity of variances, **Biometrical Journal 24**, 613-627.

[260] Tiku, M.L. (1988). Modified maximum likelihhod estimation for the bivariate normal, **Commun. Statist.** - Theor. Meth. **17**(3). 893-910.

[261] Tiku, M.L. and Balakrishnan, N. (1984). Robust multivariate classification procedures based on the MML estimators, **Commun. Statist.** - Theor. Meth. **13**(8). 967-986.

[262] Tiku, M.L. and Balakrishnan, N. (1986). A robust test for testing the correlation coefficient, **Commun. Statist.** - Simula. Comput. **15**(4). 945-971.

[263] Tiku, M.L. and Balakrishnan, N. (1985). Robust classification procedures based on the MML estimators, In **Advances in Multivariate Statistical Analysis** (Eds., S. Das Gupta and J.K. Ghosh). pp. 513-536, Indian Statistical Institute, Calcutta.

[264] Tiku, M.L. and Balakrishnan, N. (1988a). Robust Hotelling-type T^2- statistics, **Biometrical Journal 30**, 283-293.

[265] Tiku, M.L. and Balakrishnan, N. (1988b). Robust Hotelling-type T^2- statistics based on the MML estimators, **Commun. Statist.** - Theor. Meth. **17**(6). 1789-1810.

[266] Tiku, M.L. and Balakrishnan, N. (1988). Generalization of the robust bivariate T^2 statistic to multivariate populations, **Commun. Statist.** - Theor. Meth. **17**(11). 3899-3911.

[267] Tiku, M.L. and Balakrishnan, N. (1989). Robust classification procedures based on the MML estimators, **Commun. Statist.** - Theor. Meth. **18**(3). 1047-1066.

[268] Tiku, M.L., Balakrishnan, N., and Ambagaspitiya, R.S. (1989). Error rates of a robust classification procedure based on dichotomous and continuous random variables, **Commun. Statist.** - Simul. Comput. **18**(2). 571-588.

[269] Tildesley, M.L. (1921). A first study of the Burmese skull, **Biometrika, 13** 247-251.

[270] Tatsuoka, K.K. and Tatsuoka, M.M. (1987). Bug distribution and statistical pattern classification, **Psychometrika 52(2)**, 193-206.

[271] Traven, H.G.C (1991). A neural network approach to statistical pattern classification by "Semiparametric" estimation of probability density functions, **IEEE Transactions on neural networks 2(3)**, 366-377.

[272] Tukey, J.W. (1960). A survey of sampling from contaminated distributions, In **Contributions to Probability and Statistics**, (Eds., I. Olkin, S.G. Ghurye, W. Hoeffding, W.G. Madow and H.B. Mann). pp. 448-485, Stanford University Press, Stanford, California.

[273] Tukey, J.W. and McLaughlin, D.H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: trimming/Winsorization 1, **Sankhya**, Ser. A **25**, 331-352.

[274] Uematu, T. (1964). On a multidimensional linear discriminant function, **Ann. Inst. Statist. Math. 16**, 431-437.

[275] Urbakh, V.Y. (1971). Linear discriminant analysis: loss of discriminating power when a variate is omitted, **Biometrics 27**, 531-534.

[276] Van Ryzin, J. (1966). Bayes rsik consistency of classification procedures using density estimation, **Sankhya A 26**, 25-30.

[277] Von Mises, R. (1945). On the classification of observation data into distinct groups, **Ann. Math. Statist. 16**, 68-73.

[278] Wagner, H. M. (1959). Linear programming techniques for regression analysis, **J. Amer. Statist. Assoc.**, **54**, 206-212.

[279] Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups, **Ann. Math. Statist. 15**, 145-162.

[280] Wald, A. (1950). **Statistical Decision Functions**, Wiley: New York

[281] Wald, A. and Wolfowitz, J. (1950). Characterization of the minimal complete class of decision functions when the numbers of decisions and distributions is finite, **Proc. Second Berkeley Symp. on Stat. and Probability.**

[282] Weiner, J.M. and Dunn, O.J. (1966). Elimination of variates in linear discriminant analysis, **Biometrics 22**, 268-275.

[283] Welch, B.L. (1939). Note on discriminant functions, **Biometrika 31**, 218-220.

[284] Wesler, O. (1959). A classification problem involving multinomials, **Ann. Math. Statist. 30** 128-133.

[285] Yale, C. and Forsythe, A.B. (1976). Winsorized regression, **Technometrics 18**, 291-300.