STATISTICAL DISCRIMINATION IN THE AUTOMATION OF CYTOGENETICS AND CYTOLOGY

by

Simon Kirby

Thesis sumitted for the degree of Doctor of Philosophy
in the University of Edinburgh

1990

DECLARATION

The following record of research work is submitted as a thesis for the degree of Doctor of Philosophy in the University of Edinburgh, having been submitted for no other degree. Except where acknowledgement is made, the work is original.

To Barbara

and both our families

TABLE OF CONTENTS

ABSTRACT

The thesis considers two topics in the automation of cytogenetics and cytology: the automated allocation of human chromosomes to the twenty-four classes which humans possess; and the detection of abnormal cervical smear specimens.

For chromosome allocation, the following work is presented and evaluated on a number of data sets derived from chromosome preparations of different quality:

1. Three new procedures for modelling between-cell variation.
2. Six ways of combining class information on variability in multivariate Normal discrimination.
3. Covariance selection models for individual chromosome classes and an assumed common covariance structure for a number of classes.
4. Some two-stage procedures for the calculation of discriminant scores in multivariate Normal discrimination.
5. The application of some non-parametric and semi-parametric methods.
6. The modelling of band-transition sequence probabilities.

For the detection of abnormal cervical smear specimens, the use of a consensus probability of a specimen being abnormal, derived from a number of cytologists' assessments, is considered. The sequential use of multiple regression equations to try to predict the logit transformations of these consensus probabilities is described.

Finally, the sequential use of features in multivariate discrimination is considered mainly for the case of two known multivariate Normal populations with equal covariance matrices.

ACKNOWLEDGEMENTS

I would like to thank my supervisors, Dr C. M. Theobald, Dr A.D. Carothers and Mr D. Williams for their help, criticism and advice throughout and the Science and Engineering Research Council for the award of a CASE studentship. I also particularly wish to thank Mr J. Piper for his encouragement and help.

I acknowledge the support of the Commission of the European Communities through the Medical and Health Research Program, project number II.1.1/13 , for sponsoring my attendance at workshops at Llangollen and Dalhousie and a trip to the Chromosome laboratory, Rigshospitalet, Copenhagen.

Others I wish to thank for helpful discussions are Mr T. Gerdes, Mr J Maahr and Dr. C. Lundsteen.

Finally, thanks to Karen, Moira, Deena and Mark for their encouragement (on occasions!) and most of all to Barbara for all her love and support.

Chapter 1
Introduction.

In this chapter a brief background to the topics in cytogenetics and cytology considered in the thesis is given. The layout of the thesis is also outlined.

The topics in cytogenetics and cytology considered in the thesis are the allocation of human chromosomes to the twenty-four classes that humans possess and the detection of abnormal cervical smear specimens.

Chromosome analysis, the examination of the chromosome complement of a number of cells from an individual for abnormalities in number or structure is widely used in ante-natal screening, the diagnosis of haematological tumours and biological research. To examine the chromosome complement of an individual the chromosomes in a cell are each routinely allocated to one of the twenty-two autosomal and two sex classes. This can be done more quickly and more economically by the use of automation. The current state of automation is described in chapter 2 and is the result of more than twenty-five years of research (Piper et al, 1980).

The manual screening of cervical smear specimens is another time consuming procedure for the detection of clinical abnormalities. So far the procedure has not been routinely automated but there is considerable interest in the speed and accuracy which may be obtained by automation.

In chapter 2 automation in human cytogenetics and in the detection of abnormal cervical smear specimens is described. The application of statistical discrimination methods in these two areas is also reviewed.

Chapter 3 contains brief descriptions of the ten chromosome data sets and one cervical smear data set used in the thesis. The numbers of chromosomes and cervical smear specimens are also given along with brief descriptions of the features (variables) for which values were obtained.

In chapter 4 three new procedures for the modelling of between-cell variation are outlined. Such modelling is important in trying to remove this source of variation before statistical discrimination methods are used to allocate chromosomes to the twenty-four classes.

Chapter 5 considers six ways of combining class information on variability in multivariate Normal discrimination for chromosome allocation. The purposes of these methods of combining information are reduction in the computational time required to allocate the chromosomes in a cell and reduction in the number of parameters compared with the use of unrelated covariance matrices.

In chapter 6 the idea of parameter reduction in multivariate Normal discrimination for chromosome allocation is explored further with the use of covariance selection models. These may be used to model the covariance structure of individual classes or an assumed common covariance structure for a number of classes. Computational time is reduced compared with that obtained under the assumption of unrelated covariance matrices if sufficient elements of each estimated inverse covariance matrix are set equal to zero.

Another way of attempting to reduce computational time for chromosome allocation is to consider a sequential approach to the calculation of the discriminant scores. In chapter 7 some two-stage procedures are examined for multivariate Normal discrimination.

Non-parametric and semi-parametric methods of statistical discrimination make no or fewer assumptions about the forms of the individual class distributions than parametric methods. In chapter 8 four of these methods are considered for application to chromosome allocation.

In chapter 9 some models for the probabilities of band-transition sequences derived from the sequence of dark and light bands along a chromosome are outlined and applied to the allocation problem.

In chapter 10 attention is switched to the detection of abnormal cervical smear specimens. The use of a consensus probability of a cervical smear specimen being abnormal derived from a number of cytologists' opinions is considered in this chapter. The sequential use of multiple regression equations to try to predict the logit transformations of these consensus probabilities is described.

In chapter 11 the topic of sequential use of features in multivariate discrimination is considered. The motivation for this is the interest in saving feature measurement time for the allocation of objects from a cervical smear specimen to various classes. The computation required by Fu's dynamic

programming approach to obtaining an optimal varying order of feature measurement is briefly reviewed. This approach assumes that the only costs are for feature measurement and misallocation which are commensurable and that the class distributions are known. An alternative approach to obtaining an optimal varying order of feature measurement, when the feature order is free to vary, is to find an optimal fixed order of feature measurement. This approach may require less computation to obtain a solution. Results are obtained for two known multivariate Normal populations with equal covariance matrices for Fu's criterion and a new criterion, for early allocation when the feature order is fixed. Both criteria assume that the only costs are feature measurement costs and misallocation costs. The new criterion assumes that feature measurement costs and misallocation costs are not commensurable. Even the evaluation of an optimal fixed order of feature measurement may be computationally too demanding so sub-optimal approaches to obtaining a fixed order of feature measurement are also proposed for the two criteria. The evaluation of an optimal fixed order of feature measurement is also considered for the case when the cost of the calculation of the discriminant scores after a feature has been measured is included in the measurement cost of a feature. In this case the two criteria for early allocation need to be re-defined and it may not be optimal to calculate the discriminant scores after every feature measurement. It may also be true that a one-stage discriminant procedure gives a lower cost procedure than a sequential procedure. Finally, an empirical approach for the use of the new criterion for early allocation of an object is advocated and illustrated on the cervical smear data.

In chapter 12 the results obtained in the thesis are reviewed and suggestions for further work are made.

Chapter 2
Automation in cytogenetics and cytology.

2.1 Introduction.

In this chapter the current state of automation in human cytogenetics and in the detection of abnormal cervical smears is described. The work done on statistical discrimination in these areas is also reviewed.

2.2 Automation in human cytogenetics.

2.2.1 Automated systems.

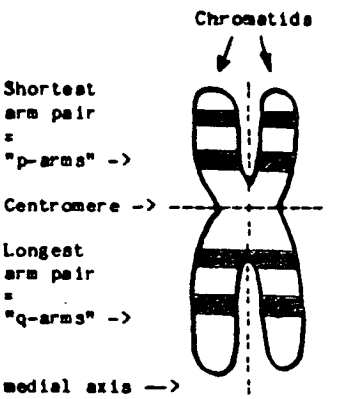
The most automated systems at present mostly proceed along similar lines for routine karyotyping (the determination of a person's chromosome complement). Firstly, good quality metaphase (metaphase being the stage of cell division at which chromosomes are most suitable for analysis) cells are found by machine scanning of a stained specimen on a slide (good is taken to mean well separated chromosomes). The specimen is usually a sample from peripheral blood or amniotic fluid. The cells are then digitised at high resolution and the chromosomes separated from the background. At this stage touching and overlapping chromosomes are separated by an operator using a light pen. Finally, chromosomes are allocated to the twenty-two autosomal or two sex classes with operator correction of errors.

For aberration scoring, the counting of structural chromosome aberrations in metaphase cells, this level of operator interaction is too much for automation to be economic. For the detection of dicentric chromosomes, chromosomes which have two centromeres, only chromosomes finally allocated to an abnormal class are presented to the operator (Piper et al, 1988 and Lörch et al, 1989).

The features used for routine karyotyping differ from system to system. Measures of size and centromeric index (how far along the chromosome the join between the chromatids, Figure 2.1 , which make up a metaphase chromosome is) are common across systems. Other important features are those based on the bands perpendicular to the longitudinal axis of a chromosome which result from staining of the cells (Figure 2.1). Taking the average density of staining across the longitudinal axis of a chromosome for a

Figure 2.1

Sketch of a chromosome showing
the banding pattern produced
by staining.



series of points along the axis gives a profile of density of staining along the length of a chromosome, Figure 2.2 . Using a number of weight functions then gives summaries of these density profiles (Figure 2.3). The weight functions are used by multiplying the average density of staining at points along a chromosome by the corresponding values of a weight function, summing the values obtained and dividing by the total average density of staining. Information from the bands may alternatively be used by considering the sequence of dark and light bands. One approach has been to divide each chromosome into 13 segments and the location of each peak, its density and the difference in density from the next light band working from the end of the short arm of a chromosome to the long arm is recorded defining a so-called band-transition sequence (Lundsteen and Granum, 1979). Another approach has been to use just the relative position along a chromosome of certain bands (van Vliet et al, 1989).

Similarly to the definition of a density profile along a chromsome a profile of shape may be defined. This is done by summing the squared distance from the longitudinal axis of the chromosome times the density of staining and dividing by the sum of the density of staining, at a series of points along the axis (Piper and Granum, 1989). As for the density profiles a number of weight functions are then used to give summaries of the shape profiles.

Because of between-cell variation which results from cells being stopped at different stages of metaphase (and hence at different stages of contraction) and from different cell preparation some feature values are normalised. For size features a multiplicative transformation is usually performed by dividing by the value for the median-sized chromosome in the same cell (Piper and Granum, 1989) or transforming to natural logs, subtracting a cell average and dividing through by the within-cell standard deviation (Lundsteen et al, 1981). Other features may or may not be transformed by subtracting a cell average and dividing by the within-cell standard deviation (Piper and Granum, 1989 and Lundsteen, Gerdes and Maahr, 1986).

For the detection of dicentric chromosomes, the features used are different because allocation of the chromosomes to classes is not required (Piper et al, 1988 and Lörch et al, 1989). These features are not described here because aberration scoring is not considered in the main part of the thesis.

Figure 2.2

Density profile along a
chromosome.
(c indicates centromere position.)

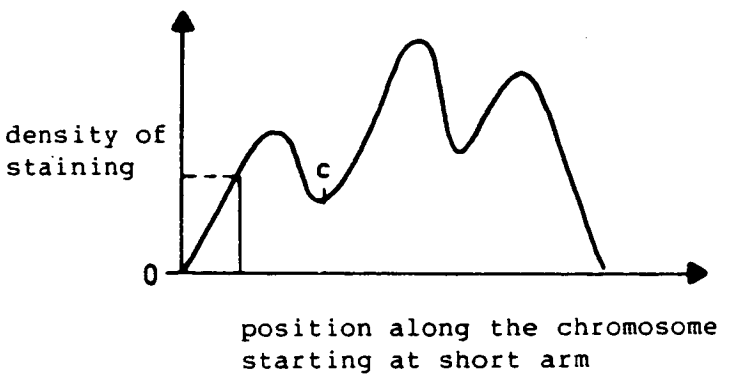
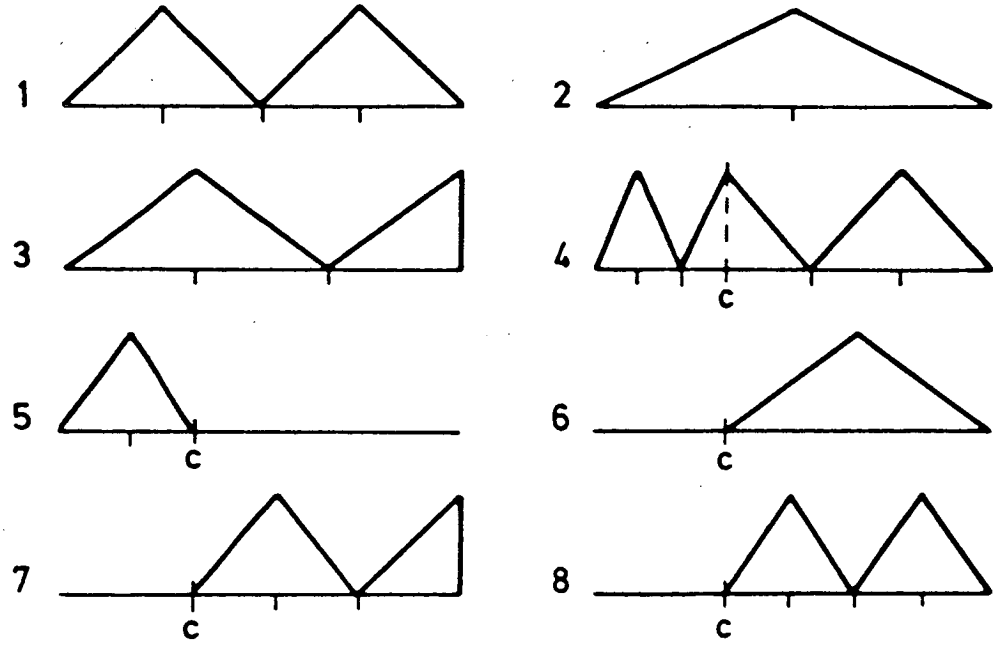


Figure 2.3

Weight functions used by
Lundsteen, Gerdes and Maahr (1986)
to obtain weighted sums of density
profile.



2.2.2 Statistical methods for automated karyotyping.

Metaphase finding.

Human chromsomes are most visible at the stage of cell division referred to as metaphase. Only some cells in a specimen will be at this stage of cell division so a first stage in chromosome analysis is to find these cells (Figure 2.4).

Because there are many cells on a slide and a specimen may have few good metaphase cells, the discrimination techniques used have been very quick and simple. Often no more than a "box" discrimination procedure is used. That is to say upper and lower limits are set separately or jointly for values of all the features by ad-hoc methods. Cells which have feature values inside all of the feature limits are accepted as "good" quality metaphase cells. This solution is sometimes done in stages with progressively more costly (in time) feature values obtained at each stage. van den Berg et al (1981) have considered a regression approach based on the assignment of a quality index for metaphases in a training sample. The quality index is regressed on features which are useful for predicting the quality of a metaphase.

Normalisation.

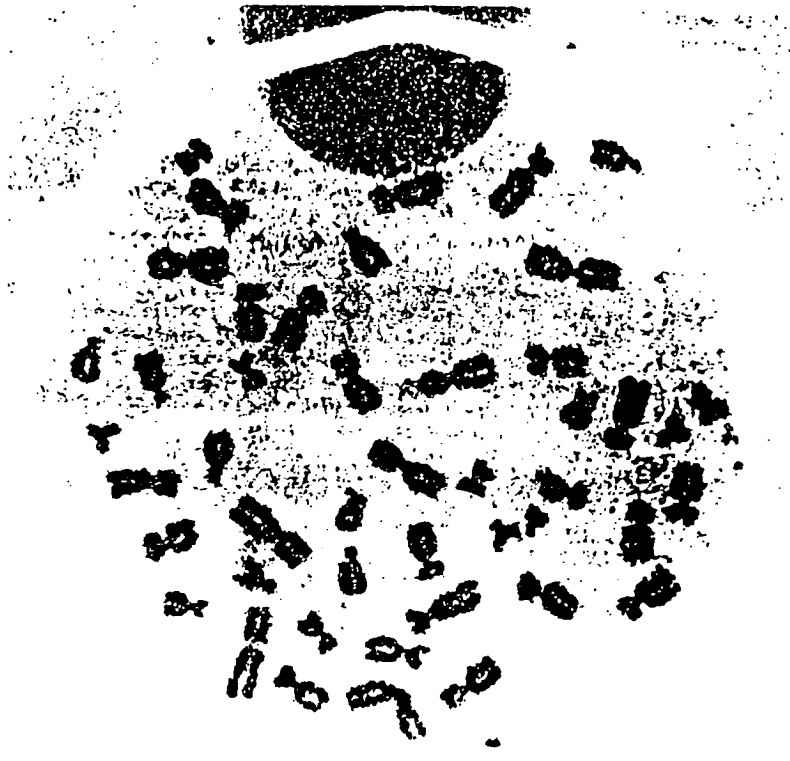
As described above, values of size features are usually normalised by multiplicative transformations based on the median-sized chromosome or the average for the chromosomes in a cell. Hilditch and Rutovitz (1972) considered multiplicative normalisation of size features by total cell size (the total size of all the chromosomes in a cell), the median of approximately equal-sized chromosomes, the size of an easily identified, chromosome or chromosomes and a weighted average based on all chromosomes identified with the weight for a chromosome from the ith class equal to

$$\mu_i^2\sigma_i^{-2}(\Sigma_i\mu_i^2\sigma_i^{-2})^{-1} \quad , \tag{2.1}$$

where μ_i and σ_i are the mean and standard deviation for normalised values for class i and the summation is over all chromosomes in a cell. The last method requires iteration between allocation and calculation of the normalising factor

Figure 2.4

Chromosomes in a cell at the metaphase stage of cell division.



starting with an initial allocation. Hilditch and Rutovitz (1972) found that judged by the inter-cell variance of ten groups of classes the weighted average gave a slightly worse result than total cell size, which was best, for unknown prior allocation for complete normal cells. However, the weighted average gave the best result for incomplete normal cells. When abnormalities were introduced the weighted average also performed best for unknown allocation of the chromosomes. In practice a multiplicative normalisation based on the median-sized chromosome in a cell or the average size of the chromosomes in a cell has been used. This is because these approaches have been found to give acceptable results without requiring the iteration needed by the weighted average considered by Hilditch and Rutovitz (1972).

Isolated allocation of human chromsomes.

Often a first step in the allocation of the chromosomes in a cell is to consider the allocation of each chromosome without regard to the allocation of the others. The method of statistical discrimination adopted has depended on the features used. When size, centromeric index, sums of weighted density profiles and sums of weighted shape profiles have been obtained (type a. features) discrimination based on the assumption of multivariate Normal distributions (Piper, 1987) or classification trees (Shepherd, Piper and Rutovitz, 1987) has been used. For band-transition sequences and the relative location of particular bands, non-parametric methods which assume all features to be independent have been used (Lundsteen et al, 1981 and van Vliet et al, 1989). Band-transition sequences have also been represented as differing length strings which are used to build a Markov network for each class (Thomason and Granum, 1986). Allocation of the chromosomes proceeds by matching a new string to a network using a cost function based on observed transition probabilities. Prior probabilities of the 24 chromosome classes have not been used for the classification trees or the Markov network for strings representing band-transition sequences but have been used for the other methods. For all the methods the cost of allocating a class i chromosome to class i' ($i \neq i'$) has been assumed the same for all i and i'. This assumption is continued throughout the thesis.

For the type a. features referred to above, both equal and unrelated covariance matrices have been assumed for multivariate Normal discrimination (Granum, 1982 and Piper, 1987). This has been done using the so-called

Estimative approach which substitutes estimates for parameter values in the probability density functions. Error rates estimated by the use of the test-set method (described in chapter 5) have shown that the asumption of unrelated class covariance matrices gives lower estimated error-rates than the assumption of equal class covariance matrices for two data sets (Granum, 1982 and Piper, 1987). Piper (1987) has shown, however, for a particular data set that assuming zero correlations between features for estimated unrelated covariance matrices gives a smaller estimated error-rate than not making this assumption and results in a large reduction in computational time and the number of parameters to be estimated. The time taken for computation is important because the operator has to wait for the allocations given by the statistical discrimination procedure.

A so-called tree classifier has also been used for the type a. features described above (Shepherd, Piper and Rutovitz, 1987). This tree classifier is called the Analogue Concept Learning System (Paterson and Niblett, 1982). This is a tree of binary splits. All the training data starts at the root of the tree and for each feature x , assumed to be quantitative, in turn the best split is determined. The splits are the divisions

$$x \leqq c_q \; \{q = 1,\ldots,n' \leqq (n - 1)\}$$

for c_q defined as mid-way between consecutive distinct values for all n ordered values of the feature. The best split over all features is then used to split the training data into two. Best was defined by Shepherd, Piper and Rutovitz (1987) as the maximal entropy gain. Splitting of descendant nodes is continued until a stopping rule is satisfied. The stopping rule used by Shepherd, Piper and Rutovitz (1987) is to stop splitting if the estimated error-rate of a test set is worsened. Nodes which could not be split further were given class labels according to which class was in the majority in the training set. A new object is allocated to a class by sending it to the appropriate descendant node of each node according to the value of the feature used in the split for that node. This continues until the object reaches a node without descendant nodes and it is then allocated to the class corresponding to the class label of this node. The results for ACLS were found to be much worse than for Estimative

multivariate Normal discrimination with equal covariance matrices. The reason for this was conjectured by Shepherd, Piper and Rutovitz (1987) to be the inadequacy of division rules based on lines perpendicular to each feature axis.

For the band-transition sequence data, a non-parametric method has been used to estimate frequencies of peak density values (values of 0-6) and density difference values (values of 0-4) for each of 14 segments (the 13 segments described earlier plus an artificial segment) for each of the 24 chromosome classes (Lundsteen et al, 1981). Values of these 28 band transition features have then been regarded as independent. This has been combined with the assumptions of independence and Normality for normalised area, area centromeric index and density centromeric index by multiplying the values of their estimated probability density functions by the estimated probabilities for the band-transition sequences. Weighting the values of the estimated probability density functions more heavily than the estimated probability for the band-transition sequence has been found to give lower estimated error-rates than the use of equal weights (Lundsteen et al, 1981).

For the measurement of the relative location of particular chromosome bands, independence of features has again been assumed. Because of the large number of possible values for these features, histograms with particular bin widths have been constructed for allocation of chromosomes (van Vliet et al, 1989). The number of bin widths used by them is derived from assuming the features to be Normally distributed and that therefore $n_i^{\frac{1}{2}}$, where n_i is the number of class i chromosomes in a training set, is approximately the correct number of bins for a feature for each class.

Allocation of chromosomes to satisfy a normal karyotype.

Because most cells looked at in automated chromosome analysis are expected to be normal it is helpful if the final allocations of the chromsomes in a cell conform to a normal karyotype. It can also be expected that an allocation which satisfies the normal class sizes will contain fewer errors on average. For a cell from a male a normal karyotype corresponds to 22 autosomal pairs, an X chromosome and a Y chromoosome whilst for a female there is a second X chromosome instead of the Y chromosome. Until recently this requirement was achieved by various sub-optimal procedures (Piper, 1986). There is now available, however, a fast algorithm which gives the allocation of the

chromosomes in a cell which maximises the product of the estimated posterior probabilities subject to the constraint that the allocations satisfy a normal karyotype (Kleinschmidt, Lee and Schannath, 1987). This algorithm works by solving the so-called transportation problem (Tso and Graham, 1983). In this context there are 24 chromosome classes which have a 'demand' for a certain number of chromosomes which must be satisfied by the 'supply' of chromosomes in a cell. The algorithm can cope with missing or additional chromosomes. The sub-optimal procedures and transportation algorithm both have been found to give slightly lower estimated error-rates than the isolated allocation of chromsomes (Piper, 1986 and Tso, 1989) and the latter has been found to outperform the former methods (Tso, 1989).

Whole-cell approaches.

Habbema has identified three models for the allocation of chromosomes (1979). These approaches are the isolated allocation of individual chromosomes, the individual allocation of the chromsomes in a cell to satisfy a normal karyotype and the simultaneous allocation of all the chromosomes in a cell. This last approach based on vectors of length n_c*p , where n_c is the number of chromosomes in the cell and p is the number of features, has not been tried in practice.

Multiple cell karyotyping.

A multiple-cell approach which recognises that allocation errors are likely in each cell has been proposed for detecting cell lines with an additional or a missing chromosome in a particular class (Carothers, Rutovitz and Granum, 1983). This uses the theory of hypothesis testing to distinguish between normal cell lines and those with a missing or extra chromosome.

2.2.3 Statistical methods for aberration scoring.

The automated detection of dicentric chromosomes has been attempted using statistical discrimination (Piper et al, 1988 and Lörch et al, 1989). The methods used have been: a sequential approach with a "box" discrimination procedure followed by four stages of linear discriminant functions (Lörch et al, 1989); and a within-cell "box" discrimination procedure (Piper et al, 1988). The "box" discrimination procedure used by Lörch et al (1989) is to eliminate objects with a small probability of being dicentric. The second stage tries to recognise non-chromosome objects. The third stage again eliminates objects

with a low probability of being dicentrics. The fourth stage allocates objects with a large probability of being dicentrics. Finally, the fifth stage separates the remaining objects. The within-cell "box" discrimination procedure described by Piper et al (1988) is a "box" discrimination procedure for which the limits are changed from cell to cell in order to adjust for between-cell variation.

2.3 Automated detection of abnormal cervical smear specimens.

2.3.1 Automated system.

The system designed at the MRC Human Genetics Unit, Edinburgh currently proceeds as follows (Carothers, 1987 and Carothers, 1988). As each object in a specimen stained with a DNA-specific absorption stain is scanned its integrated optical density (IOD) relative to the modal value for all the objects is used to allocate the object as worthy of further study or not. The IOD is used because it gives a measure of DNA content which is approximately constant in normal cells. Malignant and premalignant cells frequently contain much larger and more variable amounts of DNA (Tucker, 1979). The modal value is initially obtained from a small number of objects which are never allocated and is then updated as further objects are scanned. Because many 'abnormals' at this stage are artefacts, other features are measured to try to reduce the number of artefacts left (Tucker, 1979). These features are used sequentially in a "box" discrimination procedure as described earlier for metaphase finding. Once this has been done Estimative multivariate Normal discrimination based on an assumption of common covariance matrices is used to calculate linear discriminant functions for discrimination between abnormal and normal cells and for discrimination between abnormal cells and all non-cell material. Bivariate Normal distributions for these scores are estimated for normal cells, abnormal cells and artefacts. The 16 most likely abnormal objects are then measured to provide features at the level of the specimen. Most likely is defined as having a high rank for a ranking system monotonically related to the posterior probability of an object being an artefact (Carothers, 1987). Estimative multivariate Normal discrimination, with the assumption of common covariance matrices, is then used to allocate the specimen to one of the classes, normal, requiring attention by a cytologist or requiring interactive analysis. The classifications in the training set are given by reference diagnoses derived from a number of cytologists (Carothers, 1988). For interactive analysis the 16 objects most likely to be abnormals are allocated by the operator to one

of the classes, artefact, normal, inflammatory or abnormal. New features based on the operator allocations are then derived and again Estimative multivariate Normal discrimination with the assumption of common covariance matrices is used to declare the specimen as normal or requiring attention by a cytologist. It should be noted from this that the allocation of any object from the 16 most likely abnormals to the abnormal class by the operator is not considered sufficient to declare the whole specimen as abnormal.

2.3.2 Statistical discrimination.

The statistical methods used in the Edinburgh system are therefore sequential "box" discrimination and Estimative multivariate Normal discrimination. Other approaches described by Timmers (1987) include allocation of specimens to one of three specimen classes (normal, atypical or carcinoma) according to which cell class is in the majority. He also describes the use of nearest neighbour analysis and correspondence analysis based on the counts of cell types for each specimen for allocation of specimens to one of these classes.

Chapter 3
Chromosome and cervical smear data sets.

3.1 Introduction.

In this chapter a description is given of the chromosome and cervical smear data sets used in the thesis. Ten chromosome data sets and one cervical smear data set were used. The Edinburgh and Philadelphia chromosome data described below are from routine study and the other chromosome data were specially collected. The specially collected data sets are referred to as special except for the Copenhagen data for which this adjective is not used. This is for consistency with reference made to the data elsewhere, e.g., Piper and Granum (1989). The chromosome data sets were supplied by the MRC Human Genetics Unit, Edinburgh and the Rigshospitalet, Copenhagen. All the chromosome data sets have previously been used in allocation experiments, e.g., Granum (1982), Lundsteen, Gerdes and Maahr (1986) and Piper and Granum (1989). The cervical smear data were specially collected and supplied by the MRC Human Genetics Unit, Edinburgh.

3.2 Chromosome data sets.

3.2.1 Version a. of Edinburgh data set (normalised for between-cell variation).

These data consist of 28 feature values for each of 5548 chromosomes from 125 normal human male peripheral-blood cells. The features are given in Table 3.1 . Touching chromosomes were given special codes in this data set and so were omitted from the allocation results presented later. This leaves 4270 chromosomes. The division of the data set into two parts used in the thesis was the same as that used by Piper and Granum (1989), with the exclusion of the touching chromosomes, obtained by random allocation of cells to two groups. This division and the division of the other data sets into two parts was done for the test-set method of error rate estimation described in chapter 5. The division gives a first part of 2228 chromosomes from 66 cells and a second part of 2042 chromosomes from 59 cells.

3.2.2 Version b. of Edinburgh data set (normalised for between-cell variation).

A version of this data set was used which was not normalised for between-cell variation; this had slightly different features from the above data

Table 3.1

Feature values obtained for chromosomes in Edinburgh data set, versions a. and c. of Copenhagen data set and Philadelphia data set.

Features

1. area
2. area + length
3. density = (sum of pixel values)/(number of pixels)
4. area centromeric index
5. density centromeric index *
6. coefficient of variation of density profile
7. normalised root of sum of squared density differences (n.s.s.d.)
8. length *
9.-14. six weighted sums of density profile
15.-20. six weighted sums of shape profile
21.-26. six weighted sums of profile of absolute differences of density profile
27. length centromeric index
28. convex hull perimeter (c.h.p.)

* feature 5 replaced by ratio of mass centromeric index to area centromeric index and feature 8 replaced by m.d.r.a. (difference in mass between top and bottom half of density profile divided by area under profile) in version a. of Edinburgh data set

set and did not exclude touching chromosomes so a second comparable normalised version of the data set was used. For this version, length replaced m.d.r.a. (defined in Table 3.1) and density centromeric index replaced the ratio of centromeric indices. This data set contains 5548 chromosomes. The random split into two parts described above gives 2931 chromosomes in the first part and 2617 chromosomes in the second part.

3.2.3 Version c. of Edinburgh data set (not normalised for between-cell variation).

This data set contains the same chromosomes and has the same features as version b. of the normalised data above.

3.2.4 Version a. of Copenhagen data set (normalised for between-cell variation).

These data consist of 28 feature values for each of 8106 chromosomes from 180 normal male and female peripheral-blood cells. The features were the same as for version b. of the normalised Edinburgh data set. The random split of the data by cell into two groups used by Piper and Granum (1989) was used. This gives a first part of 3416 chromosomes from 76 cells and a second part of 4690 chromosomes from 104 cells.

3.2.5 Version b. of Copenhagen data set (normalised for between-cell variation).

These data consist of 39 feature values for each of 6989 of the chromosomes in the above data set. The features are given in Table 3.2 . The division of the cells into two parts was the same as for version a. above except that one cell was missing from the data set to give a second part of only 103 cells. There are 2941 chromosomes in the first part of the data set and 4048 chromosomes in the second part. To avoid confusion with version a. of this data set this version is referred to in later chapters as the reduced Copenhagen data set.

3.2.6 Version c. of Copenhagen data set (not normalised for between-cell variation).

The chromosomes and features were the same as for version a. of this data set.

3.2.7 Version a. of Philadelphia data set (normalised for between-cell variation).

These data consist of 28 feature values for each of 5817 chromosomes from 130 normal male and female chorionic-villus cells. The features were the same as for version b. of the Edinburgh data set. The random split by cell into

Table 3.2

Feature values obtained for chromosomes in version b. of Copenhagen data set and two special Copenhagen data sets.

Features

1. area
2. area centromeric index
3. density centromeric index
4.-11. eight weighted sums of density profiles
12.-39. density and density difference values for fourteen segments of chromosome

two parts by Piper and Granum (1989) gives a first part of 2899 chromosomes from 64 cells and a second part of 2918 chromsomes from 66 cells.

3.2.8 Version b. of Philadelphia data set (not normalised for between-cell variation).

The chromosomes and features are the same as for version a. of this data set.

3.2.9 Copenhagen special amniotic-fluid data set (normalised for between-cell variation).

These data consist of 39 feature values for each of 9396 chromosomes from 217 male and female cells from amniotic fluid. The features are the same as for version b. of the Copenhagen data set given in Table 3.2 . The data set was split into two parts by random allocation of cells to two groups. This gave a first part of 4695 chromosomes from 114 cells and a second part of 4701 chromosomes from 113 cells.

3.2.10 Copenhagen special peripheral-blood data set (normalised for between-cell varaiation).

These data consist of 39 feature values for each of 10075 chromosomes from 230 male and female cells from peripheral blood. The data were split into two parts by random allocation of cells to two groups. This gives a first part of 5170 chromosomes from 118 cells and a second part of 4905 chromosomes from 112 cells. The features are given in Table 3.2 .

3.3 Numbers of chromosomes in chromosome data sets.

Overlapped chromosomes had previously been excluded from all versions of the Edinburgh, Copenhagen and Philadelphia data sets (Piper and Granum, 1989). The reduced Copenhagen data set excluded severely bent chromosomes as well as overlapped ones. Severely bent and overlapped chromosomes were also previously excluded from the special Copenhagen data sets (Lundsteen, Gerdes and Maahr, 1986). In a small number of instances chromosomes had been missing from the original data because they had not been included in the image of the cell.

3.4 Further description of features in chromosome data sets.

A fuller description of the Edinburgh data set, of versions a. and c. of the Copenhagen data set and of the Philadelphia data set is given in Piper and Granum (1989). Further description of the features in version b. of the Copenhagen data set and the two special Copenhagen data sets is given in Lundsteen et al (1981) and Lundsteen, Gerdes and Maahr (1986). Features 1-11 in Table 3.2 for these three data sets are those used in the WDD classifier described by Lundsteen, Gerdes and Maahr (1986) and referred to in chapters 5, 6, 8 and 9.

Examination of the features described shows that in versions b. and c. of the Edinburgh data set, versions a. and c. of the Copenhagen data set and the Philadelphia data set one feature is a linear combination of two others (feature 2 = feature 1 + feature 8). Hence a maximum of two of these three features was used in the allocation experiments described later in the thesis.

3.5 Normalisation of chromosome data for between-cell variation.

The features in versions a. and b. of the Edinburgh data set, version a. of the Copenhagen data set and version a. of the Philadelphia data set were normalised in one of two ways or else left unchanged (Piper and Granum, 1989). Values of size features were divided through by the value of the median-sized chromosome in the same cell. All other features except the measures of centromeric index were normalised by subtracting the cell mean, dividing through by the within-cell standard deviation and multiplying by 100.

The other normalised data sets were obtained by subtracting the cell mean of logged area from the logged value of area for each chromosome and dividing through by the within-cell standard deviation of logged area (Lundsteen et al, 1981).

3.6 Cervical smear data set.

3.6.1 Object data.

For the first 100 objects scanned on each of 92 of the 489 slides described below, values of 18 features were obtained and a visual assessment of the class of the object was made. The object was classified by a cytologist as

belonging to one of the six classes given in Table 3.3 . The features are given in Table 3.4 .

3.6.2 Specimen data

For each of 489 slides, independent assessments were made by four cytologists. Each specimen was allocated by each cytologist to one of the nine classes given in Table 3.5 or else described as too poor a specimen to be allocated. For each of 408 of these slides 25 feature values were obtained. The features are given in Table 3.6 .

3.7 Further description of features in cervical smear data set.

The object and specimen features listed in Tables 3.4 and 3.6 are more fully described in Carothers (1987) and Tucker (1979).

Table 3.3

Classification of objects from cervical smear specimens.

Classes

1. artefact
2. normal cell without cytoplasmic staining
3. inflammatory cell
4. abnormal cell
5. normal cell with cytoplasmic staining
6. other suspicious object

Table 3.4

Feature values obtained for objects from cervical smear specimens.

Features

1. area
2. integrated optical density (i.o.d.)
3. mean i.o.d. of pixels in 'skirt' (see Tucker, 1979) round object
4. number of limbs possessed by object
5. position of peak of histogram of i.o.d. of 'normal' cells (pk)
6. variance of i.o.d. of pixels in centre of object
7. variance of i.o.d. of pixels at edge of object
8. mean of i.o.d. of pixels in centre of object
9. mean of i.o.d. of pixels at edge of object
10. number of pixels in centre of object
11. mean density
12. chord measurement (see Tucker, 1979)
13. box measurement (see Tucker, 1979)
14.-17. perimeter tests (see Tucker, 1979)
18. ellipse (see Tucker, 1979)

Table 3.5

Classification of cervical smear specimens.

Classes

1. Normal
2. II
3. IIR
4. CIN 1
5. CIN 1/2
6. CIN 2
7. CIN 2/3
8. CIN 3
9. Invasive carcinoma

Severity of disease increases with class number.

Table 3.6

Feature values obtained for cervical smear specimens.

Features

1. mean of (i.o.d./pk) for 16 objects ranked as most likely to be abnormal.
2. mean area of 16 objects in 1.
3. mean of feature 6. in Table 3.4 for 16 objects in 1.
4. mean of ranking for abnormality for 16 objects in 1.
5. minimum ranking for abnormality for 16 objects in 1.
6. number of normal cells amongst 16 objects in 1.*
7. number of inflammatory cells amongst 16 objects in 1.*
8. number of abnormal cells amongst 16 objects in 1.*
9. sum of (100 * i.o.d./pk) over all cells amongst 16 objects in 1.*
10. minimum of ranking for abnormality over all cells amongst 16 objects in 1.*
11. max. of (100 * i.o.d./pk) over all cells amongst 16 objects in 1.*
12. mean of (100 * i.o.d./pk) over 7 most abnormal cells amongst 16 objects in 1.*
13. mean area of 7 most abnormal cells amongst 16 objects in 1.*
14. mean of feature 6. in Table 3.4 for 7 cells ranked as most likely to be abnormal amongst 16 objects in 1.*
15. mean of ranking for abnormality for 7 cells ranked as most likely to be abnormal amongst 16 objects in 1.*
16. number of objects with i.o.d./pk $\geqq$ 2 passed through "box" discrimination procedure.
17. number of objects with i.o.d./pk $\geqq$ 2 and passed through stage 1 of "box" discrimination procedure.
18. number of objects with i.o.d./pk $\geqq$ 2 and rejected at stage 1 of "box" discrimination procedure.
19. number of objects with area $\geqq$ 50 and 0.5 $\leqq$ i.o.d./pk < 2 .
20. number of objects with area < 50 and 0.5 $\leqq$ i.o.d./pk < 2 .
21. number of objects with i.o.d./pk < 0.5 .
22. number of objects with 3 $\leqq$ i.o.d./pk < 5 divided by number of objects with 2 $\leqq$ i.o.d./pk < 5 .
23. number of feature 17. and feature 18. with 120 $\leqq$ area < 220 divided by number of feature 17. + feature 18. with 60 $\leqq$ area < 220 .
24. number of feature 16. with 3 $\leqq$ i.o.d./pk < 5 divided by number of feature 16. with 2 $\leqq$ i.o.d./pk < 5 .
25. number of feature 16. with 120 $\leqq$ area < 220 divided by number of feature 16. with 60 $\leqq$ area < 220 .

*features provided by intervention of operator

Chapter 4
Modelling between-cell variation for the automated allocation of human chromosomes.

4.1 Introduction.

As noted in chapter 2, the precise stage of metaphase at which a cell's development is arrested varies from cell to cell. This means that the chromosomes in different cells exhibit different amounts of contraction. Consequently, it has become standard practice to normalise at least the values of the size features for each chromosome. Current practice for two widely used systems is to subtract the cell average of logged chromosome size from the logged value for each chromosome in a cell and to divide by the within-cell standard deviation of the logged values (Lundsteen et al, 1981) or to divide through the size value for each chromosome by the size of the median-sized chromosome in the same cell (Piper and Granum, 1989).

In this chapter three new possible approaches are considered. These are:

1. The transformation of each feature to marginal Normality when cell and class effects are allowed for in a linear model on the transformed scale. This is followed by removal of the cell effect on the transformed scale and a discrimination method based on multivariate Normality.
2. The regression of size-related features on an index of size for the cell, within each chromosome class.
3. The division of cells into classes according to the degree of contraction of the chromosomes with different sets of discriminant functions for each type of cell.

The performance of these different approaches is assessed by estimating percentage error-rates for three data sets.

4.2 Transformations to marginal Normality.

For each chromosome we may consider that a feature value, x_{ikl} , for the lth chromosome from the ith class in the kth cell is given by the model

$$x_{ikl} = \mu + C_i + Ce_k + \eta_{ikl} \qquad (4.1)$$

where μ is the overall mean value for the feature, C_i represents the fixed effect of the ith chromosome class, Ce_k represents the fixed effect of the kth cell and η_{ikl} is the residual error.

A transformation to marginal Normality may be sought by finding the power transformation defined by

$$x_{ij}^{(\lambda)} = \begin{cases} (x_{ij}^{\lambda} - 1)\lambda^{-1} & (\lambda \neq 0) \\ \ln(x_{ij}) & (\lambda = 0) \end{cases} \qquad (4.2)$$

which maximises

$$\Sigma_i\{-\tfrac{1}{2}n_i \ln(\underline{S}_i^{(\lambda)}) + (\lambda - 1)\Sigma_j \ln(x_{ij})\} \qquad (4.3)$$

where x_{ij} is the value of a feature for the jth chromosome from class i , $\ln(x_{ij})$ is the natural log transformation of x_{ij} , n_i is the number of chromosomes for class i , the summation for j is from 1 to n_i and each $\underline{S}_i^{(\lambda)}$ is given by

$$\underline{S}_i^{(\lambda)} = \Sigma_j\{x_{ij}^{(\lambda)} - E(x_{ij}^{(\lambda)})\}^2 \qquad (4.4)$$

where $E(x_{ij}^{(\lambda)})$ is the expected value from model (4.1) on the transformed scale. The estimate obtained from maximising expression (4.3) is the maximum-likelihood estimate of λ. Expression (4.3) is derived from noting that the likelihood of all the data on a given feature assuming a Normal distribution for each class is

$$\prod_i (2\pi)^{-\frac{1}{2}n_i} \sigma_i^{-n_i} \exp \Sigma_j[-\{x_{ij}^{(\lambda)} - E(x_{ij}^{(\lambda)})\}^2 (2\sigma_i)^{-2}] \prod_j |dx_{ij}^{(\lambda)}/dx_{ij}| \ , \qquad (4.5)$$

where σ_i is the feature variance for chromosomes from class i . To normalise the data a cell effect may then be removed on the appropriate scale. This approach assumes that division by the within-cell standard deviation, currently done for some features in some systems, is not necessary.

Marginal Normality does not ensure multivariate Normality of the features but in many cases the presence of nonnormality is often reflected in the marginal distributions (Gnanadesikan, 1977, page 163). It is also not certain that a necessary transformation will be one of the power family, but again it may be considered that this family encompasses a reasonably wide range of possibilities.

4.3 A regression model for the features related to size.

The model given by (4.1) assumes that the cell effect is the same across all classes on the transformed or the original scale. A different approach is to define an index of size for the cell and suppose that within a cell there is a regression relationship between the size-related features of each chromosome and the size index, for each chromosome class. For a simple linear relationship we have the model

$$x_{ikl} = \alpha_i + \beta_i z_k + \eta_{ikl}, \qquad (4.6)$$

where z_k is the size index and η_{ikl} has variance σ_i^2 . This model allows a differential adjustment for cell size to be made for every class by the use of an adjusted class mean vector for each cell. Model (4.6) may be contrasted with the multiplicative one of Piper and Granum (1989) used for size features. Model (4.6) gives x_{ikl} expectation $\alpha_i + \beta_i z_k$ and variance σ_i^2 whereas the multiplicative adjustment x_{ikl}/z_k has expectation and variance μ_i and γ_i^2 and x_{ikl} has expectation $\mu_i z_k$ and variance $z_k^2\gamma_i^2$. The differences, hence, lie in the non-zero intercept in (4.6) and the dependence of the variance on the size index. Figure 4.1 suggests that a non-zero intercept is appropriate for some classes and that the increase in variance is not as big as z_k^2 . This model has previously been used with average cell size as the size index by Gerdes (1979). Here, we consider the use of the median of a number of similarly sized

chromosomes. This index rather than average cell size might be expected to be more robust to chromosomes missing from the image taken of the cell. The overall median is not considered because for a normal cell it is close to the big drop in size between chromosome classes 12 and 13 (Figures 4.7 , 4.8 and 4.9).

4.4 Division of cells into classes according to the degree of contraction of the chromosomes.

A different proposal, also recently considered by others (Gerdes, Maahr and Lundsteen, 1989), is to state that the differences in the values of size features between chromosomes of very different states of contraction cannot be explained by the simple models proposed in the previous two sections. Instead it is assumed that different sets of discriminant functions are needed for chromosomes of different states of contraction. A simple approach is to define a number of categories. Here the division into 3 classes of 'small', 'medium' and 'large' chromosomes is considered. The adjectives 'small', 'medium' and 'large' are used to describe the size of a chromosome of a particular class relative to other chromosomes in the same class.

4.5 Application to Edinburgh, Copenhagen and Philadelphia data sets.

To anticipate the results obtained in chapter 5, Estimative multivariate Normal discrimination with common covariance matrices per Denver group was used to estimate percentage error-rates for all three normalisation procedures.

For the transformations to marginal Normality, complete cells from the first part of each data set were used to estimate the required power transformations and the second part of each data set was used to estimate the percentage error-rate. The splits into two parts were those used by Piper and Granum (1989) with the number of chromosomes in each part given in chapter 3. For the other two normalisation procedures, leave-one-cell-out cross-validation, as later described in chapter 5, was used to estimate percentage error-rates.

Prior probabilities of 2/46 for chromosome classes 1-22 and 1/46 for chromosome classes 23 and 24 were used for the all-male Edinburgh data set. The prior probabilities for classes 23 and 24 were changed to 3/92 and 1/92 for the other two data sets which had cells from both sexes.

The overall estimated percentage error-rate was taken as the weighted average of the individual class percentage error-rates using the specified prior probabilities as the weights.

No re-allocation of chromosomes to satisfy a normal karyotype as described in chapter 2 was performed.

4.5.1 Transformations to marginal Normality.

The model given by equation 4.1 is 'balanced' for the Edinburgh data set for which all cells are from males but not for the other two data sets. Balanced is here defined as meaning that the cell and chromosome effects are orthogonal. The lack of balance for the other two data sets is not expected to be severe because the difference between male and female cells is just the classes of the two sex chromosomes. The procedure followed here has been to ignore the lack of balance and to define the estimated cell and class effects for these two data sets as given by

$$(\bar{x}_{k.} - \bar{x}_{..}) \quad (k=1,...,K) \tag{4.7}$$

and

$$\Sigma_k \Sigma_l \{x_{ikl} - (\bar{x}_{k.} - \bar{x}_{..}) - \bar{x}_{..}\}(n_i)^{-1} \quad (i=1,...,24) \tag{4.8}$$

where K is the number of cells and a dot with a bar indicates summation over the particular index followed by division through by the number in the summation. This approach may be justified on the grounds that the current normalisation procedures make no use of the knowledge of the sex of a cell and therefore it is comparable with them. In some instances, also, the sex of the person from whom the cell has come will be unknown.

The values of λ considered for each feature were restricted to the range -5 to 5 at intervals of 0.5 . If the 95% confidence interval for λ , based on the asymptotic distribution of the maximum likelihood estimator, contained the value $\lambda = 1$ this value was assumed.

The omission of the division through by the within-cell standard deviation for each cell for the current normalisation is also examined.

4.5.2 A regression model for the features related to size.

Four of the twenty-seven available features which contain no exact linear dependence can be seen to be related to size when the values for the chromosomes in a cell are plotted against a size index for a cell. Three of these four features, area, length and c.h.p. are measures of chromosome size. Figures 4.1-4.6 show plots for the Edinburgh data set of:

1. Area versus the average area of the 25th and 26th smallest chromosomes in the same cell.
2. N.s.s.d. versus the average area of the 25th and 26th smallest chromosomes in the same cell.
3. N.s.s.d. versus the average length of the 25th and 26th smallest chromosomes in the same cell.
4. N.s.s.d. versus the average c.h.p. of the 25th and 26th smallest chromosomes in the same cell.
5. Length versus the average length of the 25th and 26th smallest chromosomes in the same cell.
6. C.h.p. versus the average c.h.p. of the 25th and 26th smallest chromosomes in the same cell.

The average of the 25th and 26th smallest chromosomes was chosen because it is the middle of a range of four similarly sized chromosomes for area, length and c.h.p. for complete male cells (Figures 4.7, 4.8 and 4.9) and the middle of a range of five similarly sized chromosomes for complete female cells.

Because the true chromosome class of a chromosome to be allocated is unknown, the regression equations were used to adjust the size-related elements of the mean feature vectors for every class for each new cell for which chromosomes are to be allocated. This contrasts with the current approach of normalising the feature vector for every chromosome. The simple linear regression model (4.1) was used for area, length and c.h.p. and an additional squared term was included for n.s.s.d. . This additional squared term was included because of the curvature evident in Figures 4.2, 4.3 and 4.4 . The explanatory feature or features were the average area, length and c.h.p. of the 25th and 26th smallest chromosomes in the same cell for area, length and c.h.p. respectively and the average area, length or c.h.p. (and the same feature

Figure 4.1
Edinburgh data
y axis - area of chromosomes in class
x axis - average area of 25th and 26th smallest chromosomes in corresponding cell

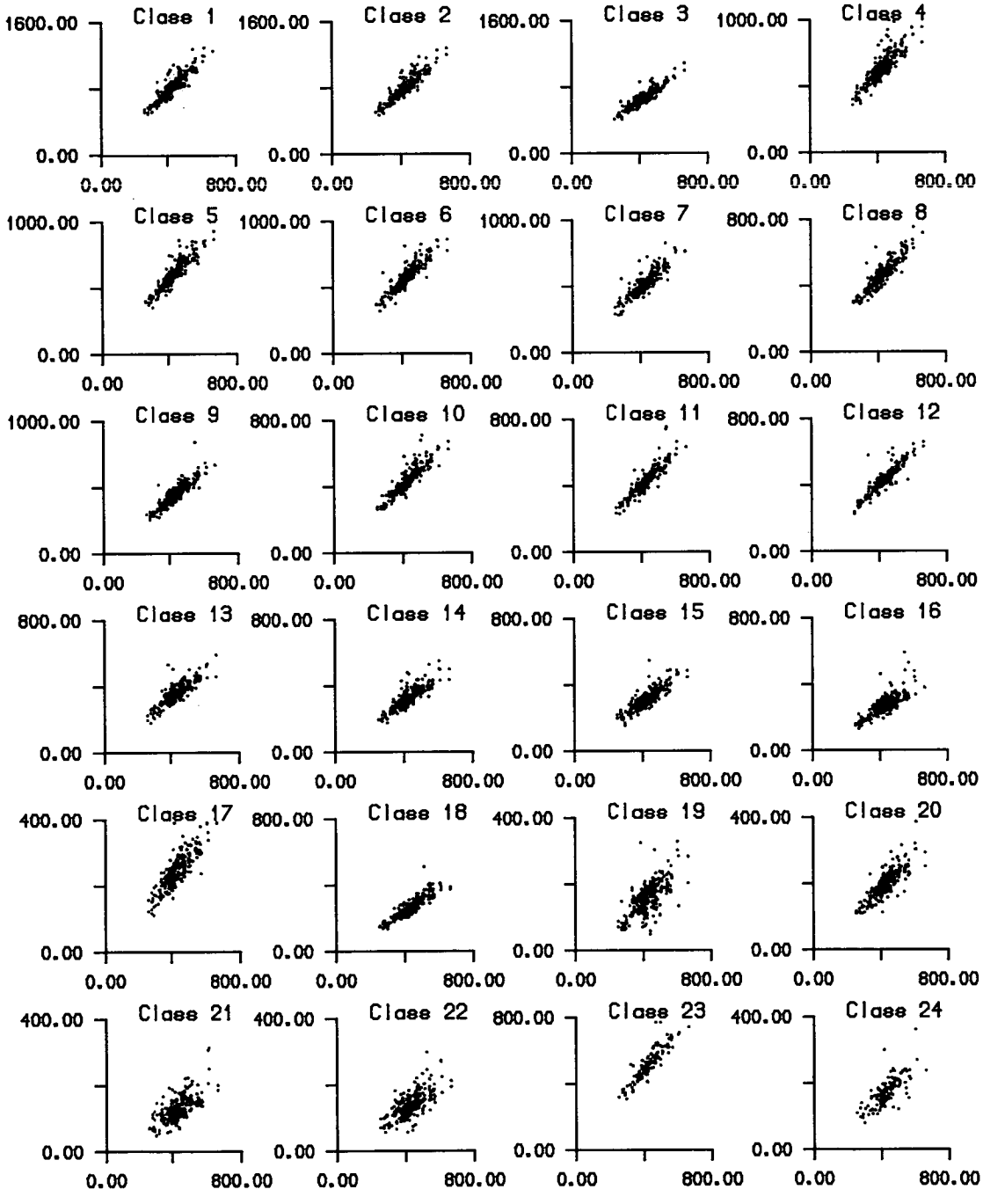


Figure 4.2
Edinburgh data
y axis - n.s.s.d. of chromosomes in class
x axis - average area of 25th and 26th smallest chromosomes in corresponding cell

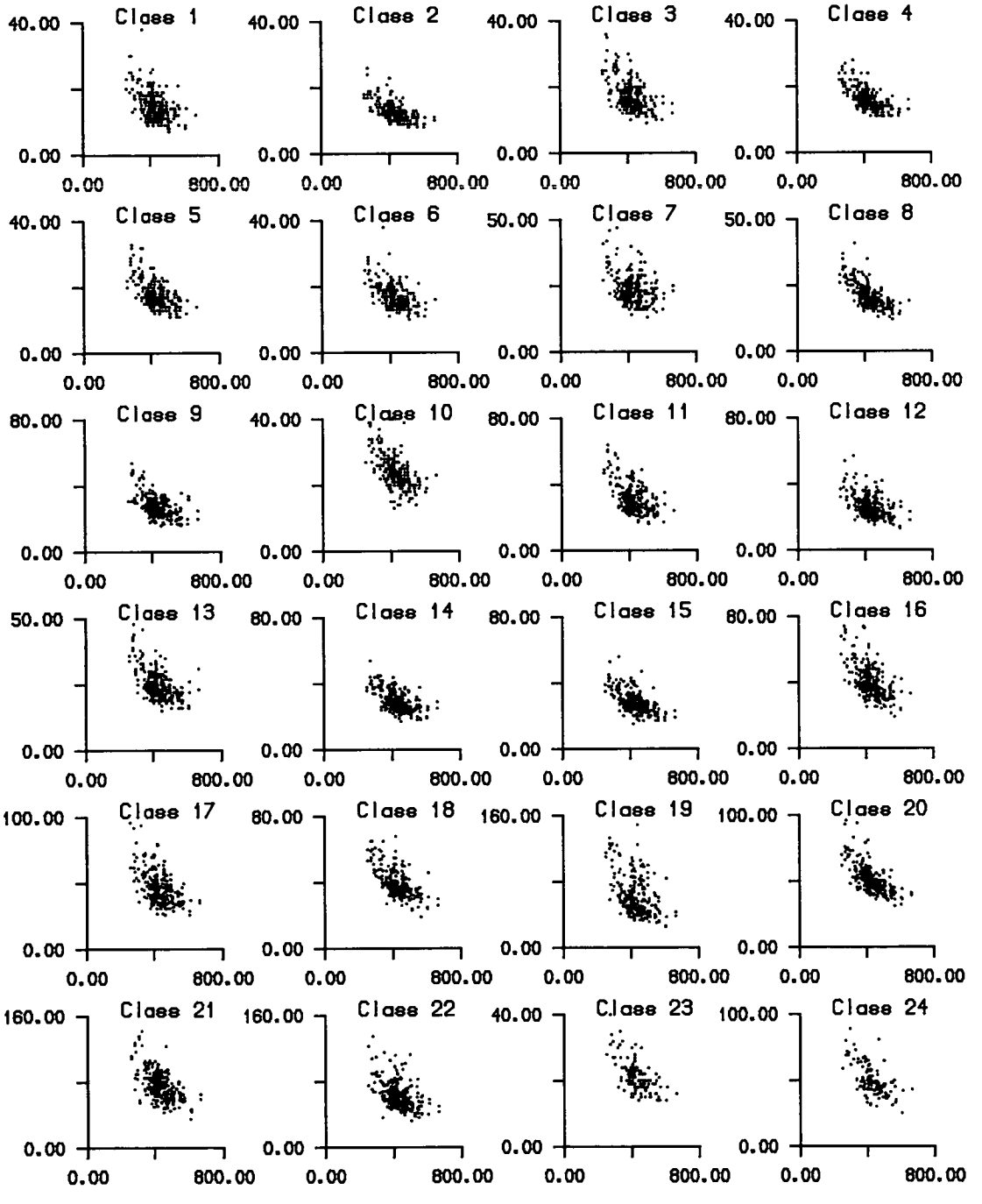


Figure 4.3
Edinburgh data
y axis - n.s.s.d. of chromosomes in class
x axis - average length of 25th and 26th smallest chromosomes in corresponding cell

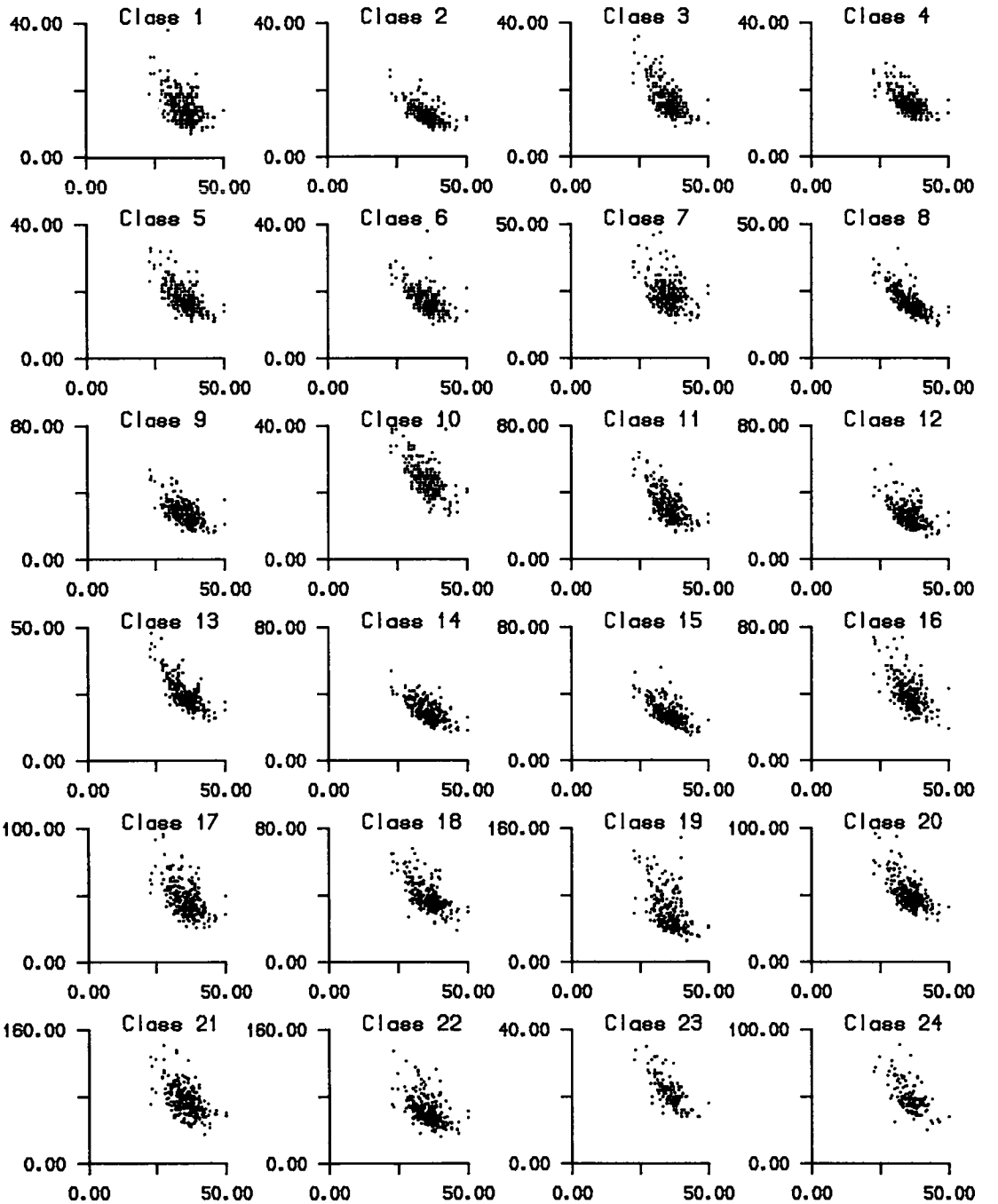


Figure 4.4
Edinburgh data
y axis - n.s.e.d. of chromosomes in class
x axis - average c.h.p. of 25th and 26th smallest chromosomes in corresponding cell

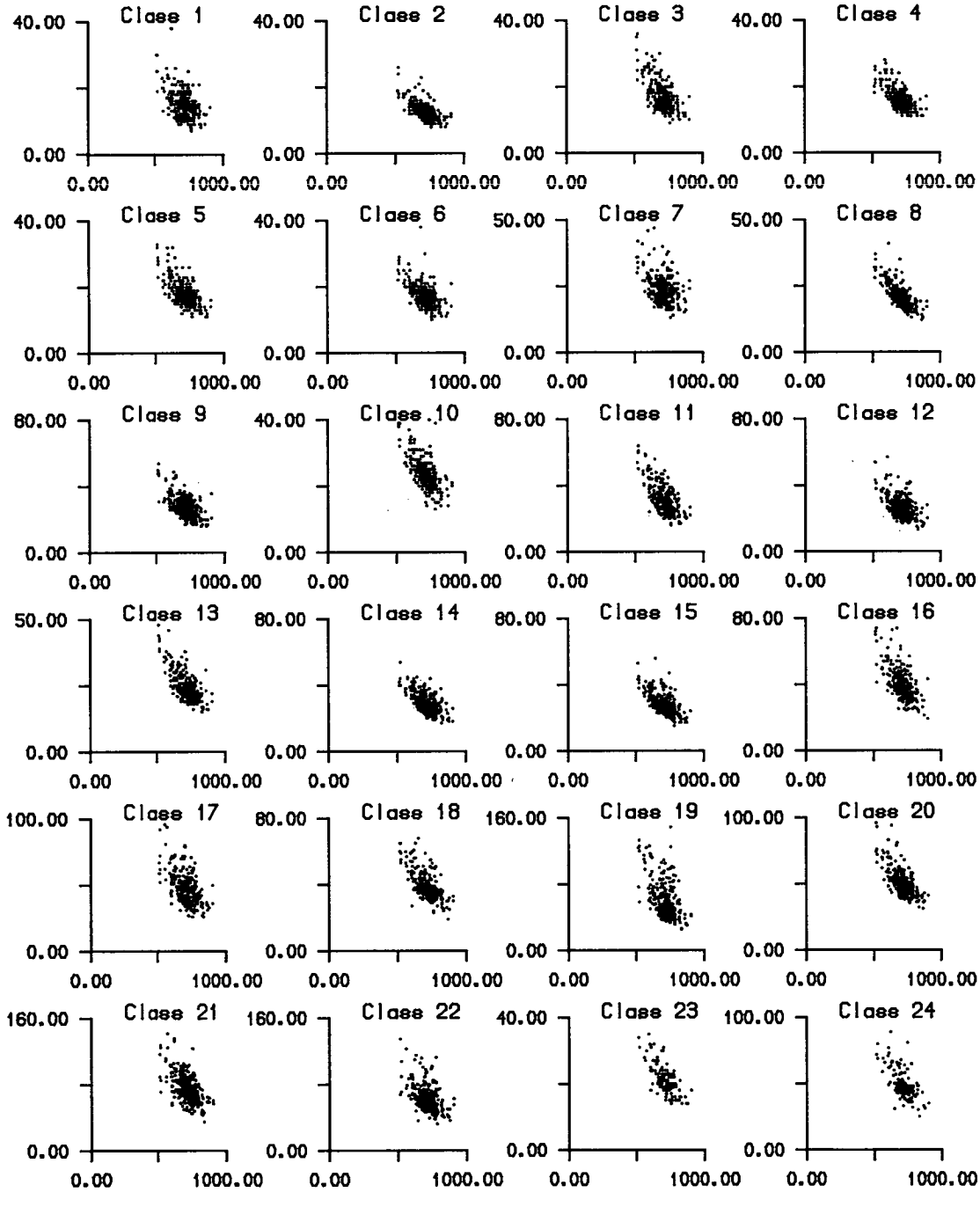


Figure 4.5
Edinburgh data
y axis - length of chromosomes in class
x axis - average length of 25th and 26th smallest chromosomes in corresponding cell

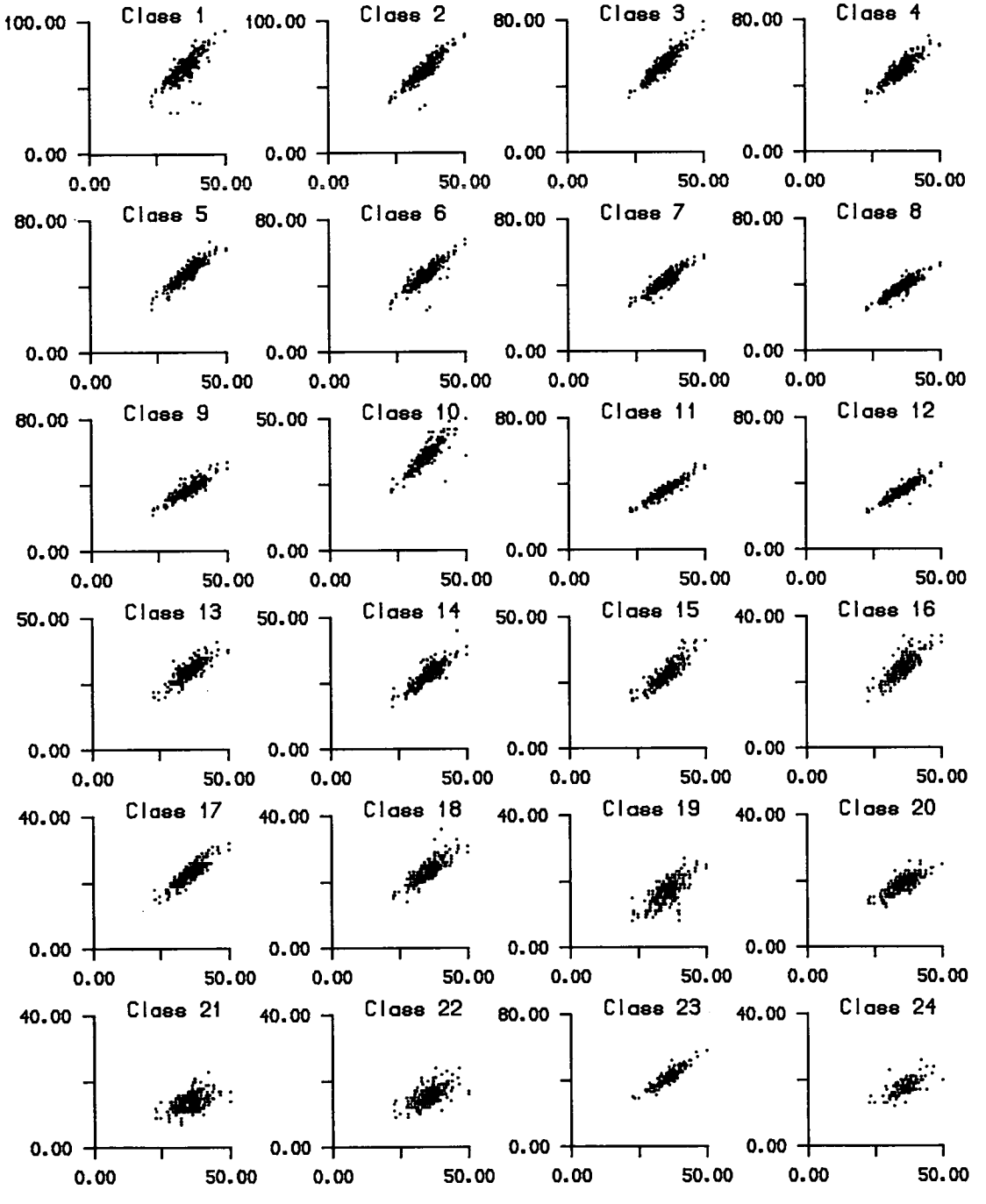


Figure 4.6
Edinburgh data
y axis - c.h.p. of chromosomes in class
x axis - average c.h.p. of 25th and 26th smallest chromosomes in corresponding cell

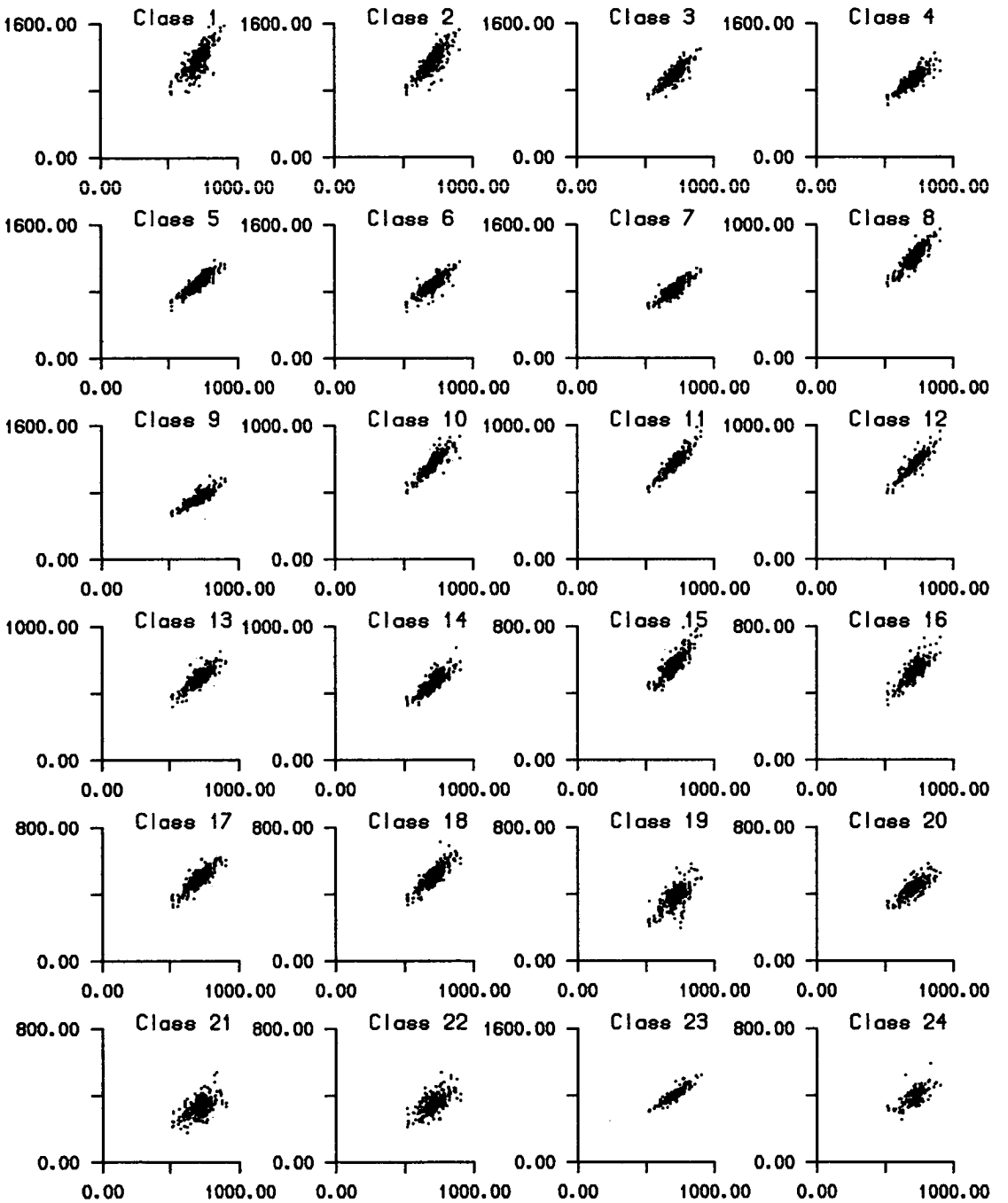


Figure 4.7
Edinburgh data
Complete cells

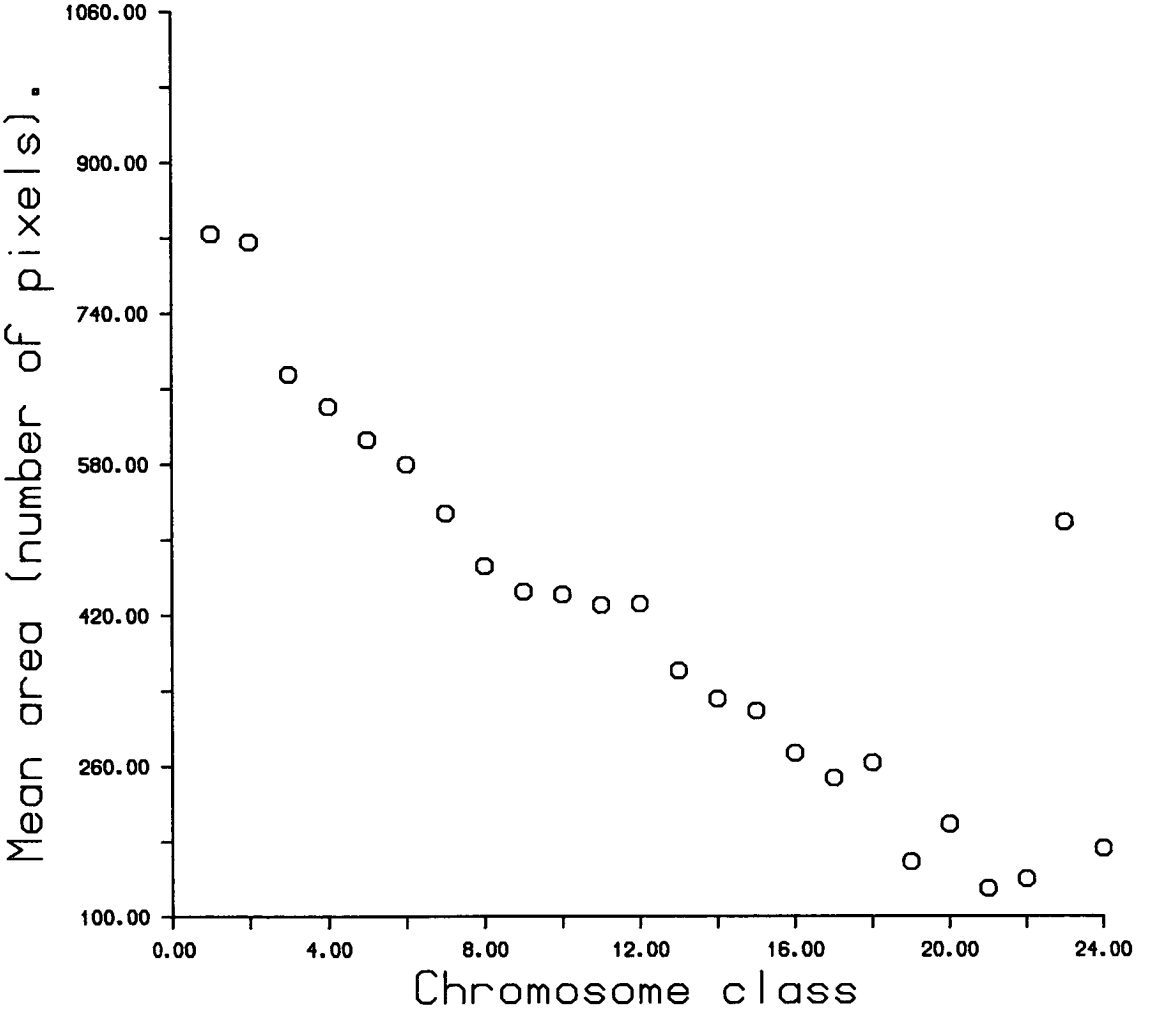


Figure 4.8
Edinburgh data
Complete cells

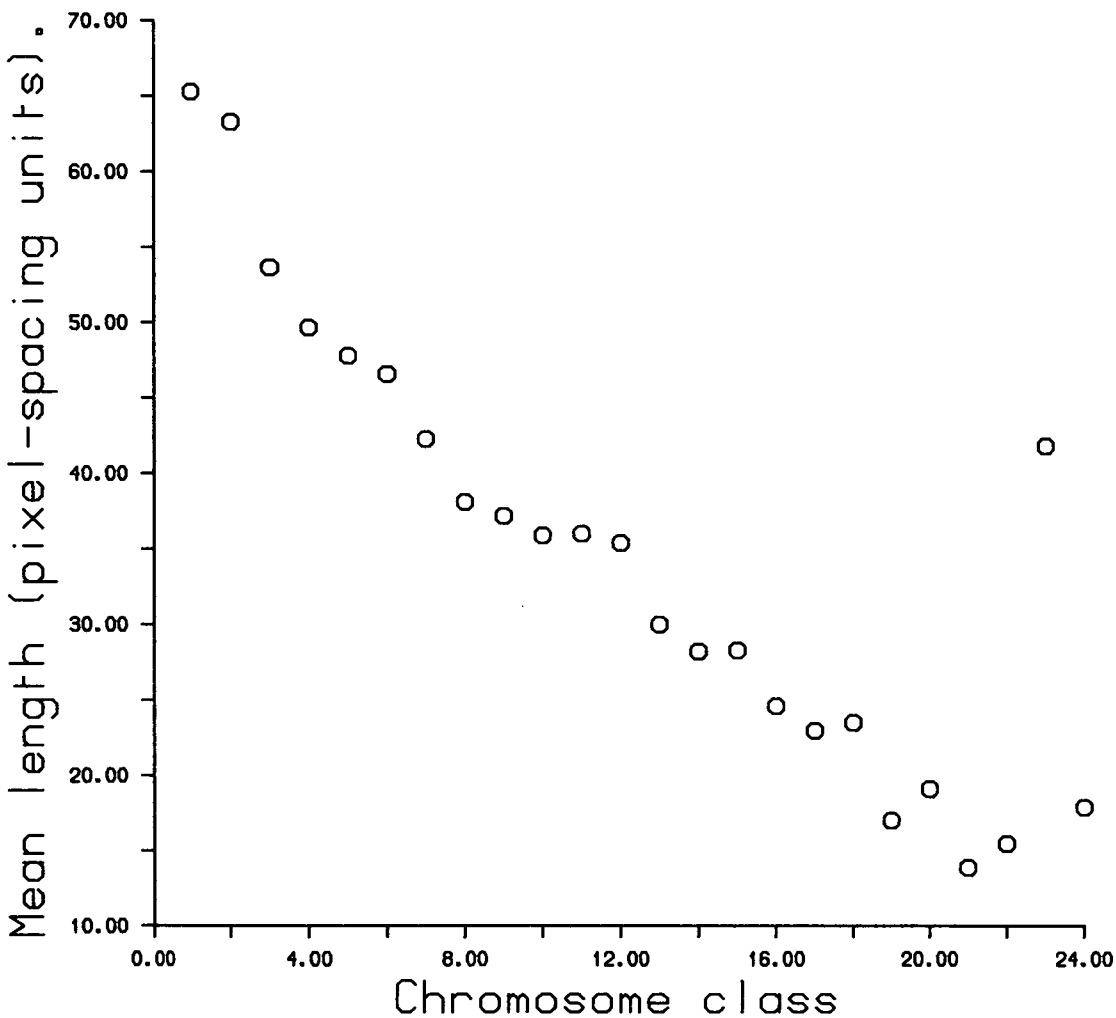
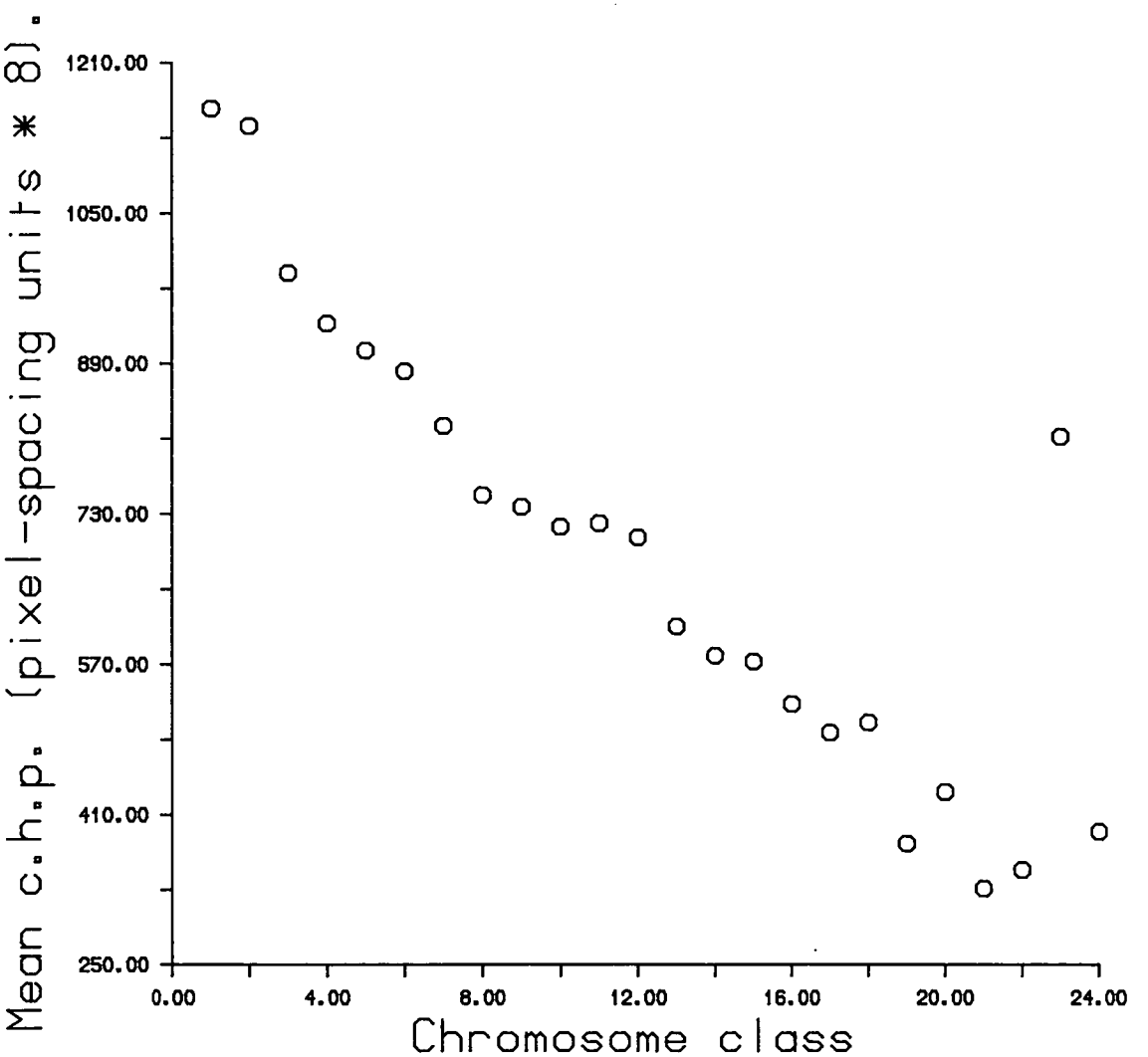


Figure 4.9
Edinburgh data
Complete cells



squared) of the 25th and 26th smallest values for n.s.s.d. . For the other features the current normalisation was used. Individual covariance matrices were calculated using

$$\Sigma_j(\underline{x}_{ij} - \bar{\underline{x}}_j)(\underline{x}_{ij} - \bar{\underline{x}}_j)^T(n_i - 1)^{-1} \quad , \tag{4.9}$$

where $\underline{x}_{ij}$ is the feature vector for the jth chromosome from class i and the elements of $\bar{\underline{x}}_j$ are the adjusted means for size features and the means of normalised features using the current normalisation for other features; the summation is over all chromsomes except the ones in the cell currently being allocated. The chromosomes in the cell being allocated were also excluded from the parameter estimates for the regression equations.

4.5.3 Division of cells into classes according to the degree of contraction of the chromosomes.

For each unnormalised data set the complete cells were divided into three approximately equal parts according to the mean length of all the chromosomes (Edinburgh data set) or the mean length of the autosomal chromosomes (Copenhagen and Philadelphia data sets). The number of complete cells was 50, 112 and 77 in the Edinburgh, Copenhagen and Philadelphia data sets respectively. For comparison, complete cells in the three data sets were also each randomly divided into 3 parts containing approximately equal numbers of cells with 'small', 'medium' and 'large' chromosomes. This random division was done to check that any effect on error rate was not due to the result of a reduction in sample size. The normalised values of the chromosomes using the current normalisation procedure were then used in the discrimination in both cases.

4.6 Results.

4.6.1 Transformations to marginal Normality.

Table 4.1 gives the transformations estimated by maximising expression (4.3). Table 4.2 gives the estimated percentage allocation error-rates using these combinations of transformations and removing a cell effect along with the estimated percentage error-rates for the usual normalisation procedure.

Table 4.1

Estimated power transformation for each feature.
(A 0 indicates natural log transformation.)

	Data sets		
	Edinburgh	Copenhagen	Philadelphia
Feature			
1	0	0	0
3	1	0	0.5
4	1	1.5	2
5	1	1	1.5
6	-0.5	0.5	-1
7	0	0	0
8	0	0	0
9	1	0.5	1.5
10	1	1	1
11	1	2	1.5
12	1	1	1
13	1	1.5	1
14	1	1	1
15	1	1	1.5
16	1	1	1
17	1	1	2
18	1	1	1
19	1	1	1
20	1	1	1.5
21	1	1	1.5
22	1	1.5	1
23	1	1.5	1.5
24	1	1	1
25	1	1	1
26	1	1	1
27	1.5	1.5	2
28	0	0	0

Table 4.2

Estimated percentage error-rates for transformations given in Table 4.1 using test-set method. (Estimated percentage error-rates for usual normalisation in brackets.)

Edinburgh data set

15.8(14.6)

Copenhagen data set

4.8(4.9)

Philadelphia data set

14.9(16.1)

Table 4.3 gives the estimated percentage error-rates for transforming just the size-related features and removing a cell effect. Finally, Table 4.4 gives the estimated percentage error-rates for the current normalisation procedure without the division through by the within-cell standard deviation.

4.6.2 A regression model for the features related to size.

The estimated percentage error-rates are given in Table 4.5 for this normalisation procedure for all four size-related features, using each of area, length or c.h.p. (and the same feature squared) as the explanatory features for n.s.s.d. , along with the current normalisation. The estimated percentage error-rates for this procedure excluding n.s.s.d. , which is not a size feature and for which the current normalisation was used, are given in Table 4.6 .

4.6.3 Division of cells into classes according to the degree of contraction of the chromosomes.

Table 4.7 gives the estimated percentage error-rates for the split into 'small', 'medium' and 'large' chromosomes and the random splits into three parts containing approximately equal numbers of cells with 'small', 'medium' and 'large' chromosomes.

4.7 Discussion.

The only agreement on transformations in Table 4.1 across the three data sets is for the four size-related features where a log transformation is indicated. This agrees with the multiplicative transformation currently done for area, length and c.h.p. . Table 4.2 shows that only for the Philadelphia data set does the use of all the transformations produce an interesting reduction in estimated percentage error-rate. Table 4.3 shows that use of just the transformations for the size-related features gives similar results to the use of all the transformations.

Table 4.4 indicates that the division through of features which are not measures of size or centromeric index by the within-cell standard deviation is unlikely to be important other than for numerical convenience. Indeed, this gives bigger estimated percentage error-rates for the Edinburgh and Philadelphia data sets.

Tables 4.5 and 4.6 show that the regression approach for all four or just three of the size-related features does not give smaller estimated percentage

Table 4.3

Estimated percentage error-rates for transformations given in Table 4.1 for size-related features only using test-set method. (Estimated percentage error-rates for usual normalisation in brackets.)

Edinburgh data set

16.0(14.6)

Copenhagen data set

5.0(4.9)

Philadelphia data set

14.8(16.1)

Table 4.4

Estimated percentage error-rates for current normalisation without the division by the within-cell standard deviation using the leave-one-out method.
(Estimated percentage error-rates for usual normalisation in brackets.)

Edinburgh data set	Copenhagen data set	Philadelphia data set
11.1(13.2)	5.1(4.4)	16.8(17.6)

Table 4.5

Estimated percentage error-rates for regression relationships for all size-related features and current normalisation for other features using leave-one-out method.
(Estimated percentage error-rates in order are for area, length and c.h.p. and the same feature squared as explanatory features for n.s.s.d. with estimated percentage error-rate for usual normalisation in brackets.)

Edinburgh data set

15.2,15.3,15.3(13.2)

Copenhagen data set

5.0, 5.4, 5.0(4.4)

Philadelphia data set

17.9,18.7,17.8(17.6)

Table 4.6

Estimated percentage error-rates for regression relationships for all size-related features except n.s.s.d. and current normalisation for other features using leave-one-out method. (Estimated percentage error-rates for usual normalisation in brackets.)

Edinburgh data set	Copenhagen data set	Philadelphia data set
15.0(13.2)	4.6(4.4)	18.0(17.6)

Table 4.7

Estimated percentage error-rates for 'small', 'medium' and 'large' chromosomes using leave-one-out method.
(Percentage error-rates for random splits into three given beneath.)

	Data sets		
	Edinburgh	Copenhagen	Philadelphia
Small	24.3	4.1	26.8
Medium	22.0	5.9	21.8
Large	20.1	6.2	25.1
Average	22.1	5.4	24.6
Random splits	26.5	5.3	24.2
	27.3	6.4	21.4
	18.3	5.4	25.1
Average	24.0	5.7	23.6
Figures for unsplit data	13.2	4.4	17.6

error-rates than the current normalisation procedure for any of the data sets. This may be because its assumptions about error variances are incorrect rather than because its assumptions about expectations are incorrect.

Finally, Table 4.7 shows some evidence for the Edinburgh and Copenhagen data sets of a reduction in average percentage error-rate for chromosomes divided into 'small', 'medium' or 'large' groups compared with random splits containing approximately equal numbers of 'small', 'medium' and 'large' chromosomes. Against this result must be set the increase in percentage error-rate due to reduced training set size as shown by comparison with the results for the total data sets also given in this table.

Chapter 5
Combining class information on variability in multivariate Normal discrimination for the automated allocation of human chromosomes.

5.1 Introduction.

As described in chapter 2, the time taken to calculate the discriminant scores is an important consideration for the automated allocation of human chromosomes as well as the error rate. The use of unequal covariance matrices in Estimative multivariate Normal discrimination has been found to give lower estimated error-rates than the use of a common covariance matrix using the test-set method of error rate estimation (Granum, 1982 and Piper, 1987) but the former is a computationally expensive procedure (Piper, 1987). One approach to reducing computation is to replace the estimated unrelated covariance matrices by just their main diagonal elements. For the features used by the Edinburgh system this has been found to produce a smaller estimated error-rate for the test-set method of error rate estimation for a typical data set (Piper, 1987). An alternative approach is to assume that the covariance matrices are related in ways which can reduce the computation required to evaluate the associated discriminant functions. Such assumptions also reduce the number of parameters. The statistical validity of the assumed relationships may be formally tested but in practice the estimated error-rates and computational time for typical data sets are more appropriate criteria. This is because the bias in the predicted distributions may be outweighed by the reduction in sampling variation due to the smaller number of parameters. Flury (1988, page 164) in particular has noted that estimates of constrained covariance matrices which are biased may give lower error rates in discrimination than unbiased estimates.

In this chapter six possible assumptions about relationships between covariance matrices are outlined. All of these may be easily incorporated into the Estimative approach to discrimination and for two of them a Bayesian or approximate Bayesian predictive approach to discrimination is also available (Aitchison, Habbema and Kay, 1977, Moran and Murphy, 1979 and Hawkins and Raath, 1982). The number of parameters in the predicted densities is given for the resulting eight procedures. For comparison, the number of parameters in

the predicted densities is also given for the Estimative approach with a common covariance matrix and unrelated diagonal and non-diagonal covariance matrices and for the Bayesian predictive approach with a common and unrelated covariance matrices. For all thirteen procedures, the number of calculations required to allocate a new object, when discriminant scores are calculated using formulae described later, is given. Figures of estimated percentage error-rate versus the square-root of average computational time for 46 chromosomes in a cell are given for all of these procedures for five human chromosome data sets.

5.2 Six assumed relationships between covariance matrices.

In the following we assume that the vectors of features for each class i are distributed as $N_p(\underline{\mu}_i, \underline{\Sigma}_i)$ where the $\underline{\Sigma}_i$ satisfy the assumed relationships between the class covariance matrices.

5.2.1 A common covariance matrix for the classes in a group.

We assume that, in general, we have c classes which can be put into a smaller number, g , of disjoint groups of classes, and then suppose that the covariance matrices $\underline{\Sigma}_i$ are the same for all the classes in a group.

For automated human karyotyping, the 24 chromosome classes may be allocated to seven so-called Denver groups (Book et al, 1960) on the basis of size and centromeric index. These groups are chromosome classes, 1-3, 4-5, 6-12 plus X, 13-15, 16-18, 19-20, 21-22 plus Y.

This assumption might be expected to be reasonable if the differences between the chromosome class covariance matrices are mainly due to the differences in variation of size or centromeric index.

5.2.2 Proportional covariance matrices.

We assume that the covariance matrix for class i is given by $c_i\underline{\Omega}$ (e.g., Manly and Rayner, 1987).

The assumption of proportionality between the chromosome class covariance matrices might be expected to be reasonable for size-related features.

5.2.3 Proportional covariance matrices within each of g groups.

The previous two assumptions may be combined so that proportionality of the covariance matrices is only assumed within each of g groups.

Again this combination of the previous two assumptions might be expected to be reasonable for size-related features.

5.2.4 Proportional common covariance matrices.

A different combination of the assumed relationships in 5.2.1 and 5.2.2 is to assume proportionality between the common covariance matrices for the g groups.

As before this assumption might be expected to be reasonable for size-related features.

5.2.5 Proportional diagonal covariance matrices.

We assume that the covariance matrix for class i is given by $d_i\underline{\Theta}$ where $\underline{\Theta}$ is diagonal.

This assumption might be expected to be reasonable for the size-related features if the correlations between the features are small.

5.2.6 Common principal components.

It is assumed that the covariance matrix for class i can be expressed as

$$\underline{B\Lambda_i B}^T \quad , \tag{5.1}$$

where $\underline{\Lambda}_i$ is diagonal and $\underline{B}$ is an orthogonal matrix (Flury, 1984). This assumption will be reasonable if there are common orthogonal linear combinations of the features which explain the variation in the feature values. This assumption does not seem likely to be true but the method reduces the computational time required to allocate a chromosome for the number of features considered later and also reduces the number of parameters.

5.3 Estimators for Estimative discrimination.

In the following, n_i denotes the number of observations for the ith class and $\underline{S}_i$ is the usual unbiased estimate of the covariance matrix for class i given by

$$\Sigma_j(\underline{x}_{ij} - \bar{\underline{x}}_i)(\underline{x}_{ij} - \bar{\underline{x}}_i)^T(n_i - 1)^{-1} \qquad (5.2)$$

where the summation is from j = 1 to j = n_i , $\underline{x}_{ij}$ is a vector of observed values and $\bar{\underline{x}}_i$ is the mean of the observed vectors for class i .

5.3.1 A common covariance matrix for the classes in a group (GC).

The common covariance matrix for the classes in a group may be estimated from the training data set by the usual unbiased estimate

$$\underline{S}_k = v_k^{-1}\Sigma_l(n_{kl} - 1)\underline{S}_{kl}, \qquad (5.3)$$

where kl refers to the lth class within the kth group and v_k is the number of degrees of freedom given by $\Sigma_l(n_{kl} - 1)$.

5.3.2 Proportional covariance matrices (P).

Eriksen (1987) has proved that unique maximum-likelihood estimates of proportional covariance matrices may be obtained by iterating till convergence between the following two equations with all the $\hat{c}_i$ initially set equal to one and c_1 constrained equal to one

$$\hat{\underline{\Omega}} = \Sigma_i(n_i - 1)\underline{S}_i\{\hat{c}_i(n_0 - c)\}^{-1} \qquad (5.4)$$

and

$$\hat{c}_i = tr(\hat{\underline{\Omega}}^{-1}\underline{S}_i)\ (p)^{-1} \qquad (i=2,...,24)\ , \qquad (5.5)$$

where

$$n_0 = \Sigma_i n_i \ ,$$

tr represents the trace of a matrix and p is the number of features.

5.3.3 Proportional covariance matrices within each of g groups (GP).

The estimator defined in 5.3.2 may be used within each of the g groups.

5.3.4 Proportional common covariance matrices (PG).

$(n_i - 1)$, $\underline{S}_i$ and $\hat{c}_i$ in (5.4) and (5.5) are replaced by v_k, $\underline{S}_k$ and $\hat{c}_k$ respectively.

5.3.5 Proportional diagonal covariance matrices (PD).

The log likelihood for the $\underline{S}_i$, excluding constant terms, is given by

$$-\tfrac{1}{2}\Sigma_i(n_i - 1)\{ \ln|d_i\underline{\Theta}| + tr(\underline{S}_i\underline{\Theta}^{-1})d_i^{-1}\} \ . \qquad (5.6)$$

Differentiating this with respect to d_i gives

$$-\tfrac{1}{2}(n_i - 1)\{pd_i^{-1} - tr(\underline{S}_i\underline{\Theta}^{-1})d_i^{-2}\} \qquad (5.7)$$

and equating to zero gives

$$\hat{d}_i = tr(\underline{S}_i\underline{\Theta}^{-1})\ p^{-1} \ . \qquad (5.8)$$

Differentiating (5.6) with respect to Θ_m , the mth diagonal element of $\underline{\Theta}$, gives

$$-\tfrac{1}{2}(n_0 - c)\Theta_m^{-1} + \tfrac{1}{2}\Sigma_i\Sigma_j(x_{ijm} - \bar{x}_{im})^2\ (\Theta_m^{2}d_i)^{-1} \ . \qquad (5.9)$$

Equating to zero gives

$$\hat{\theta}_m = \Sigma_i \Sigma_j (x_{ijm} - \bar{x}_{im})^2 \{(n_0 - c)d_i\}^{-1} \quad . \tag{5.10}$$

Hence,

$$\underline{\hat{\theta}} = \Sigma_i (n_i - 1)\mathrm{diag}(\underline{S}_i)\{(n_0 - c)d_i\}^{-1} \quad . \tag{5.11}$$

As for the estimation of proportional non-diagonal covariance matrices, the equations may be solved iteratively with all the $\hat{d}_i$, (i=2,...,c), initially set equal to one and d_1 constrained equal to one.

5.3.6 Common principal components (E).

Flury (1984) suggests that maximum-likelihood estimates of $\underline{B}$ and the $\underline{\Lambda}_i$ may be obtained by solving

$$\underline{B}_m^T \{\Sigma_i n_i (\lambda_{im} - \lambda_{ir})(\lambda_{im}\lambda_{ir})^{-1} \underline{S}_i\} \underline{B}_r = 0 \qquad (m,r = 1,...,p \; ; \; m \neq r) \tag{5.12}$$

where $\underline{B}_m$ is the mth column of $\underline{B}$ and λ_{im} is the mth diagonal element of $\underline{\Lambda}_i$, subject to the orthonormality conditions

$$\underline{B}^T \underline{B} = \underline{I}_p \tag{5.13}$$

and also to the conditions

$$\lambda_{im} = \underline{B}_m^T \underline{S}_i \underline{B}_m \quad (i=1,...24, \; m=1,...,p) \quad . \tag{5.14}$$

Algorithms to solve these equations are available (Flury and Constantine, 1985 and Clarkson, 1988a and Clarkson, 1988b).

5.4 Bayesian predictive densities.

The Bayesian predictive approach to discrimination uses predictive densities derived from a prior distribution for the parameters and the data (Aitchison, Habbema and Kay, 1977). The predictive density for a class c_i is given by

$$\int_{\Theta} p(\underline{x}|c_i\ ,\ \underline{\Theta})\ p(\underline{\Theta}|z)\ d\underline{\Theta} \tag{5.15}$$

where $p(\underline{x}|c_i\ ,\ \underline{\Theta})$ is the probability density function of $\underline{x}$ for class i with parameter vector $\underline{\Theta}$ and z is given data. The $p(\underline{\Theta}|z)$ is a posterior density function for $\underline{\Theta}$ based on a prior distribution for the parameters and the data. In the following sub-sections we only consider the procedures resulting from the use of vague prior information about $\underline{\Theta}$ for the conjugate prior distribution.

5.4.1 A common covariance matrix for the classes in a group (BGC).

Assuming a vague Normal-Wishart prior distribution for $(\underline{\mu}_{kl},\underline{\Sigma}_k)$ (Aitchison and Dunsmore, 1975, page 21), the mean for the lth class in the kth group and the covariance matrix for the kth group, the predictive density for the lth class in the kth group is given by

$$c_{kp}|\underline{R}_{kl}|^{-\frac{1}{2}}\{1 + (\underline{x} - \underline{\bar{x}}_{kl})^T\underline{R}_{kl}^{-1}(\underline{x} - \underline{\bar{x}}_{kl})\}^{-\frac{1}{2}(v_k+1)} \tag{5.16}$$

where

$$c_{kp} = \Gamma\{\tfrac{1}{2}(v_k + 1)\}\pi^{-\frac{1}{2}p}\Gamma\{\tfrac{1}{2}(v_k - p + 1)\} \tag{5.17}$$

and

$$\underline{R}_{kl} = v_k \underline{S}_k (1 + n_{kl}^{-1}) \tag{5.18}$$

(Aitchison, Habbema and Kay, 1977).

5.4.2 Proportional covariance matrices (BP).

We re-define the proportional covariance matrix assumption so that the covariance matrix for class i is given by $a_i \underline{\Delta}$, where $\Sigma_i a_i = 1$. The predictive density obtained by Hawkins and Raath (1982) for the ith class using a vague Normal-Wishart prior distribution for $(\underline{\mu}_i, \underline{\Delta})$ is

$$C[n_i/\{\hat{a}_i(n_i + 1)\}]^{\frac{1}{2}p}[\{1 + (\underline{x} - \bar{\underline{x}}_i)^T \underline{T}^{-1}(\underline{x} - \bar{\underline{x}}_i)n_i\}/\{\hat{a}_i(n_i + 1)\}]^{-\frac{1}{2}r} , \tag{5.19}$$

where C is a constant, $\hat{a}_i$ is an estimate of the proportionality factor for the ith class, $\underline{T}$ is

$$\Sigma_i (n_i - 1)\underline{S}_i \hat{a}_i^{-1} \tag{5.20}$$

and $r = n_0 - c + 1$. Here, a_i appears as a parameter to be estimated because for the exchangeable Dirichlet prior distribution with parameter α for the set $\{a_i\}$ used by Hawkins and Raath (1982) , i.e.,

$$f(a_1, a_2, \ldots, a_k) \propto \begin{cases} \prod_i a_i^{\alpha - 1}, & \Sigma a_i = 1 \\ 0, & \text{otherwise} \end{cases} , \tag{5.21}$$

it may not in general be integrated out.

5.5 The number of calculations required to allocate one new object and the number of parameters in the predicted density, for each procedure.

The number of calculations required to allocate one new object, when the discriminant scores are calculated as described in sub-sections 5.5.1 and 5.5.2 , is given to indicate the likely c.p.u. time required compared with the use of an estimated common covariance matrix (C), estimated unrelated non-diagonal (U) and diagonal (UD) covariance matrices and a common (BC) and unrelated covariance matrices (BU) in the Bayesian predictive approach. As mentioned in chapter 2, it is assumed that the cost of misallocating a class i chromosome to class i' ($i \neq i'$) is the same for all i and i' .

5.5.1 Estimative procedures.

For the Estimative procedures, except C, the discriminant score to be calculated for each class is taken to be

$$- \ln |\hat{\underline{\Sigma}}_i| - (\underline{x} - \bar{\underline{x}}_i)^T \hat{\underline{\Sigma}}_i^{-1} (\underline{x} - \bar{\underline{x}}_i) + 2*\ln(P_i) \qquad (5.22)$$

where $\hat{\underline{\Sigma}}_i$ represents any of the estimators of the covariance matrix for class i discussed above and P_i is the prior probability for class i . For procedure C the discriminant score to be calculated for each class is taken to be

$$\underline{x}^T \hat{\underline{\Sigma}}^{-1} \bar{\underline{x}}_i - \tfrac{1}{2} \bar{\underline{x}}_i^T \hat{\underline{\Sigma}}^{-1} \bar{\underline{x}}_i + \ln(P_i) \qquad (5.23)$$

where $\hat{\underline{\Sigma}}$ is the usual unbiased estimate of a common covariance matrix for all classes given by

$$\hat{\underline{\Sigma}} = \Sigma_i (n_i - 1) \underline{S}_i (n_o - c)^{-1} \quad . \qquad (5.24)$$

These discriminant functions are derived from the formulae for the estimated posterior probabilities for each class by taking natural logs and omitting terms

which are constant for all classes.

For procedures U, GC, P, GP and PG each discriminant score can be calculated more quickly if $\hat{\underline{\Sigma}}_i^{-1}$ is expressed in terms of its Cholesky decomposition. Expansion of the quadratic form for procedures GC, P, GP, PG, PD and E also results in quicker computation of the set of discriminant scores for the chromosome data for the numbers of features considered in section 5.6 . This is because of the equivalence of some of the resulting quadratic forms and also because parts of the expansion do not depend on the data for the new object and may therefore be stored. The first term in the expansion of the quadratic form for procedure E is calculated by working out

$$\underline{w}^T = \underline{x}^T \hat{\underline{B}}$$

and storing this vector with each element squared.

5.5.2 Bayesian predictive procedures

The Bayesian predictive formulae for BC and BU are similar to that for BGC with v_k , $\underline{S}_k$ and $\bar{\underline{x}}_{kl}$ replaced by $v = \Sigma_i(n_i - 1)$, $\hat{\underline{\Sigma}}$ and $\bar{\underline{x}}_i$ or $v_i = (n_i - 1)$, $\underline{S}_i$ and $\bar{\underline{x}}_i$ respectively and n_{kl} replaced by n_i . For procedures BU and BGC a discriminant score for each class can be calculated by multiplying the predictive density by P_i . For procedures BC and BP a score may be calculated for each object by multiplying the predictive density by P_i and taking the $2/(n_0 - c + 1)$ power. Again, the number of calculations for the set of discriminant scores for each procedure may be reduced by use of the Cholesky decomposition and, for the chromosome data for the number of features considered below, by expansion of the quadratic form for procedures BC, BGC and BP.

5.5.3 Summary of number of calculations required to allocate one new object and the number of parameters in the predicted density, for each procedure.

Table 5.1 summarises the number of multiplications, power operations, divisions and additions and subtractions required to allocate one new object for each of the thirteen procedures described above when the sets of discriminant scores are calculated as described in the previous sub-sections. It is assumed in Table 5.1 that the number of additions required to evaluate

Table 5.1

Numbers of calculations to allocate one object when the sets of discriminant scores are calculated as described in sub-sections 5.5.1 and 5.5.2 .
(g groups, c classes and p features)

	Multiplications, power operations and divisions.		
Procedure	Multiplications	Power operations	Divisions
C	cp	–	–
BC	$\frac{1}{2}p(p+3+2c)+c$	–	c
U	$\frac{1}{2}cp(p+3)$	–	–
BU	$\frac{1}{2}cp(p+3)+c$	c	–
UD	2cp	–	–
GC	$\frac{1}{2}p\{gp+(3g+2c)\}$	–	–
BGC	$\frac{1}{2}p\{gp+(3g+2c)\}+2c$	c	–
P	$\frac{1}{2}p(p+3+2c)+c-1$	–	–
BP	$\frac{1}{2}p(p+3+2c)+c$	–	c
GP	$\frac{1}{2}p\{gp+(3g+2c)\}+c-g$	–	–
PG	$\frac{1}{2}p(p+3+2c)+g-1$	–	–
PD	$(c+2)p+c-1$	–	–
E	$p(p+c+1)$	–	cp

Table 5.1 (continued)

Number of calculations to allocate one object when the sets of discriminant scores are calculated as described in sub-sections 5.5.1 and 5.5.2 .
(g groups, c classes and p features)

Procedure	Additions and subtractions
C	$cp+c$
BC	$\frac{1}{2}p(p+3+2c)+2c$
U	$\frac{1}{2}cp(p+3)+cp+c$
BU	$\frac{1}{2}cp(p+3)+c+cp$
UD	$2cp+c$
GC	$\frac{1}{2}p\{gp+(3g+2c)\}+2c$
BGC	$\frac{1}{2}p\{gp+(3g+2c)\}+2c$
P	$\frac{1}{2}p(p+3+2c)+2c$
BP	$\frac{1}{2}p(p+3+2c)+2c$
GP	$\frac{1}{2}p\{gp+(3g+2c)\}+2c$
PG	$\frac{1}{2}p(p+3+2c)+2c$
PD	$(c+1)p+2c$
E	$p(p+2c)+2c$

$$\underline{z}^T \underline{L}\underline{L}^T \underline{z}$$

where $\underline{L}$ is a lower-triangular matrix, is equal to the number of multiplications. Table 5.2 summarises the number of parameters in the predicted density for each of the thirteen procedures.

5.6 Application of the thirteen procedures to five human chromosome data sets.

5.6.1 Five data sets

The thirteen procedures described above were applied to the Edinburgh, Copenhagen , Philadelphia and two special Copenhagen data sets described in chapter 3. For all the data sets the normalisation of features for between-cell variation was that currently used and described in chapter 3.

5.6.2 Estimation of percentage error-rates.

Two methods of estimating percentage error-rates were used (McLachlan, 1986). The first, subsequently referred to as the leave-one-out method, leaves out all the chromosomes in a cell when they are allocated from the parameter estimates or the data used to derive the predictive densities. The leave-one-out method provides an almost unbiased estimate of the percentage error-rate. The second, subsequently referred to as the test-set method, splits the data into two sets and uses one set to estimate parameters or derive predictive densities and the other set to estimate percentage error-rates.

The leave-one-out method was chosen instead of the bootstrap method despite the fact that it may have a larger variance (Efron, 1979) because of the heavy computational burden of the latter method for such large data sets. The bootstrap method operates by taking random samples of the same size as the original training data from the training data with replacement. For each sample, all the observations in a class are allocated using the discriminant functions derived from all the observations in the sample. This gives the so-called apparent error rate for the class for the sample. The bootstrap estimate of the bias in this estimate of error for the sample is then given by

Table 5.2

Summary of number of parameters in predicted densities for each procedure.
(g groups, c classes and p features)

PROCEDURE	PARAMETERS
C	$cp+\frac{1}{2}p(p+1)$
BC	$cp+\frac{1}{2}p(p+1)$
U	$cp+\frac{1}{2}cp(p+1)$
BU	$cp+\frac{1}{2}cp(p+1)$
UD	$2cp$
GC	$cp+\frac{1}{2}gp(p+1)$
BGC	$cp+\frac{1}{2}gp(p+1)$
P	$cp+\frac{1}{2}p(p+1)+c-1$
BP	$cp+\frac{1}{2}p(p+1)+c-1$
GP	$cp+\frac{1}{2}gp(p+1)+c-g$
PG	$cp+\frac{1}{2}p(p+1)+g-1$
PD	$cp+p+c-1$
E	$cp+p^2+cp$

$$\hat{e}_i - App_i \qquad (5.25)$$

where $\hat{e}_i$ is the error rate estimate for the original data for class i using the discriminant functions derived from the bootstrap sample and App_i is the apparent error rate for class i . Averaging the estimated bias over all the bootstrap samples then gives the bootstrap estimate of the bias for the apparent error rate for the original data. It is thought that the number of bootstrap samples for a satisfactory estimate of bias should be at least 100 (Jain, Dubes and Chen, 1987).

5.6.3 Leave-one-out formulae

Leaving out object j from the ith class the estimate of the covariance matrix becomes

$$(n_i - 2)^{-1} \{(n_i - 1)\underline{S}_i - (\underline{x}_{ij} - \bar{\underline{x}}_i)(\underline{x}_{ij} - \bar{\underline{x}}_i)^T n_i (n_i - 1)^{-1}\} \qquad (5.26)$$

and has inverse

$$(n_i - 2)\,[\underline{A}_i^{-1} + \underline{A}_i^{-1}(\underline{x}_{ij} - \bar{\underline{x}}_i)(\underline{x}_{ij} - \bar{\underline{x}}_i)^T \underline{A}_i^{-1}\{(n_i - 1)n_i^{-1} - d^2_{ji}(n_i - 1)^{-1}\}^{-1}] \qquad (5.27)$$

where $\underline{A}_i = (n_i - 1)\underline{S}_i$ and d^2_{ji} is the estimated Mahalanobis distance between $\underline{x}_{ij}$ and $\bar{\underline{x}}_j$. The determinant becomes

$$(n_i - 2)^{-p}(n_i - 1)^p |\underline{S}_i| \{1 - d^2_{ji} n_i (n_i - 1)^{-2}\} \ . \qquad (5.28)$$

The corresponding results for grouped classes when the ith class is in the kth group are

$$(v_k - 1)^{-1}\{v_k \underline{S}_k - (\underline{x}_{ij} - \bar{\underline{x}})(\underline{x}_{ij} - \bar{\underline{x}}_i)^T n_i (n_i - 1)^{-1}\} \qquad (5.29)$$

and

$$(v_k - 1)[\underline{A}_k^{-1} + \underline{A}_k^{-1}(\underline{x}_{ij} - \bar{\underline{x}}_j)(\underline{x}_{ij} - \bar{\underline{x}}_j)^T\underline{A}_k^{-1}\{(n_i - 1)n_i^{-1} - d^2_{ji}v_k^{-1}\}^{-1}] \qquad (5.30)$$

where

$$\underline{A}_k = \Sigma_l(n_{kl} - 1)\underline{S}_{kl}$$

and the estimated Mahalanobis distance, d^2_{ji} now uses $\underline{S}_k^{-1}$ instead of the $\underline{S}_i^{-1}$ above and

$$(v_k - 1)^{-p}v_k^p|\underline{S}_k|\{1 - d^2_{ji}n_i(n_i - 1)^{-1}v_k^{-1}\} \quad . \qquad (5.31)$$

These formulae were repeatedly used to update $\underline{S}_i$, $\underline{S}_i^{-1}$, $|\underline{S}_i|$, $\underline{S}_k$, $\underline{S}_k^{-1}$ and $|\underline{S}_k|$ as the chromosomes in a cell were sequentially omitted.

For UD, the determinant and inverse of diag($\underline{S}_i$) excluding the jth object were simply obtained by calculating the diagonal elements of (5.26) and then multiplying the elements and inverting each element respectively.

5.6.4 Estimation of proportionality factor for procedure BP.

Maximum-likelihood estimates of the "plug-in" parameters a_i were used. The method suggested by Hawkins and Raath (1982) for estimating the a_i using the mode of the marginal posterior distribution of the a_i was tried for various values of α for their Dirichlet prior but convergence could not be achieved.

5.6.5 Convergence criterion for procedures P, BP, GP, PG and PD.

The convergence criterion for estimating the proportionality factors for these procedures was that the absolute sum of the differences for two iterations was less than 0.0001 times the number of parameters + 1 being estimated.

5.6.6 Feature subset selection.

For the Edinburgh, Copenhagen and Philadelphia data sets subsets of size 3, 6, 9, 12, 16, 20 and 24 were chosen from the full number of 28 or 27 available features containing no exact linear dependencies. For the Edinburgh data set, these subsets were chosen by the forward-selection procedure MSEPCOR described in (Piper, 1987) and a forward selection version of a method described by Fatti and Hawkins (1986) whilst for the Philadelphia and Copenhagen data sets only the former method was used because of its generally superior performance. The version of the method described by Fatti and Hawkins (1986) adopted was that which uses Fisher's approach to combining independent test statistics for the three test statistics derived by them for each feature. Features were selected using the complete data sets. The orders in which the features were chosen are given in Table 5.3 . The different orders in which the features were chosen reflects the different sources of the three data sets (Piper and Granum, 1989).

All of the 11 features used in the WDD classifier described by Lundsteen, Gerdes and Maahr (1986) and specified in chapter 3 were used for the two special Copenhagen data sets.

5.6.7 Prior probabilities and overall percentage error-rate.

Prior probabilities of 2/46 for chromosome classes 1-22 and 1/46 for chromosome classes 23 and 24 were used for the Edinburgh data set which contains only cells from males. The prior probabilities for classes 23 and 24 were changed to 3/92 and 1/92 for all the other data sets which had cells from both sexes.

The overall estimated percentage error-rate was taken as the weighted average of the individual class percentage error-rates using the specified prior probabilities as the weights. No re-allocation of chromosomes to satisfy a normal karyotype as described in chapter 2 was performed.

5.7 Results.

To show the trade-off between the computational time to allocate 46 chromosomes in a cell and percentage error-rate, estimated percentage error-rate was plotted against the square-root of the average of ten computer c.p.u. times obtained using the same operands (Figures 5.1-5.12). The

Table 5.3

Orders in which features were selected by feature-selection procedures.
(Fatti-Hawkins procedure in brackets)

	Data sets		
	Edinburgh	Copenhagen	Philadelphia
Order			
1	2(2)	2	7
2	12(27)	14	12
3	22(4)	5	11
4	11(7)	22	18
5	14(12)	13	5
6	4(22)	11	14
7	10(1)	12	10
8	13(25)	18	20
9	7(14)	20	8
10	18(10)	26	22
11	20(13)	25	25
12	23(26)	24	21
13	21(8)	21	23
14	24(6)	7	3
15	26(11)	16	24
16	25(23)	23	26
17	6(21)	6	17
18	19(5)	10	13
19	5(15)	17	6
20	17(19)	19	16
21	16(9)	3	9
22	3(24)	9	19
23	9(17)	15	15
24	8(3)	27	27
25	15(20)	4	28
26	27(18)	8	4
27	28(16)	28	2
28	1(28)	_	_

Figure 5.1

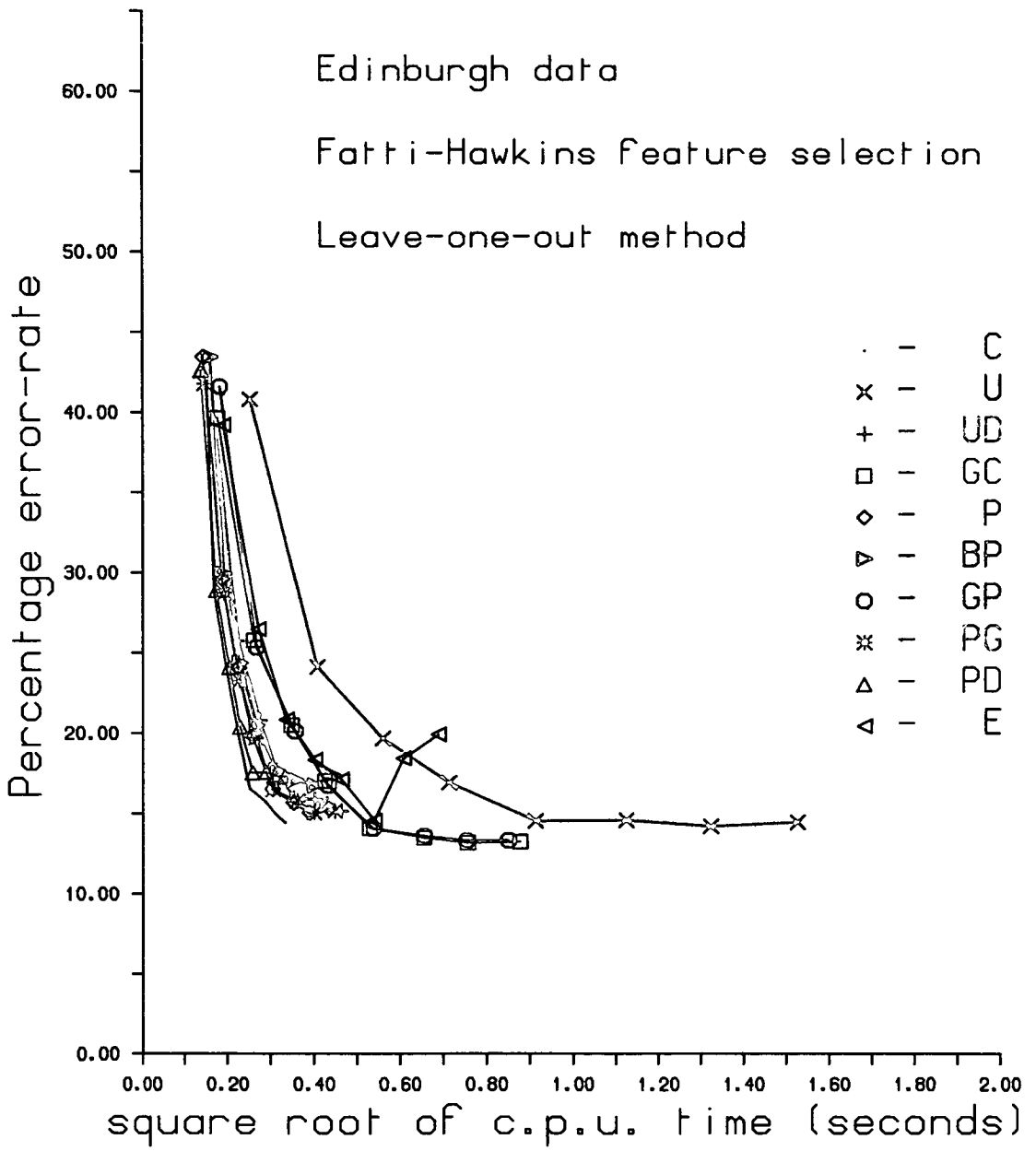

Edinburgh data
Fatti-Hawkins feature selection
Leave-one-out method
Percentage error-rate
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)
. - C
× - U
+ - UD
□ - GC
◇ - P
▷ - BP
○ - GP
※ - PG
△ - PD
◁ - E

Figure 5.2

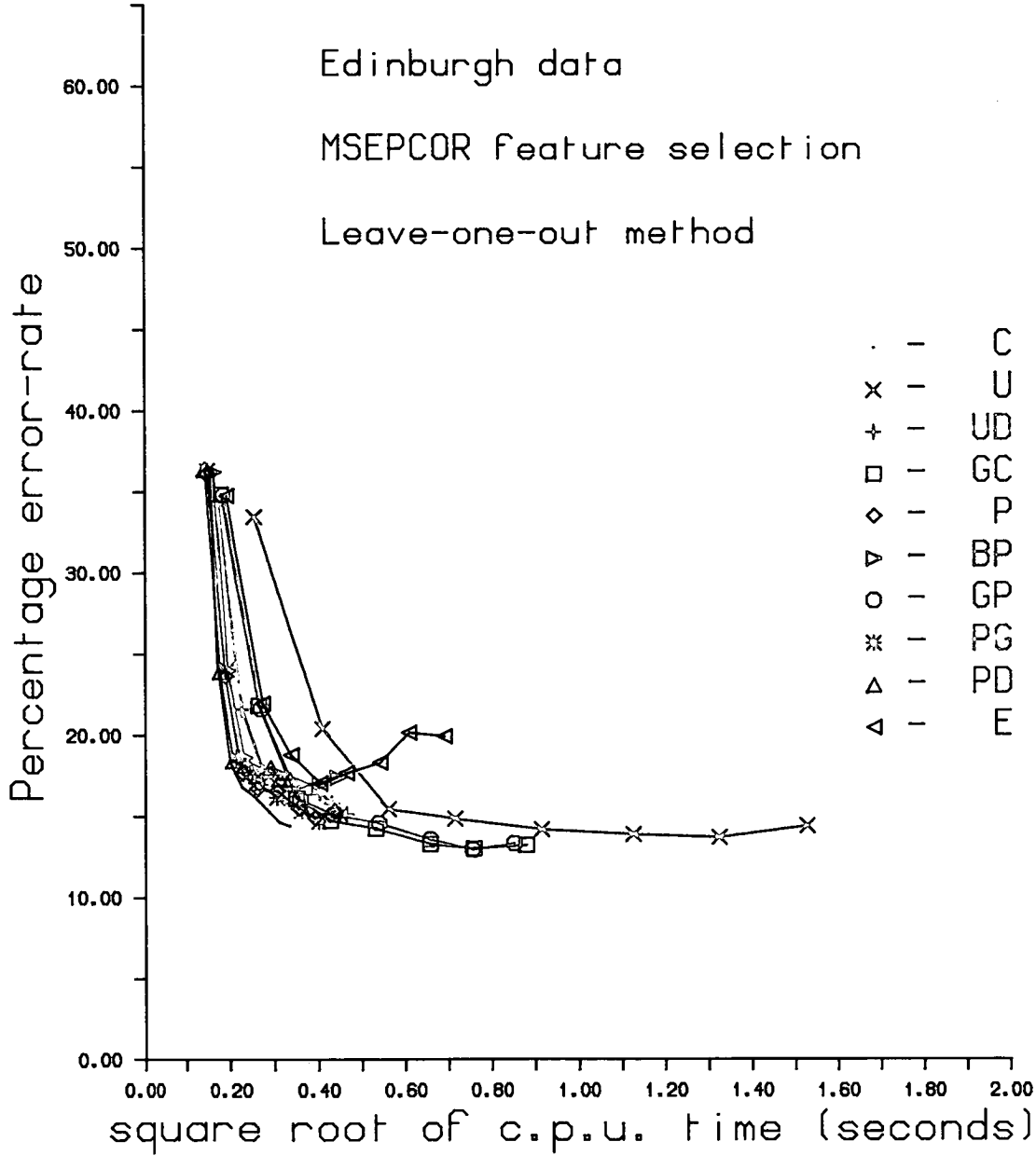

Edinburgh data
MSEPCOR feature selection
Leave-one-out method
C
U
UD
GC
P
BP
GP
PG
PD
E
Percentage error-rate
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)

Figure 5.3

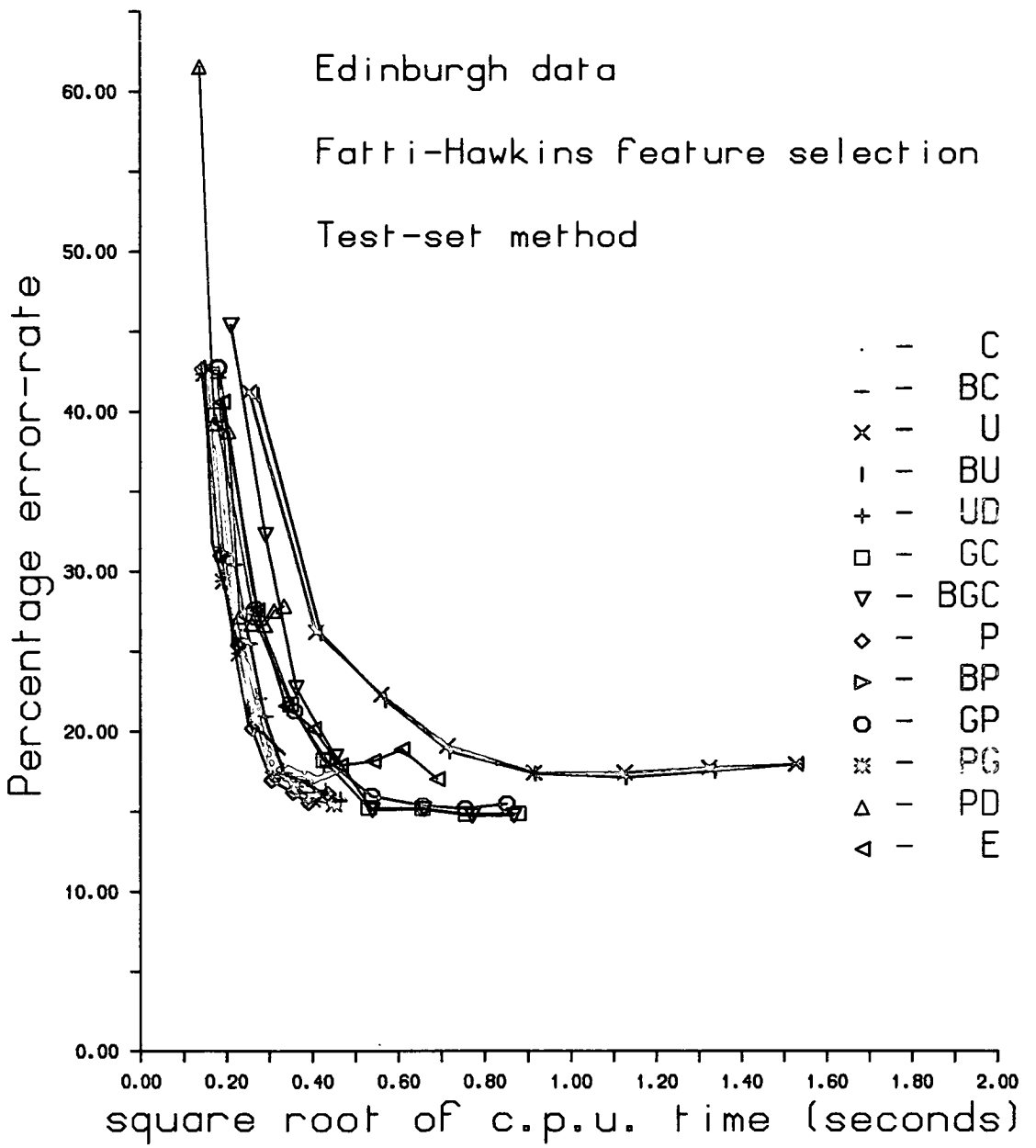

Edinburgh data
Fatti-Hawkins feature selection
Test-set method
Percentage error-rate
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)
. - C
- - BC
× - U
ı - BU
+ - UD
□ - GC
▽ - BGC
◇ - P
▷ - BP
○ - GP
⁜ - PG
△ - PD
◁ - E

Figure 5.4

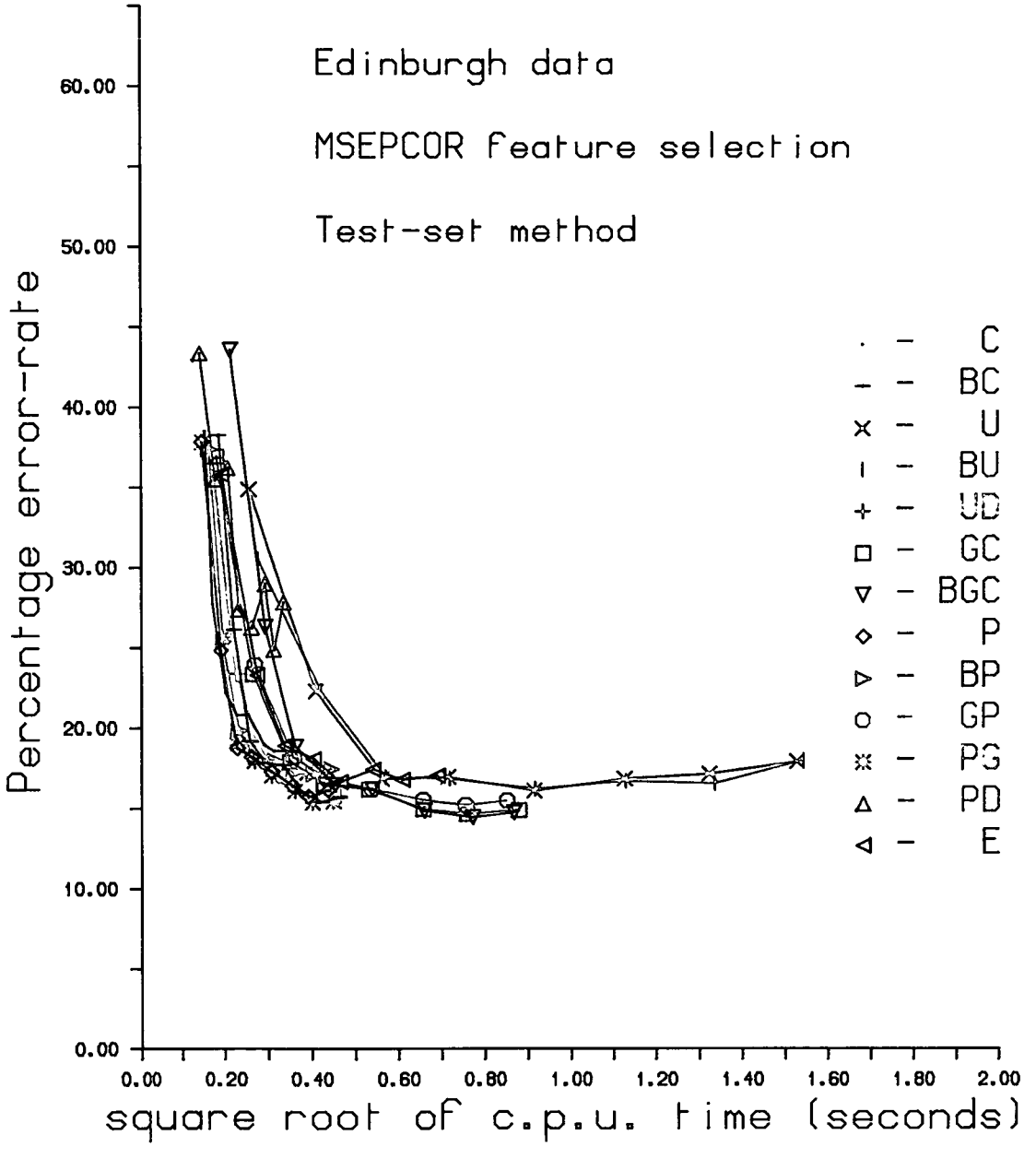

Edinburgh data
MSEPCOR feature selection
Test-set method
Percentage error-rate
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)
. - C
- - BC
x - U
| - BU
+ - UD
□ - GC
▽ - BGC
◇ - P
▷ - BP
○ - GP
※ - PS
△ - PD
◁ - E

Figure 5.5

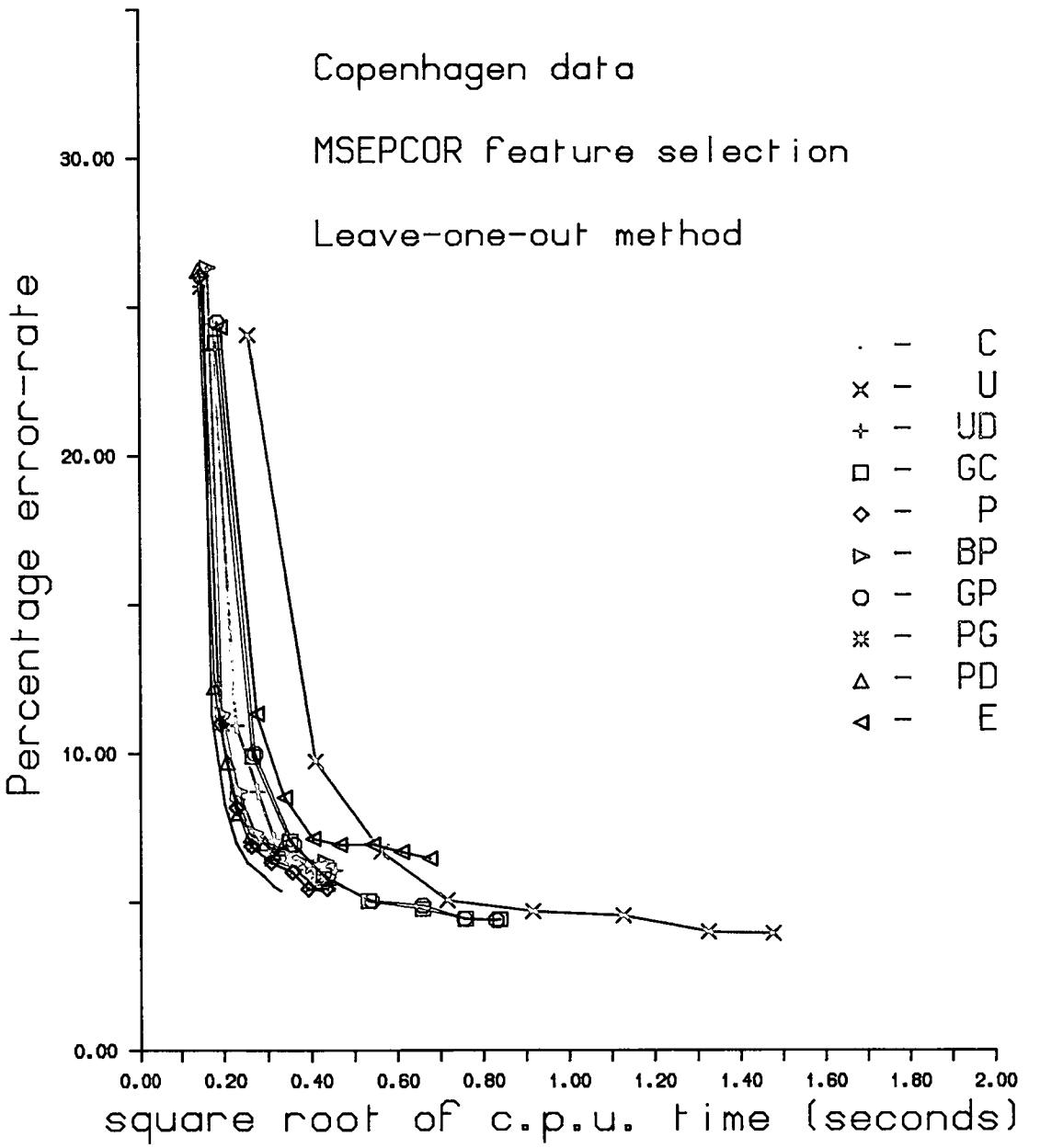

Copenhagen data
MSEPCOR feature selection
Leave-one-out method
Percentage error-rate
30.00
20.00
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)
. - C
× - U
+ - UD
□ - GC
◇ - P
▷ - BP
o - GP
※ - PG
△ - PD
◁ - E

Figure 5.6

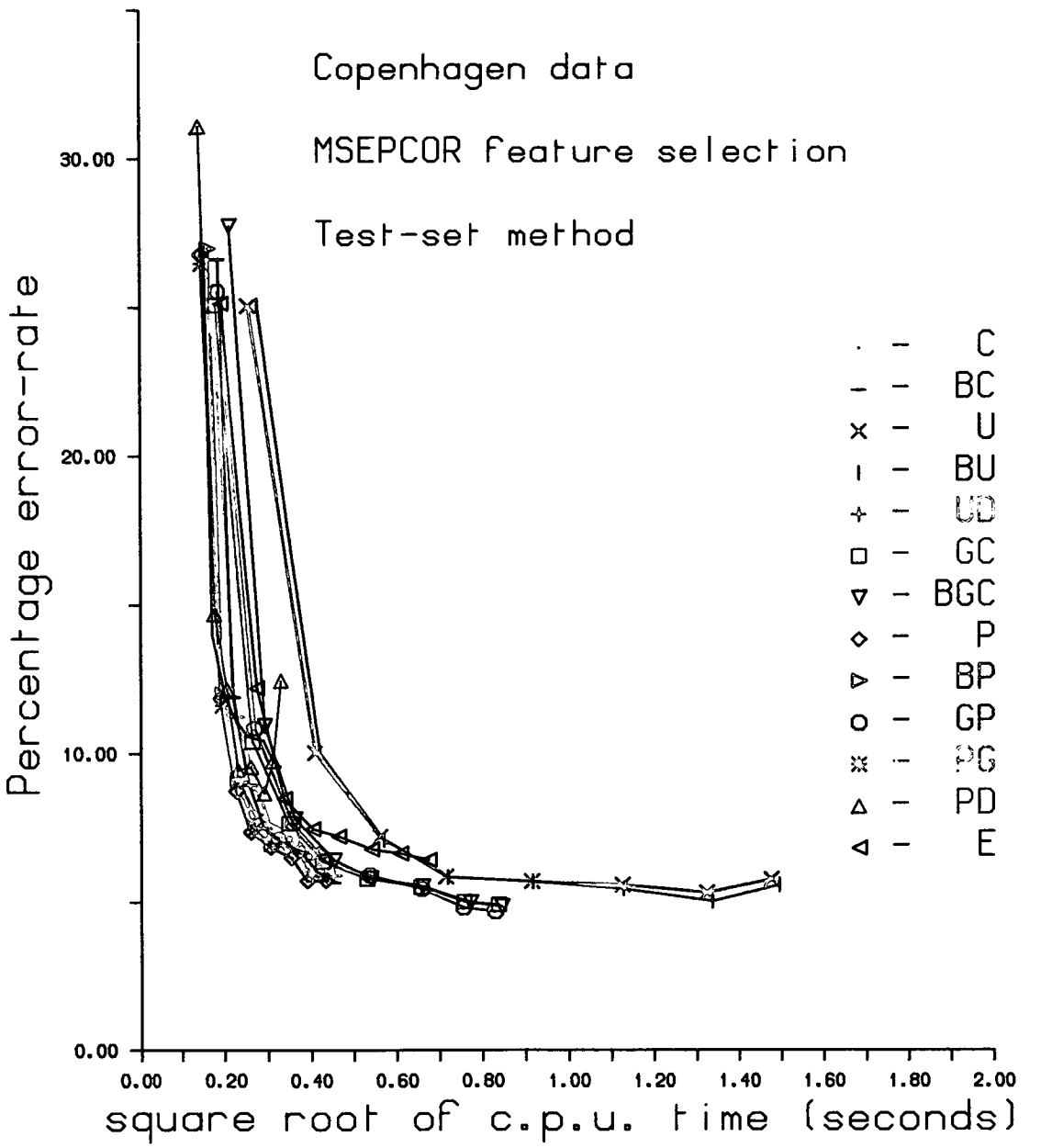

Copenhagen data
MSEPCOR feature selection
Test-set method
Percentage error-rate
30.00
20.00
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)
C
BC
U
BU
UD
GC
BGC
P
BP
GP
PG
PD
E

Figure 5.7

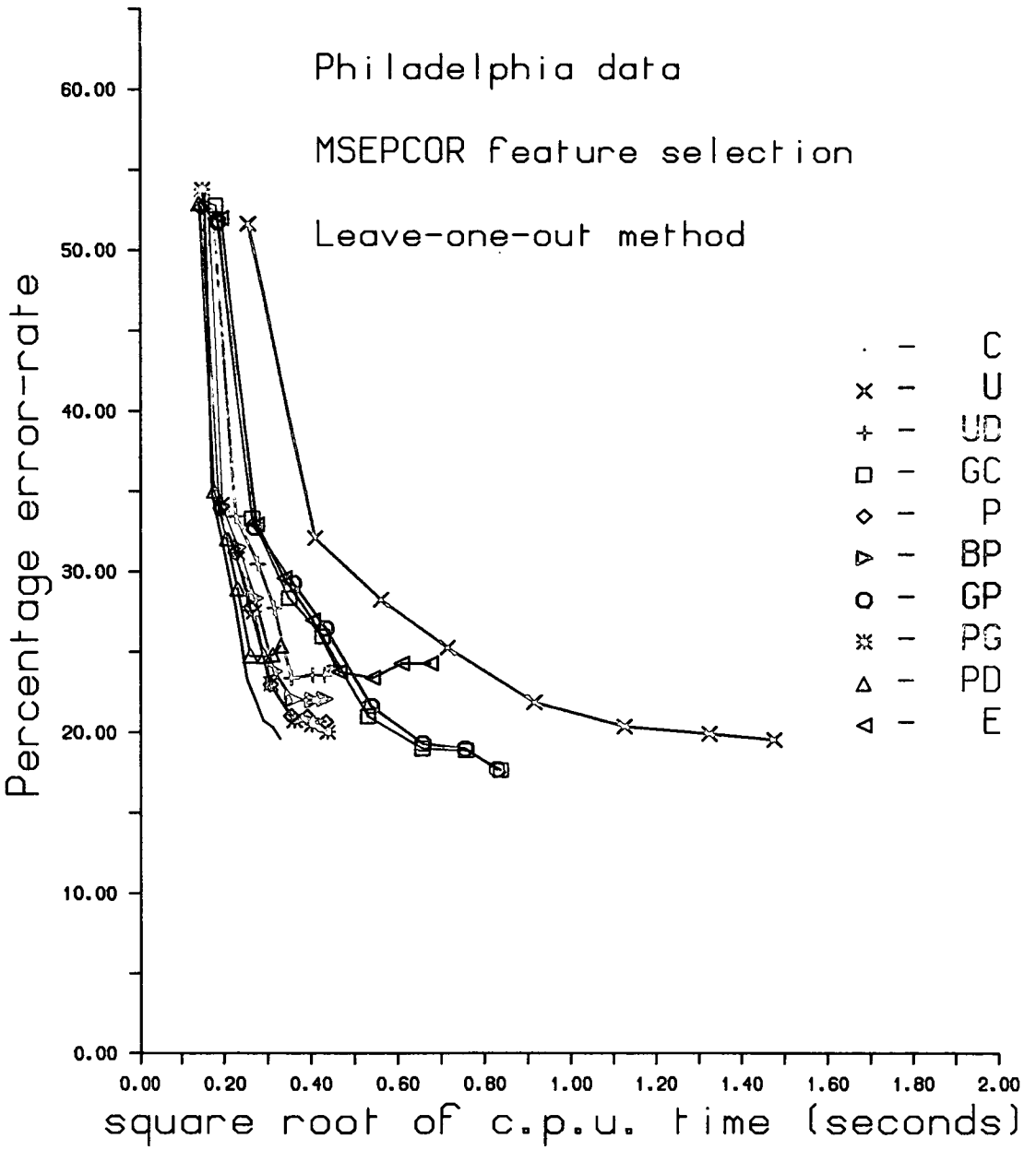

Philadelphia data
MSEPCOR feature selection
Leave-one-out method
C
U
UD
GC
P
BP
GP
PG
PD
E
Percentage error-rate
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)

Figure 5.8

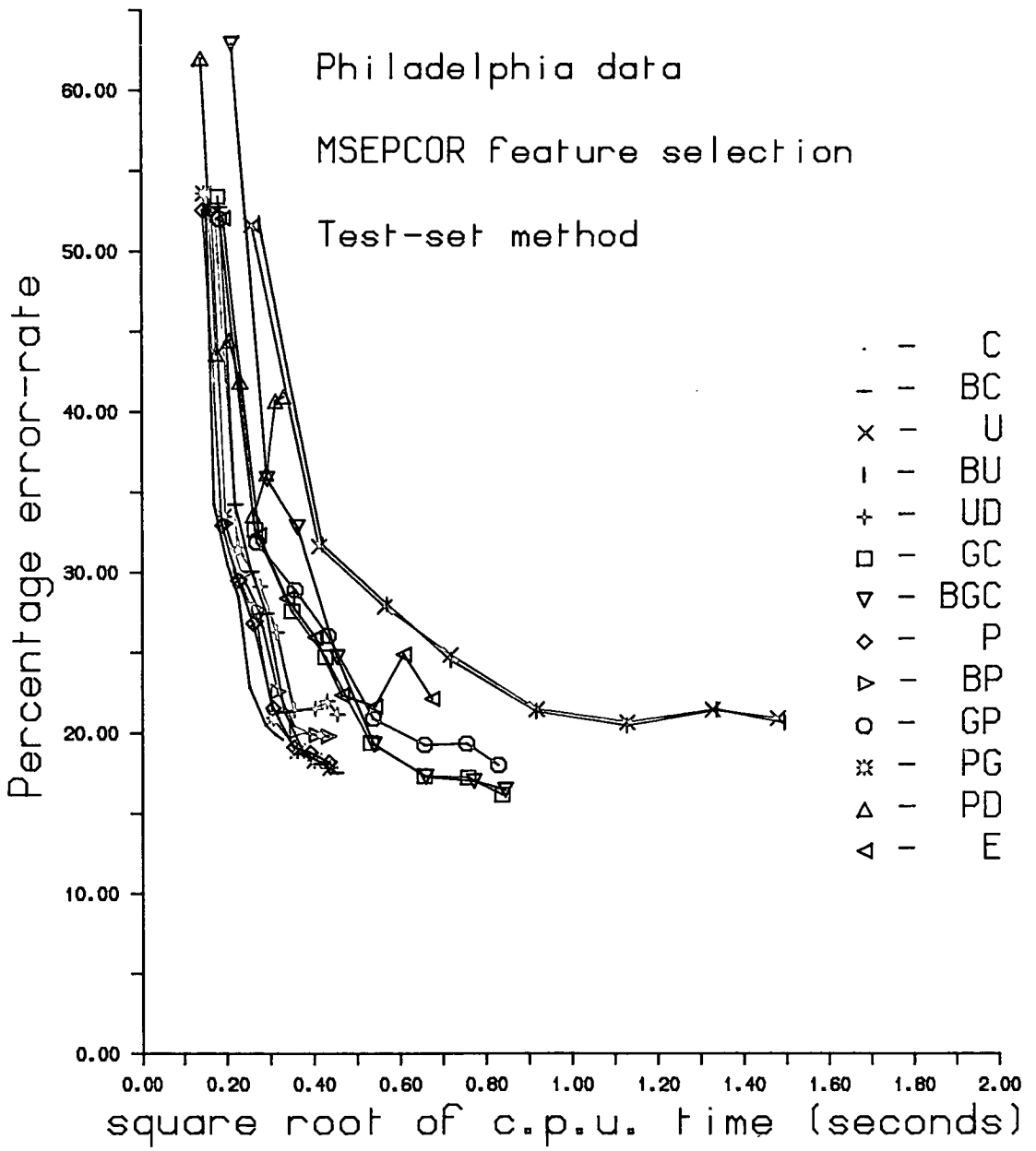

Philadelphia data
MSEPCOR feature selection
Test-set method
Percentage error-rate
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)
C
BC
U
BU
UD
GC
BGC
P
BP
GP
PG
PD
E

Figure 5.9

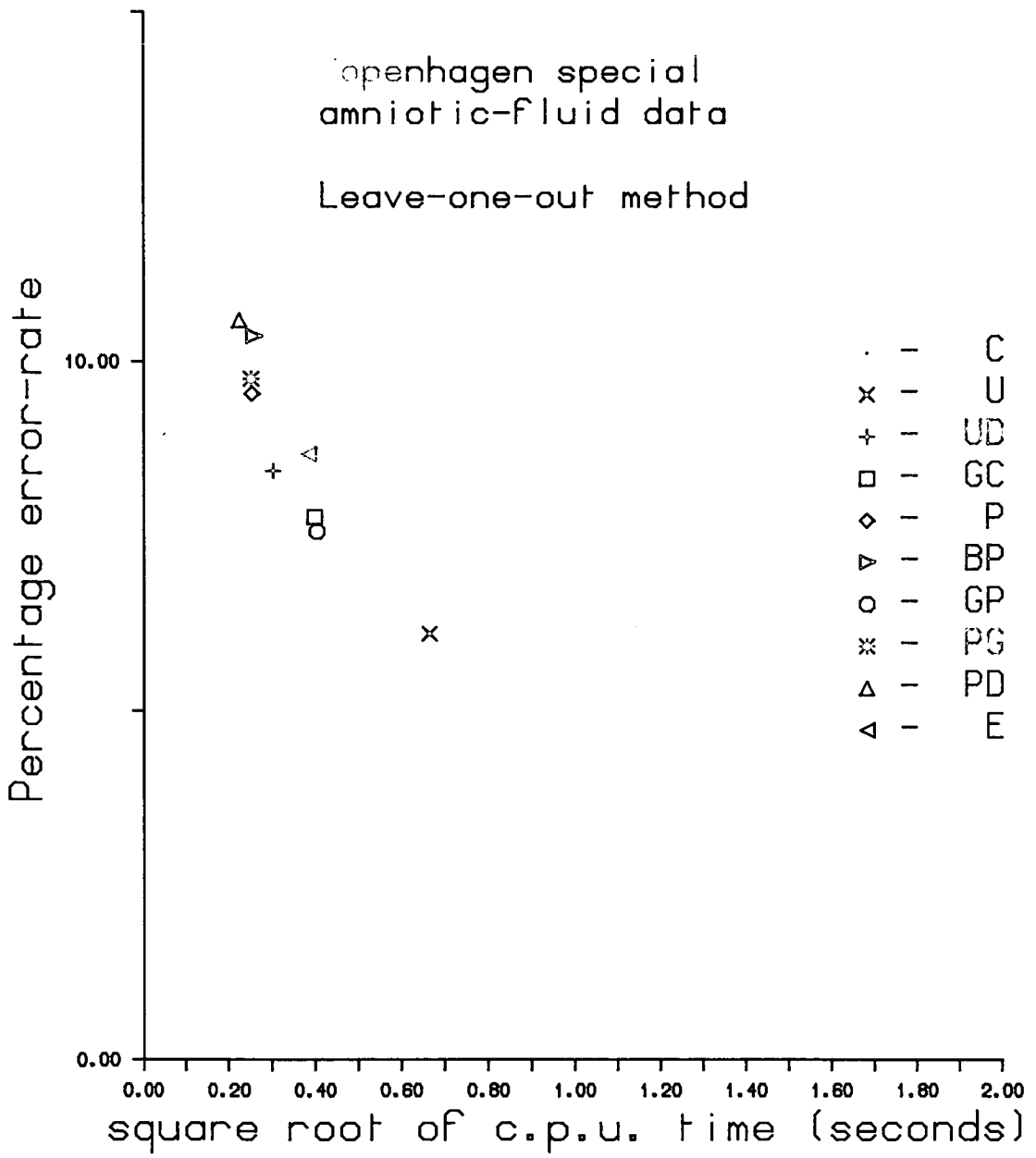

openhagen special
amniotic-fluid data
Leave-one-out method
Percentage error-rate
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)
C
U
UD
GC
P
BP
GP
PG
PD
E

Figure 5.10

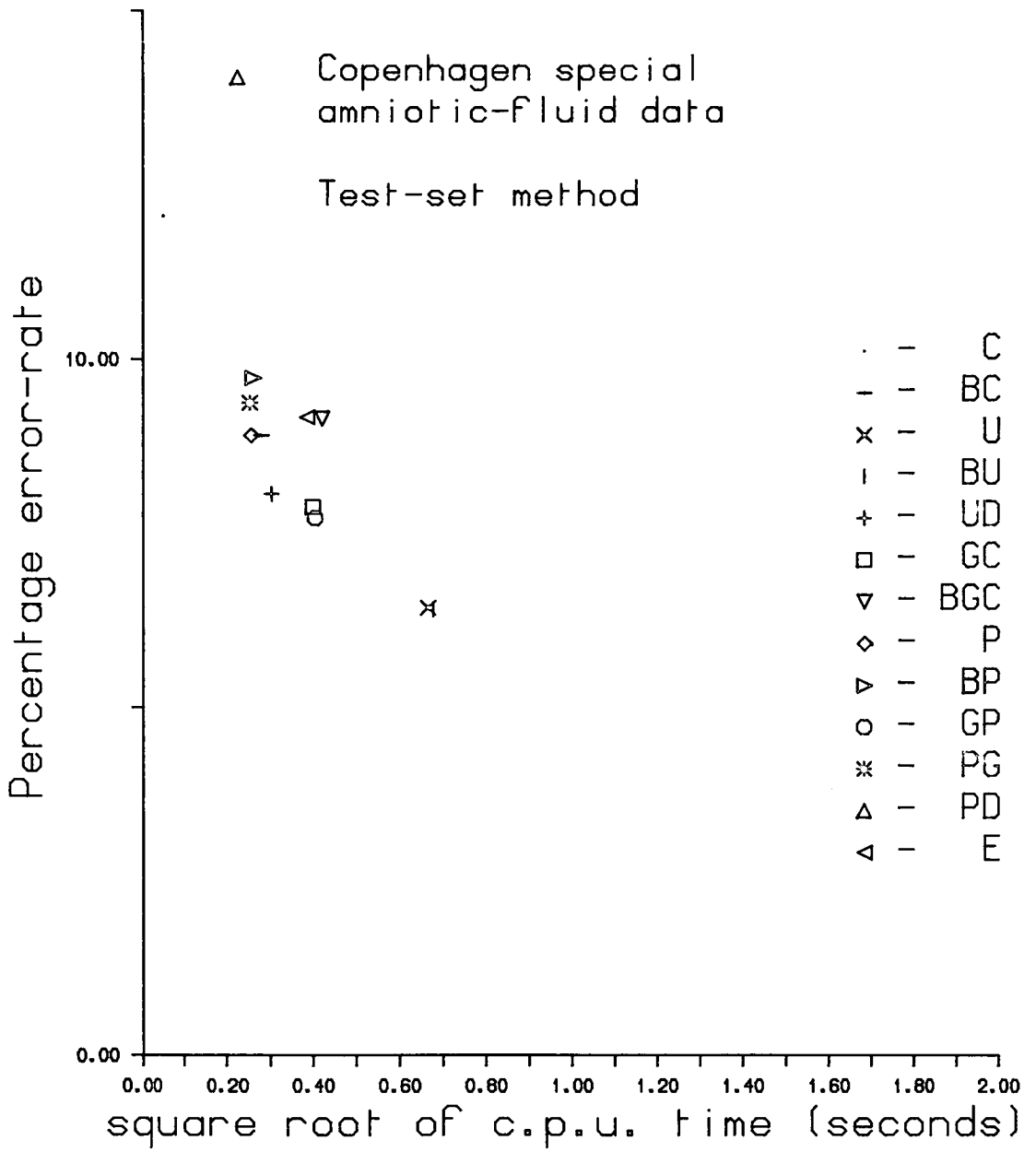

Copenhagen special
amniotic-fluid data
Test-set method
Percentage error-rate
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)
C
BC
U
BU
UD
GC
BGC
P
BP
GP
PG
PD
E

Figure 5.11

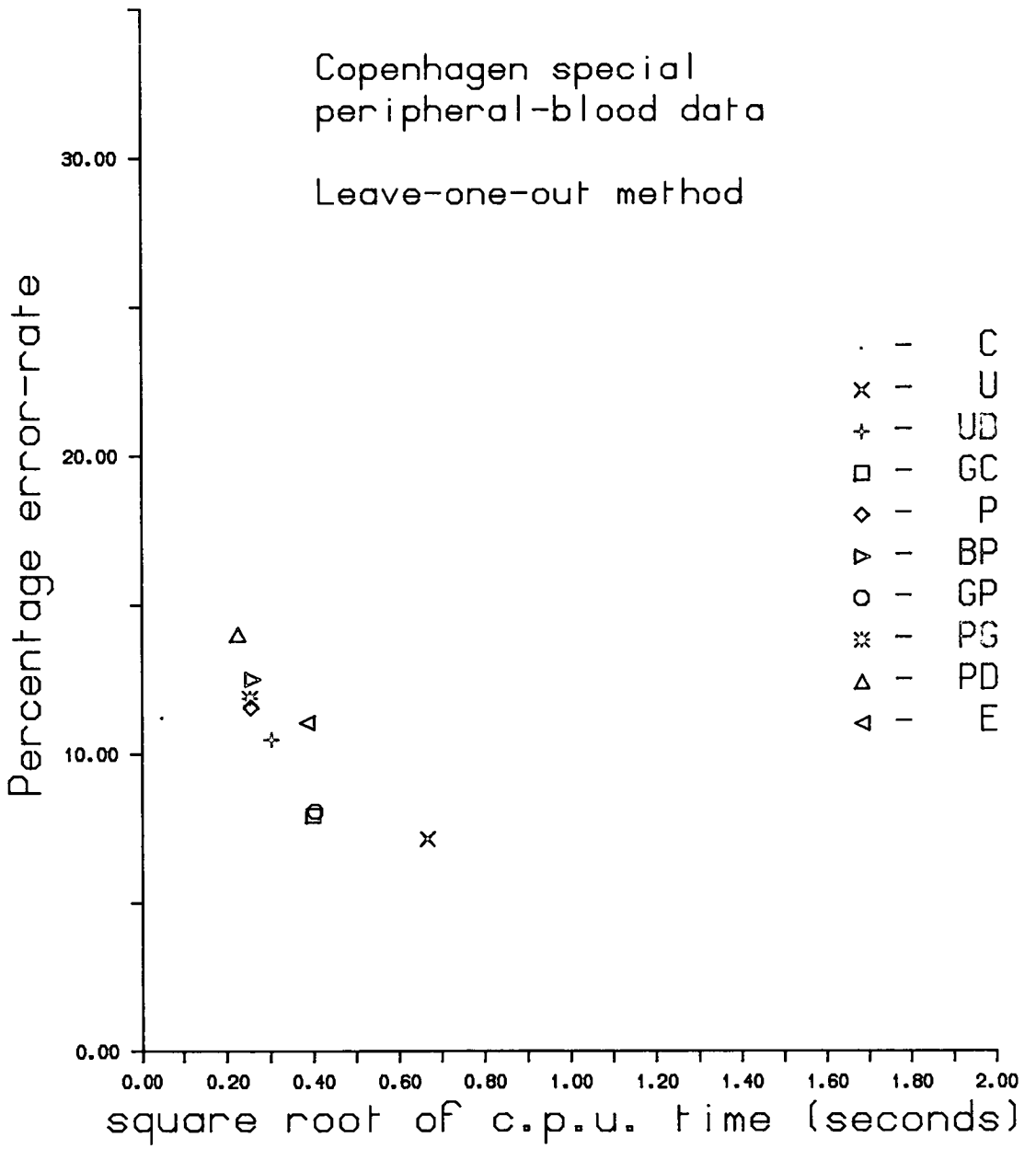
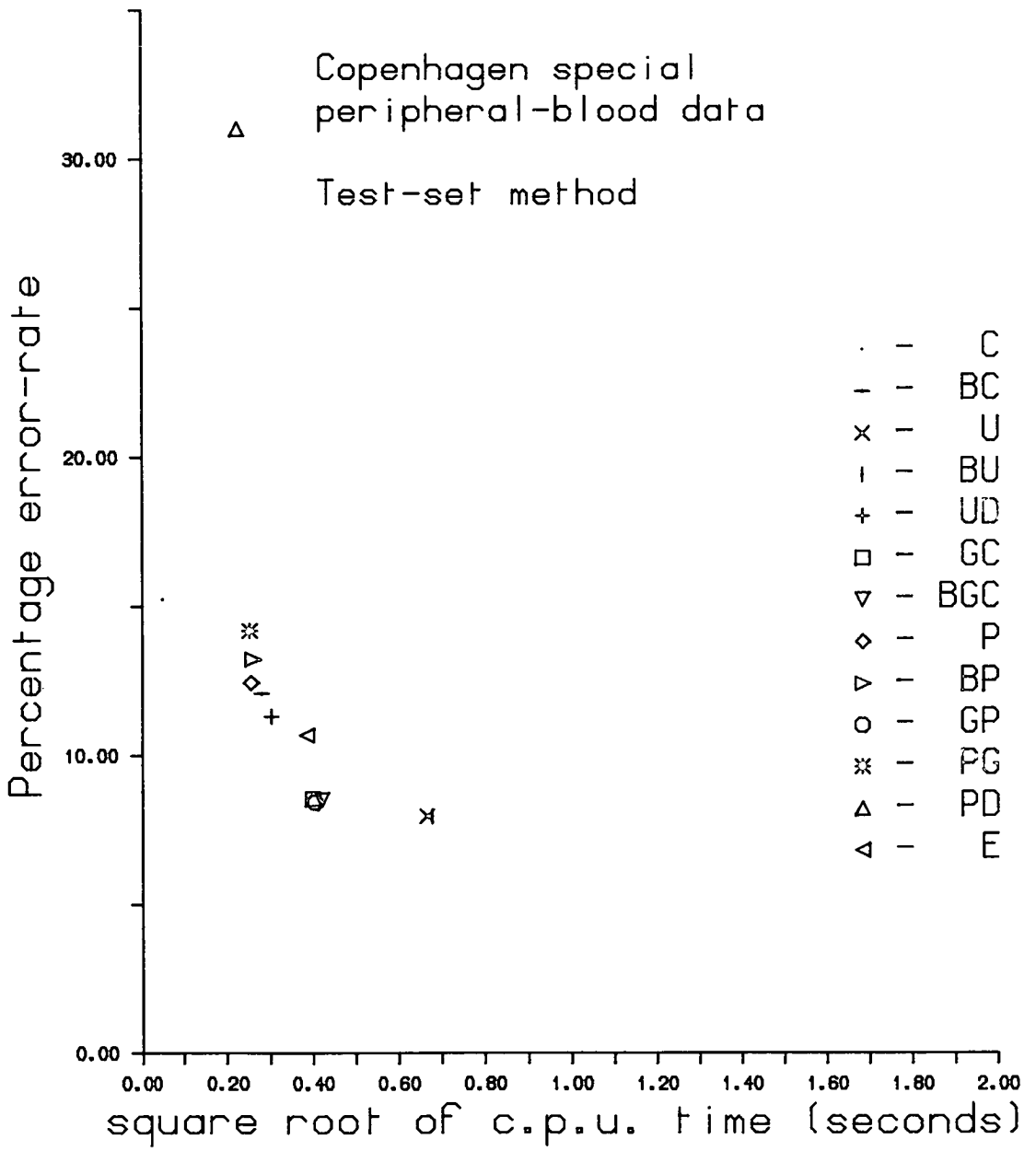

Copenhagen special
peripheral-blood data
Leave-one-out method
. – C
× – U
+ – UD
□ – GC
◇ – P
▷ – BP
○ – GP
※ – PG
△ – PD
◁ – E
Percentage error-rate
30.00
20.00
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)

Figure 5.12


Copenhagen special
peripheral-blood data
Test-set method
Percentage error-rate
30.00
20.00
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)
. - C
- - BC
× - U
| - BU
+ - UD
□ - GC
▽ - BGC
◇ - P
▷ - BP
○ - GP
※ - PG
△ - PD
◁ - E

discriminant scores were calculated as described in sub-sections 5.5.1 and 5.5.2 . The programs were written in Fortran 77 using double-precision arithmetic and executed on Edinburgh University's NAS computer. This computer is considerably more powerful than those commonly used for automated karyotyping but it shows the likely proportional savings in time. The square-root of average c.p.u. time was used because the number of calculations is approximately proportional to the square of the number of features except for procedures C, UD and PD. The results for the Estimative and Bayesian predictive procedures were in general so similar, except for procedures P and BP, for the leave-one-out method that they are not plotted on the figures but are instead given in Tables 5.4-5.7 .

The figures for the leave-one-out results show that procedures C, U, GC and GP are candidate procedures in the trade-off of estimated percentage error-rate against computational time for the Copenhagen data set. A candidate procedure is defined as one for which no other procedure which has the same or a quicker allocation time has a smaller estimated percentage error-rate. Procedures C, GC and GP are candidates for the Edinburgh and Philadelphia data sets and procedures C, U, UD, GC and GP are candidates for the special Copenhagen data sets. The figures also show that the estimated percentage error-rates for procedure C are considerably better than those reported earlier for the test set method for the Edinburgh and Copenhagen data sets (Granum, 1982 and Piper, 1987). Tables 5.4-5.7 show that there is, in general, very little difference between the estimated percentage error-rates for the Estimative and Bayesian procedures, excluding P and BP, for the sample sizes considered here.

The figures for the test-set method show that procedures C, BC, U, UD, GC, BGC, P, GP, PG and E are all candidates in the trade-off of estimated percentage error-rate against computational time for at least one of the five data sets.

Also of interest is the comparison between test-set and leave-one-out results. The leave-one-out percentage error-rates are in general lower and it may be conjectured that this is because of the more adequate size of the data sets for the estimation of parameters or derivation of predictive densities. It is also apparent that procedures C and PD give markedly lower percentage error-rates for the leave-one-out method. The reason for this is not clear but it is possible that the test set results for these procedures are more influenced

Table 5.4

Comparison of Bayesian and Estimative leave-one-out percentage error-rates for the Edinburgh data set.

(Percentage error-rate for MSEPCOR feature selection followed by percentage error-rate for Fatti-Hawkins feature selection.)

Number of features	Procedure			
	C	BC	GC	BGC
3	36.3,41.9	36.3,41.9	34.8,39.6	34.8,39.6
6	24.5,28.7	24.5,28.8	21.8,25.7	21.8,25.7
9	18.2,23.7	18.2,23.7	16.1,20,4	16.1,20.4
12	16.7,19.8	16.7,19.8	14.7,16.9	14.6,16.9
16	16.2,16.5	16.2,16.4	14.2,14.0	14.1,14.0
20	15.2,15.7	15.2,15.7	13.2,13.4	13.2,13.3
24	14.6,14.9	14.6,14.9	13.0,13.1	13.0,13.2
28	14.5	14.4	12.8	12.8

Number of features	Procedure	
	U	BU
3	33.3,40.7	33.4,40.6
6	20.4,24.0	20.4,23.8
9	15.4,19.6	15.4,19.6
12	14.8,16.8	15.0,16.6
16	14.2,14.4	14.2,14.4
20	13.8,14.5	13.4,14.3
24	13.7,14.1	13.4,13.5
28	14.4	13.7

Table 5.5

Comparison of Bayesian and Estimative leave-one-out percentage error-rates for the Copenhagen data set.

Number of features	Procedure			
	C	BC	GC	BGC
3	25.7	25.7	23.8	23.8
6	11.2	11.2	9.9	9.9
9	8.3	8.3	7.0	7.0
12	7.0	7.0	5.9	5.9
16	6.3	6.3	5.0	5.0
20	5.9	5.9	4.8	4.8
24	5.5	5.5	4.4	4.4
27	5.4	5.4	4.4	4.4

Number of features	Procedure	
	U	BU
3	24.1	24.0
6	9.7	9.7
9	6.7	6.7
12	5.1	5.1
16	4.7	4.7
20	4.5	4.5
24	4.0	4.0
27	3.9	4.0

Table 5.6

Comparison of Bayesian and Estimative leave-one-out percentage error-rates for the Philadelphia data set.

Number of features	Procedure			
	C	BC	GC	BGC
3	53.3	53.4	52.7	52.7
6	34.7	34.7	33.3	33.3
9	31.0	31.0	28.3	28.4
12	27.7	27.7	25.9	25.9
16	23.2	21.7	20.9	20.9
20	20.7	20.0	18.9	18.9
24	20.2	19.8	18.8	18.7
27	19.5	19.2	17.6	17.5

Number of features	Procedure	
	U	BU
3	51.6	51.5
6	32.0	32.0
9	28.2	28.1
12	25.2	25.3
16	21.8	21.7
20	20.3	20.0
24	19.8	19.8
27	19.4	19.2

Table 5.7

Comparison of Bayesian and Estimative leave-one-out percentage error-rates for the Copenhagen special data sets.

Data set	Procedure			
	C	BC	GC	BGC
Amniotic fluid	8.9	9.1	7.8	7.8
Peripheral blood	11.2	11.2	7.9	7.9

Data set	Procedure	
	U	BU
Amniotic fluid	6.1	6.1
Peripheral blood	7.1	6.7

by extreme observations than for the other procedures.

5.8 Discussion.

The results in this chapter show that some of the methods of combining class information on variability considered in this chapter give possible choices in the trade-off of estimated percentage error-rate versus allocation time for 46 chromosomes in a cell. It is noticeable that on some occasions these methods even give lower estimated percentage error-rates than the use of unrelated covariance matrices for each class. It is conjectured that this is due to the bias in the predicted densities being outweighed by the reduction in sampling variation. This advantage may be expected to disappear as the training set size increases unless any of the assumptions about the relationships between the covariance matrices are actually true.

A subset of the results given here is presented in Kirby et al (accepted for publication).

Chapter 6
Covariance selection models for the automated allocation of human chromosomes.

6.1 Introduction.

In the previous chapter six methods of combining class information on variability in multivariate Normal discrimination were proposed for the automated allocation of human chromosomes. These reduced the computational time required to allocate one chromosome for the numbers of features considered and also reduced the number of parameters, compared with the assumption of unrelated covariance matrices for each class. Another way the number of computations required to allocate one chromosome may be reduced and the number of parameters will be reduced for multivariate Normal data is to model the covariance structure for each class using covariance selection models. These approaches may also be used together by combining class information on variability and then modelling the resulting covariance structure. This method of modelling the covariance structure for each class or group may reduce the number of calculations required for the quadratic form in the discriminant score (5.22). The idea of using covariance selection models to reduce the number of parameters to be estimated has previously been considered for the analysis of repeated measurements by Kenward (1987) who considers a restricted class of covariance selection models.

In this chapter a brief outline of these covariance selection models is given. This is followed by a consideration of the number of calculations required to allocate one object and the number of parameters to be estimated. The method is first used to model the covariance structure for each class for the five data sets used in chapter 5. It is then used to model the covariance structure for the seven Denver groups for each data set.

6.2 Covariance selection models

Covariance selection models were introduced by Dempster (1972) as a method of giving a more parsimonious description of the covariance structure of a multivariate Normal population than the usual unbiased or maximum-likelihood estimates of a covariance matrix.

Dempster (1972) proposed that the number of parameters to be estimated for a positive definite covariance matrix, $\frac{1}{2}p(p + 1)$ where p is the number of features, could be reduced by setting certain elements of the inverse of the covariance matrix equal to zero. This can be done on the basis of a priori information or data-based tests. The interpretation of setting a particular element of the inverse, σ^{rs} , referred to as a concentration equal to zero is that features r and s are conditionally independent given fixed values of the remaining features. This is because setting σ^{rs} equal to zero is equivalent to setting ρ^{rs} , the corresponding element of the inverse of the correlation matrix, equal to zero and ρ^{rs} is a multiple of the partial correlation between features r and s. Dempster (1972) showed the existence of a unique estimate $\hat{\underline{\Sigma}}_{cs}$ of $\underline{\Sigma}$ for any set of index pairs (r,s) ($1 \leqq r < s \leqq p$) for which $\sigma^{rs} = 0$. He further showed that this estimate is maximum likelihood.

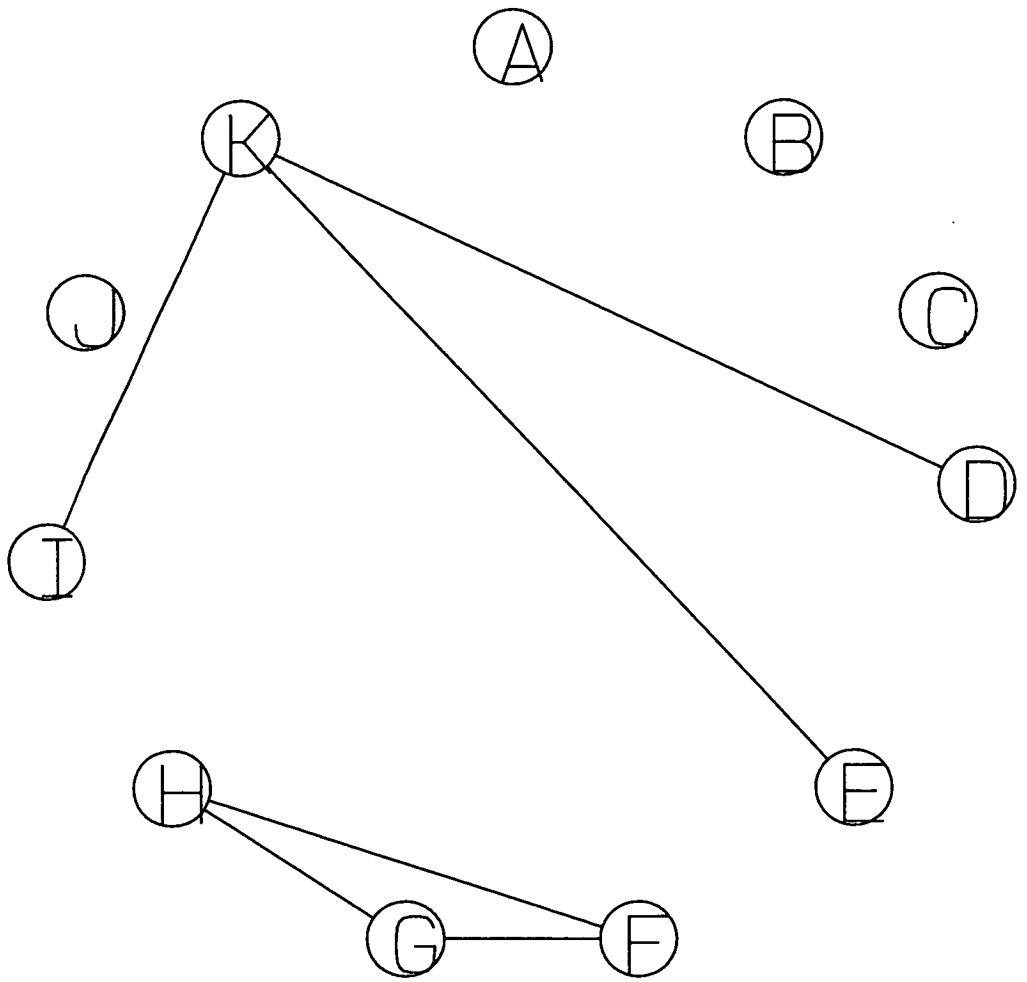
In the absence of a priori information, the elements σ^{rs} to be set equal to zero can be determined by the stepwise selection procedures analogous to those used for the selection of regressors in multiple regression. The change in twice the log likelihood for the addition or deletion of a concentration parameter is approximately a χ^2 variable on one degree of freedom under the null hypothesis that the concentration is equal to zero. Improved likelihood ratio statistics have been derived by Porteous (1985).

Two algorithms to obtain the estimates $\hat{\underline{\Sigma}}_{cs}$, given a set of concentrations to be set equal to zero, have been derived by Speed and Kiiveri (1986). A special case of the first algorithm has been programmed by Wermuth and Scheidt (1977) whilst the second algorithm is analogous to iterative proportional scaling for contingency tables.

A fitted model may be conveniently and uniquely represented by an interaction graph. This is an undirected graph whose vertices correspond to features and whose edges represent conditional dependencies. Thus no edge is drawn between two vertices if the corresponding features are conditionally independent. An example is given in Figure 6.1 which shows the interaction graph corresponding to the model fitted for chromosome class 1 in Table 6.7 obtained by setting concentrations equal to zero if the significance level exceeds 0.0001 in the likelihood ratio test of zero concentration versus the saturated model. The features listed as numbers 1 to 11 in chapter 3 have been coded as letters A to K in this graph.

Figure 6.1

Interaction graph for covariance selection model for chromosome class 1 in Table 6.7 obtained by setting concentrations equal to zero if significance level exceeds 0.0001 in test of zero concentration versus saturated model



6.3 Number of calculations required to allocate one new object and the number of parameters to be estimated.

6.3.1 An unrelated covariance matrix for each class.

Defining v_i as the number of concentrations to be set equal to zero for class i then we must have $v_i \leqq \frac{1}{2}p(p-1)$. The number of multiplications required to allocate one new object using the discriminant scores defined by (5.22), with $\hat{\underline{\Sigma}}_i$ representing the estimator of the covariance selection model for class i , is

$$\Sigma_i 2\{\tfrac{1}{2}p(p+1) - v_i\} \quad . \tag{6.1}$$

The number of additions and subtractions is

$$\Sigma_i\{\tfrac{1}{2}p(p+1) - v_i\} + cp + c \tag{6.2}$$

for c classes. These formulae assume that only the the non-zero elements of the lower or upper triangle of each estimated inverse covariance matrix, with off-diagonal elements multiplied by 2 , are stored and used in the computation. For the number of multiplications to be less than for the calculation of discriminant scores for procedure U as defined in chapter 5 we require that

$$\Sigma_i 2\{\tfrac{1}{2}p(p+1) - v_i\} < \tfrac{1}{2}cp(p+3) \tag{6.3}$$

and for the number of additions and subtractions to be less we require

$$\Sigma_i\{\tfrac{1}{2}p(p+1) - v_i\} + cp + c < \tfrac{1}{2}cp(p+3) + cp + c \quad . \tag{6.4}$$

The reduction in the number of multiplications obtained by using the Cholesky decomposition of the inverse of the estimated covariance matrix for procedure U may not in general be used here. This means that the number of multiplications may be greater than for procedure U.

The total number of parameters to be estimated for c classes is $cp + \Sigma_i\{\frac{1}{2}p(p + 1) - v_i\}$.

6.3.2 A common covariance matrix for the classes in each of g groups.

The numbers of multiplications and additions and subtractions to allocate one object using the discriminant scores defined by (5.22), with $\hat{\underline{\Sigma}}_i$ representing the estimator of the covariance selection model for the group containing the ith class, become

$$\Sigma_k 2\{\tfrac{1}{2}p(p + 1) - v_k\} + cp \qquad (6.5)$$

and

$$\Sigma_k\{\tfrac{1}{2}p(p + 1) - v_k\} + cp + 2c \qquad (6.6)$$

respectively. Here, v_k is the number of concentrations to be set equal to zero from group k. These formulae again assume that only the non-zero elements of the lower or upper triangle of each estimated estimated inverse, with off-diagonal elements multiplied by 2, are stored and used in the computation. They also assume that the quadratic form in the discriminant score (5.22) is expanded.

The total number of parameters to be estimated for c classes in g groups is $cp + \Sigma_k\{\frac{1}{2}p(p + 1) - v_k\}$.

6.4 Application to the five data sets used in chapter 5.

6.4.1 Features used.

For the Edinburgh, Copenhagen and Philadelphia data sets the first 24 features given by the MSEPCOR feature selection procedure shown in Table 5.3 were used. Only 24 features were used and not the full 28 or 27 containing no exact linear dependence because this was the maximum allowed by the program MIM used to fit the models and described further below. As seen in

chapter 5 the use of 24 features gives percentage error-rates close to those obtained with all 28 or 27 available features containing no exact linear dependence. The 11 features used in the WDD classifier described in chapter 3 were used for the special Copenhagen data sets. The normalisation of features for between-cell variation was that currently used and described in chapter 3.

6.4.2 Selection of concentrations to be set equal to zero.

Initially the algorithm given by Wermuth and Scheidt (1977) for setting one concentration to zero was used iteratively to set a number of concentrations to zero in a backwards elimination procedure. However it was found on a number of occasions that this approach led to divergence of the parameter estimates before the convergence criterion was met. Consequently, to make use of existing software the interactive computer program MIM (Edwards, 1987) was used to fit the covariance selection models for each class or group. This program fits the broader class of hierarchical interaction models using an iterative proportional scaling algorithm (Frydenberg and Edwards, 1989). Significance tests are based on the asymptotic likelihood ratio test or deviance. Because of the length of time taken to fit the models using this program concentrations were set to zero whenever the null hypothesis of zero concentration relative to the saturated model gave a significance level greater than α . This is a cruder procedure than the usual stepwise backwards elimination procedure but the results still illustrate the possible value of covariance selection models.

6.4.3 Fitted models.

Models were fitted to the first part of each data set, using the same random split by cell as described in chapter 3, so that the percentage error-rates could be estimated using the second part of each data set.

6.4.4 Estimated percentage error-rates.

Estimated percentage error-rates were obtained by weighting the estimated percentage error-rates for each class by the prior probabilities for each class. For the Edinburgh data set, the prior probabilities were those for all cells from males whilst for the remaining data sets the prior probabilities were those for equal numbers of male and female cells. As in chapters 4 and 5 no re-arrangement of allocations to achieve a normal karyotype was done.

6.5 Results.

6.5.1 An unrelated covariance matrix for each class.

Two sets of covariance selection models were fitted to see if a set of models could be found which gave a candidate procedure for the trade-off of estimated percentage error-rate and the c.p.u. time required to allocate 46 chromosomes in a cell. The discriminant scores were calculated as described in sub-sections 6.3.1 and 6.3.2 . As in chapter 5, the c.p.u. times were the average of ten times for the same operands using programs written in Fortran 77 with double-precision arithmetic executed on Edinburgh University's NAS computer. The sets of covariance selection models were obtained by using significance levels of 0.01 and 0.0001 to decide which concentrations to set equal to zero. Such extreme significance levels were chosen because of the good performance of procedure UD in chapter 5. Summaries of model fits are given in Tables 6.1, 6.3, 6.5, 6.7 and 6.9. Estimated percentage error-rates for the two sets of covariance selection models and procedures UD and U are given in Tables 6.2, 6.4, 6.6, 6.8 and 6.10. Procedures UD and U may be viewed as two extreme covariance selection models corresponding to setting concentrations equal to zero if a significance level of 0 or 1 respectively is exceeded for the likelihood ratio test of zero concentration versus the saturated model. The average number of concentrations set equal to zero out of the maximum possible 276 and 55 for the models in Tables 6.1, 6.3, 6.5, 6.7 and 6.9 , which were obtained from using a significance level of 0.01 above which concentrations are set equal to zero, is 222, 214, 219, 29 and 31 respectively to the nearest integer. The average number of concentrations set equal to zero for the second set of models in these tables, which were obtained from using a significance level of 0.0001 above which concentrations are set equal to zero, is 256, 247, 251, 42 and 39 respectively to the nearest integer.

6.5.2 A common covariance matrix for the classes in each of g groups.

Because of the success of procedure GC in chapter 5 a covariance selection model was fitted to the common covariance model for the classes in each Denver group. The same two significance levels to decide which concentrations to set equal to zero as used for modelling the covariance structure for each class were used here. Summaries of model fits are given in Tables 6.11, 6.13, 6.15, 6.17 and 6.19. Estimated percentage error-rates are given in Tables 6.12, 6.14, 6.16, 6.18 and 6.20. The estimated percentage error-rate for procedure GC

Table 6.1

Summaries of covariance selection model fits for each chromosome class.
(Concentrations set equal to zero if significance level exceeds 0.01 (set 1) or 0.0001 (set 2) in likelihood ratio test of zero concentration versus saturated model.)

Edinburgh data set.

	Set 1		Set 2	
Class	deviance	df	deviance	df
1	606.9	236	900.2	261
2	621.4	250	795.1	269
3	543.5	225	959.1	258
4	525.3	240	751.2	261
5	551.2	222	940.6	259
6	573.1	213	979.5	249
7	640.3	218	1057.8	261
8	683.0	215	1111.2	249
9	878.6	230	1258.1	257
10	733.0	220	1076.8	257
11	744.9	219	1390.0	255
12	735.5	203	1298.5	254
13	791.7	218	1279.5	249
14	733.3	232	1007.8	253
15	800.7	216	1311.0	253
16	682.4	211	1400.4	251
17	680.3	223	1242.6	252
18	649.9	211	1164.3	250
19	865.6	232	1433.8	255
20	682.1	224	1317.6	259
21	667.0	213	1692.1	254
22	691.2	190	1546.2	243
23	672.3	242	808.8	267
24	853.3	232	1092.2	260

Table 6.2

Covariance selection models for each chromosome class.

Edinburgh data set.

Size of test	Estimated percentage error-rate
1	17.1
0.01	15.6
0.0001	15.6
0	16.5

Table 6.3

Summaries of covariance selection model fits for each chromosome class.
(Concentrations set equal to zero if significance level exceeds 0.01 (set 1) or 0.0001 (set 2) in likelihood ratio test of zero concentration versus saturated model.)

Copenhagen data set.

	Set 1		Set 2	
Class	deviance	df	deviance	df
1	642.5	219	827.1	247
2	566.5	231	846.1	256
3	627.6	233	1001.5	257
4	535.4	214	917.1	254
5	581.8	218	1042.7	254
6	652.0	209	1093.9	247
7	614.0	220	924.0	249
8	712.9	215	1281.6	257
9	625.5	221	1042.8	249
10	634.8	219	919.3	248
11	737.7	218	1432.8	248
12	593.2	207	843.0	232
13	823.5	200	1477.8	240
14	763.0	218	1243.2	248
15	683.7	205	1128.6	240
16	558.6	213	1028.5	248
17	1041.7	219	1402.2	246
18	686.2	196	1189.9	234
19	848.8	211	1586.0	242
20	730.3	216	1208.1	239
21	970.0	199	1234.4	244
22	985.3	217	1413.9	243
23	649.5	218	996.6	249
24	591.4	208	874.3	255

Table 6.4

Covariance selection models for each chromosome class.

Copenhagen data set.

Size of test	Estimated percentage error-rate
1	5.3
0.01	5.0
0.0001	5.0
0	6.2

Table 6.5

Summaries of covariance selection model fits for each chromosome class.
(Concentrations set equal to zero if significance level exceeds 0.01 (set 1) or 0.0001 (set 2) in likelihood ratio test of zero concentration versus saturated model.)

Philadelphia data set.

	Set 1		Set 2	
Class	deviance	df	deviance	df
1	559.5	225	722.6	248
2	685.0	221	967.6	253
3	520.2	234	796.1	256
4	664.7	231	990.5	256
5	674.3	229	1043.7	254
6	803.9	221	1185.9	255
7	640.4	225	976.9	252
8	698.0	210	1225.6	248
9	697.7	211	1231.0	253
10	677.6	223	1055.5	255
11	680.4	202	1029.6	241
12	998.2	228	1342.6	252
13	809.9	211	1336.2	248
14	638.1	206	1426.1	249
15	720.9	221	1089.0	247
16	709.1	203	1193.5	246
17	832.6	224	1282.1	251
18	800.9	214	1451.0	251
19	837.7	212	1209.9	247
20	764.8	223	1552.9	255
21	812.0	210	1459.6	244
22	806.1	229	1240.4	252
23	812.7	228	971.7	250
24	788.8	220	999.1	260

Table 6.6

Covariance selection models for each chromosome class.

Philadelphia data set.

Size of test	Estimated percentage error-rate
1	21.4
0.01	20.3
0.0001	19.7
0	21.9

Table 6.7

Summaries of covariance selection model fits for each chromosome class.
(Concentrations set equal to zero if significance level exceeds 0.01 (set 1) or 0.0001 (set 2) in likelihood ratio test of zero concentration versus saturated model.)

Copenhagen special amniotic-fluid data set.

	Set 1		Set 2	
Class	deviance	df	deviance	df
1	178.5	41	292.1	49
2	117.3	34	221.9	44
3	116.3	33	404.1	45
4	128.8	28	392.4	42
5	56.2	25	425.6	40
6	72.7	26	365.8	43
7	76.3	27	178.8	39
8	47.5	21	245.3	42
9	225.5	31	473.2	42
10	65.1	25	378.1	36
11	206.3	34	502.1	46
12	91.4	25	373.2	37
13	141.1	30	306.2	39
14	175.9	36	285.0	40
15	140.7	30	332.3	38
16	65.5	30	650.6	47
17	46.5	20	324.9	40
18	64.0	20	235.8	30
19	89.0	33	349.7	43
20	121.3	35	313.7	44
21	89.3	31	374.3	47
22	81.8	27	438.0	39
23	62.2	24	291.3	39
24	126.9	41	185.7	50

Table 6.8

Covariance selection models for each chromosome class.

Copenhagen special amniotic-fluid data set.

Size of test	Estimated percentage error-rate
1	6.4
0.01	6.4
0.0001	7.5
0	8.1

Table 6.9

Summaries of covariance selection model fits for each chromosome class.
(Concentrations set equal to zero if significance level exceeds 0.01 (set 1) or 0.0001 (set 2) in likelihood ratio test of zero concentration versus saturated model.)

Copenhagen special peripheral-blood data set.

	Set 1		Set 2	
Class	deviance	df	deviance	df
1	165.0	40	204.7	47
2	205.2	32	318.6	40
3	237.6	34	419.5	44
4	112.8	31	285.9	37
5	84.4	26	164.9	33
6	185.0	30	375.3	36
7	93.5	28	162.9	35
8	58.4	25	167.6	34
9	107.9	30	252.5	36
10	58.3	31	222.1	38
11	144.6	27	430.5	34
12	161.3	30	253.4	36
13	154.6	29	297.1	36
14	159.9	31	198.4	36
15	188.5	31	387.6	38
16	217.1	34	305.1	40
17	204.0	30	302.9	35
18	98.8	25	341.8	36
19	107.5	30	262.0	43
20	146.3	36	258.9	44
21	70.4	24	259.4	35
22	185.7	28	479.6	38
23	118.1	33	200.9	44
24	194.1	43	307.0	49

Table 6.10

Covariance selection models for each chromosome class.

Copenhagen special peripheral-blood data set.

Size of test	Estimated percentage error-rate
1	7.9
0.01	8.1
0.0001	8.6
0	11.3

Table 6.11

Summaries of covariance selection model fits for each Denver group.
(Concentrations set equal to zero if significance level exceeds 0.01 (set 1) or 0.0001 (set 2) in likelihood ratio test of zero concentration versus saturated model.)

Edinburgh data set.

	Set 1		Set 2	
Group	deviance	df	deviance	df
1	524.9	212	940.2	244
2	625.1	222	886.1	248
3	610.2	158	1298.7	202
4	667.0	194	1122.8	227
5	600.0	175	1315.8	218
6	771.7	203	1338.1	245
7	773.7	196	1256.6	232

Table 6.12

Covariance selection models for each Denver group.

Edinburgh data set.

Size of test	Estimated percentage error-rate
1	14.6
0.01	15.0
0.0001	14.6

Table 6.13

Summaries of covariance selection model fits for each Denver group.
(Concentrations set equal to zero if significance level exceeds 0.01 (set 1) or 0.0001 (set 2) in likelihood ratio test of zero concentration versus saturated model.)

Copenhagen data set.

	Set 1		Set 2	
Group	deviance	df	deviance	df
1	700.6	215	973.3	240
2	649.6	201	995.5	241
3	624.6	157	1213.4	189
4	797.4	188	1429.5	221
5	663.8	181	1118.8	210
6	742.2	187	1535.7	225
7	750.8	192	1570.5	238

Table 6.14

Covariance selection models for each Denver group.

Copenhagen data set.

Size of test	Estimated percentage error-rate
1	5.0
0.01	4.7
0.0001	5.1

Table 6.15

Summaries of covariance selection model fits for each Denver group.
(Concentrations set equal to zero if significance level exceeds 0.01 (set 1) or 0.0001 (set 2) in likelihood ratio test of zero concentration versus saturated model.)

Philadelphia data set.

	Set 1		Set 2	
Group	deviance	df	deviance	df
1	688.9	203	1042.6	237
2	829.1	209	1172.2	241
3	1110.8	152	1753.6	189
4	816.9	186	1371.3	220
5	820.0	183	1618.1	220
6	730.4	191	1355.2	234
7	1026.2	202	1304.8	228

Table 6.16

Covariance selection models for each Denver group.

Philadelphia data set.

Size of test	Estimated percentage error-rate
1	17.2
0.01	18.0
0.0001	17.9

Table 6.17

Summaries of covariance selection model fits for each Denver group.
(Concentrations set equal to zero if significance level exceeds 0.01 (set 1) or 0.0001 (set 2) in likelihood ratio test of zero concentration versus saturated model.)

Copenhagen special amniotic-fluid data set.

	Set 1		Set 2	
Group	deviance	df	deviance	df
1	56.8	27	146.9	38
2	77.0	22	131.6	27
3	143.4	16	247.5	24
4	155.3	26	264.7	34
5	43.1	18	213.7	29
6	102.3	27	246.0	36
7	163.1	23	189.6	29

Table 6.18

Covariance selection models for each Denver group.

Copenhagen special amniotic-fluid data set.

Size of test	Estimated percentage error-rate
1	7.9
0.01	8.2
0.0001	8.3

Table 6.19

Summaries of covariance selection model fits for each Denver group.
(Concentrations set equal to zero if significance level exceeds 0.01 (set 1) or 0.0001 (set 2) in likelihood ratio test of zero concentration versus saturated model.)

Copenhagen special peripheral-blood data set.

	Set 1		Set 2	
Group	deviance	df	deviance	df
1	161.7	27	456.3	35
2	73.2	26	103.8	30
3	85.6	17	179.0	28
4	94.4	22	239.4	30
5	55.8	25	293.5	33
6	91.6	30	153.1	34
7	91.4	18	335.0	28

Table 6.20

Covariance selection models for each Denver group.

Copenhagen special peripheral-blood data set.

Size of test	Estimated percentage error-rate
1	8.5
0.01	11.2
0.0001	10.8

is also given in these tables. Procedure GC may be viewed as an extreme covariance selection model corresponding to setting concentrations equal to zero if a significance level of 1 is exceeded for the likelihood ratio test of zero concentration versus the saturated model. The average number of concentrations set equal to zero out of the maximum possible 276 and 55 for the models in Tables 6.11, 6.13, 6.15, 6.17 and 6.19 , obtained from using a significance level of 0.01 above which concentrations are set equal to zero, is 194, 189, 189, 23 and 24 respectively to the nearest integer. The average number of concentrations set equal to zero for the second set of models in these tables, obtained from using a significance level of 0.0001 above which concentrations are set equal to zero, is 231, 223, 224, 31 and 31 respectively to the nearest integer.

6.6 Discussion.

The results for a covariance selection model for each class for the Edinburgh, Copenhagen and Philadelphia data sets do not give candidate procedures in the trade-off of estimated percentage error-rate against c.p.u. time. These models are coded as CS1 (concentration set equal to zero when null hypothesis of zero concentration relative to the saturated model gives a significance level greater than 0.01) and CS2 (concentration set equal to zero when null hypothesis of zero concentration relative to the saturated model gives a significance level greater than 0.0001) on Figures 6.2 to 6.6. Both results for the Copenhagen special amniotic-fluid data set and the result for significance level 0.01 for the Copenhagen special peripheral-blood data set do, however, give candidate procedures for the trade-off of estimated percentage error-rate against c.p.u. time (Figures 6.5 and 6.6).

For the covariance selection models for each Denver group for the Edinburgh and Copenhagen data sets both sets of models give candidate procedures for the trade-off of estimated percentage error-rate against c.p.u. time. These models are coded as CS3 (concentration set equal to zero when null hypothesis of zero concentration relative to the saturated model gives a significance level greater than 0.01) and CS4 (concentration set equal to zero when null hypothesis of zero concentration relative to the saturated model gives a significance level greater than 0.0001). For the Copenhagen special data sets only the set of models obtained from using a significance level of

Figure 6.2

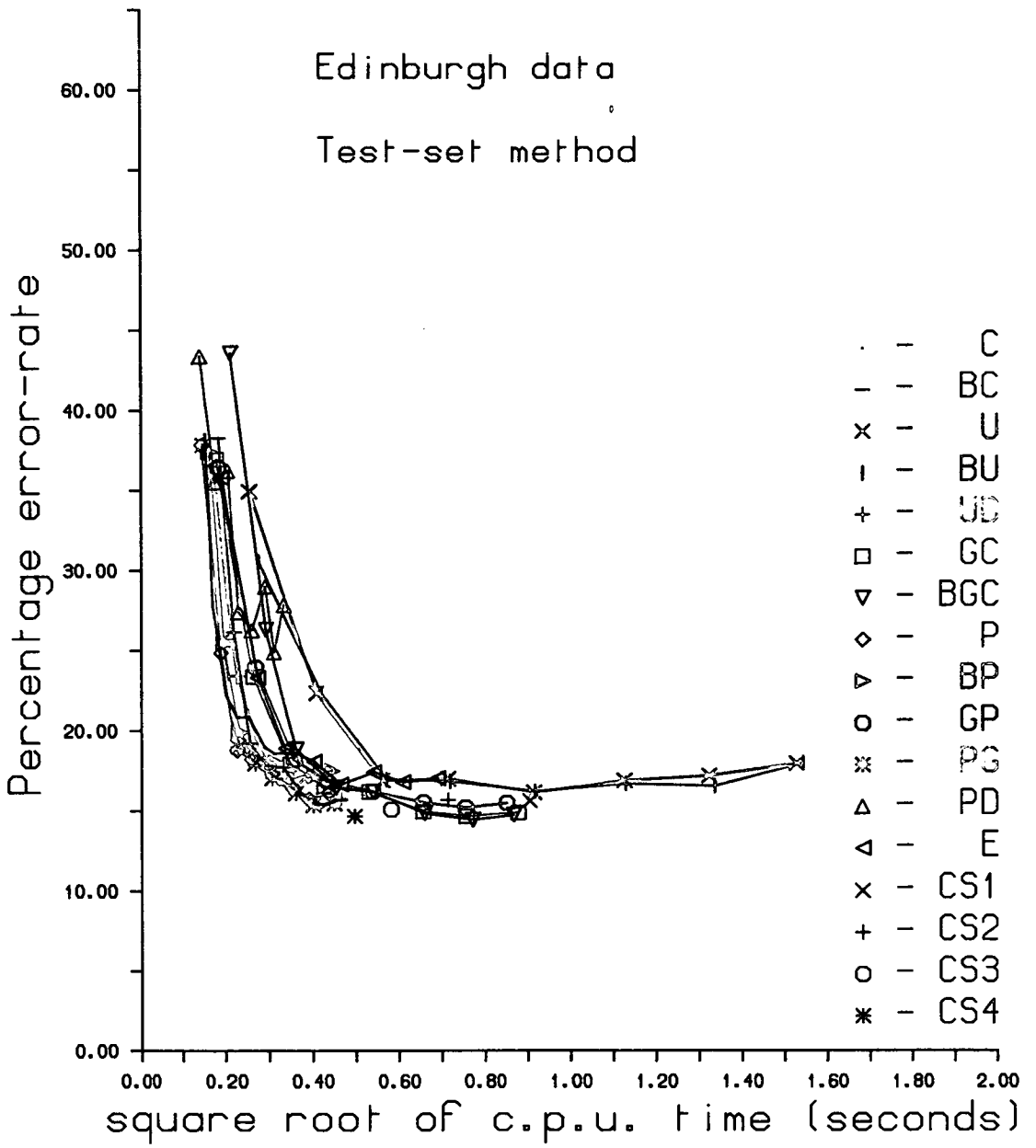

Edinburgh data
Test-set method
Percentage error-rate
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)
. - C
– - BC
× - U
ı - BU
+ - UD
□ - GC
▽ - BGC
◇ - P
▷ - BP
o - GP
⋇ - PG
△ - PD
◁ - E
× - CS1
+ - CS2
o - CS3
✳ - CS4

Figure 6.3

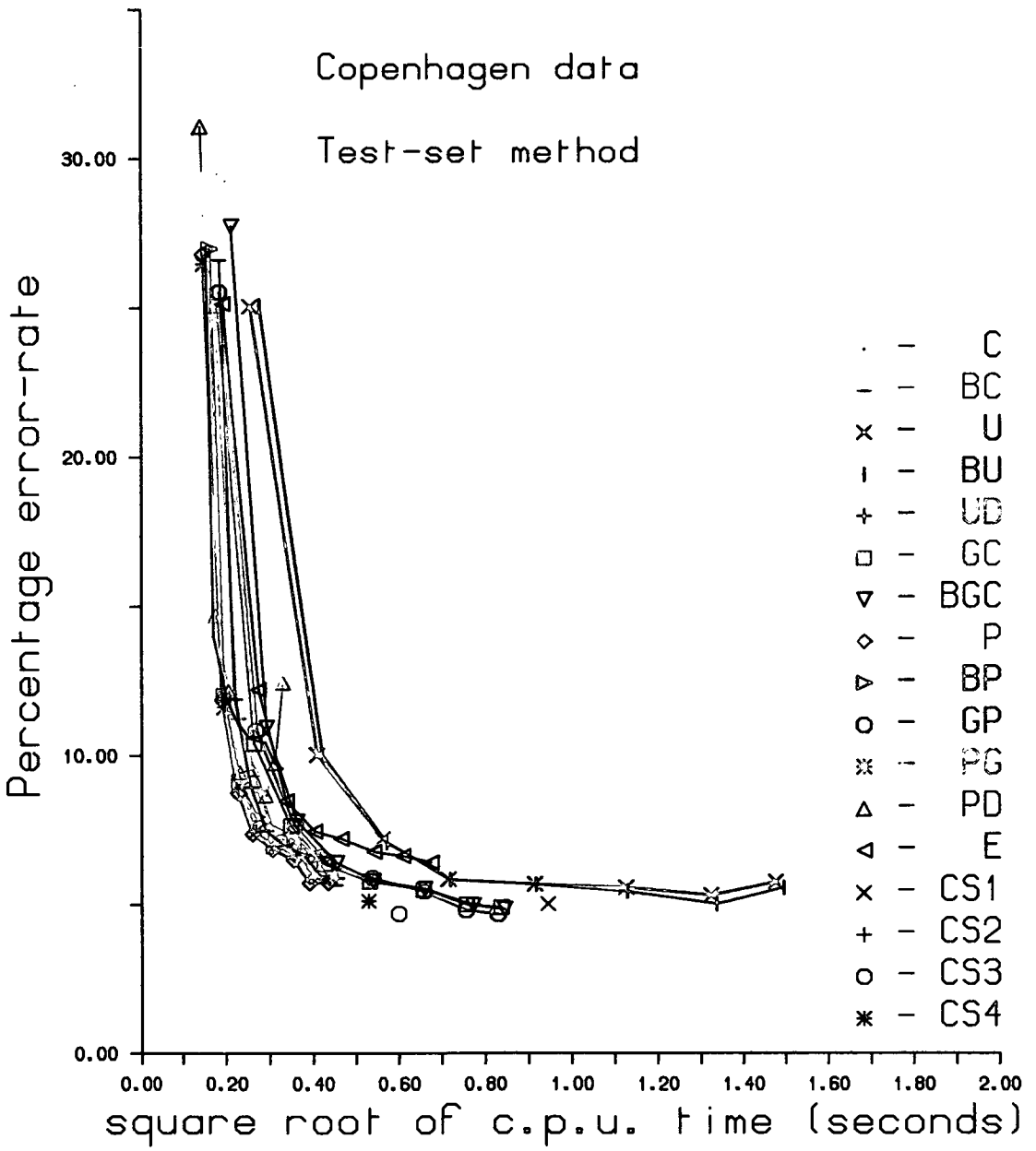

Copenhagen data
Test-set method
Percentage error-rate
30.00
20.00
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)
. - C
- - BC
× - U
ı - BU
+ - UD
□ - GC
▽ - BGC
◇ - P
▷ - BP
o - GP
⁂ - PG
△ - PD
◁ - E
× - CS1
+ - CS2
o - CS3
* - CS4

Figure 6.4

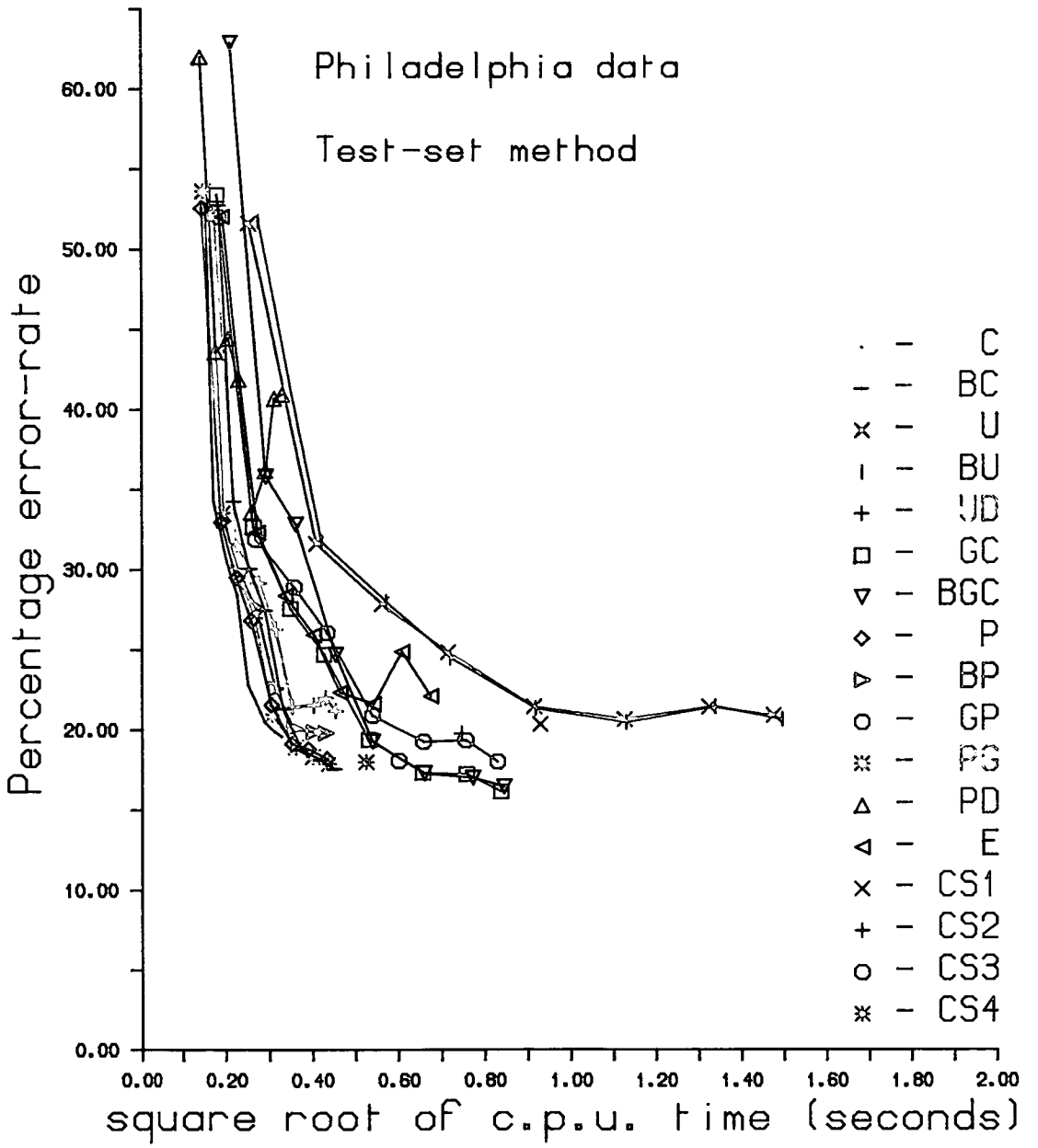

Philadelphia data
Test-set method
Percentage error-rate
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)
. - C
- - BC
x - U
| - BU
+ - UD
□ - GC
▽ - BGC
◇ - P
▷ - BP
○ - GP
※ - PG
△ - PD
◁ - E
× - CS1
+ - CS2
○ - CS3
※ - CS4

Figure 6.5

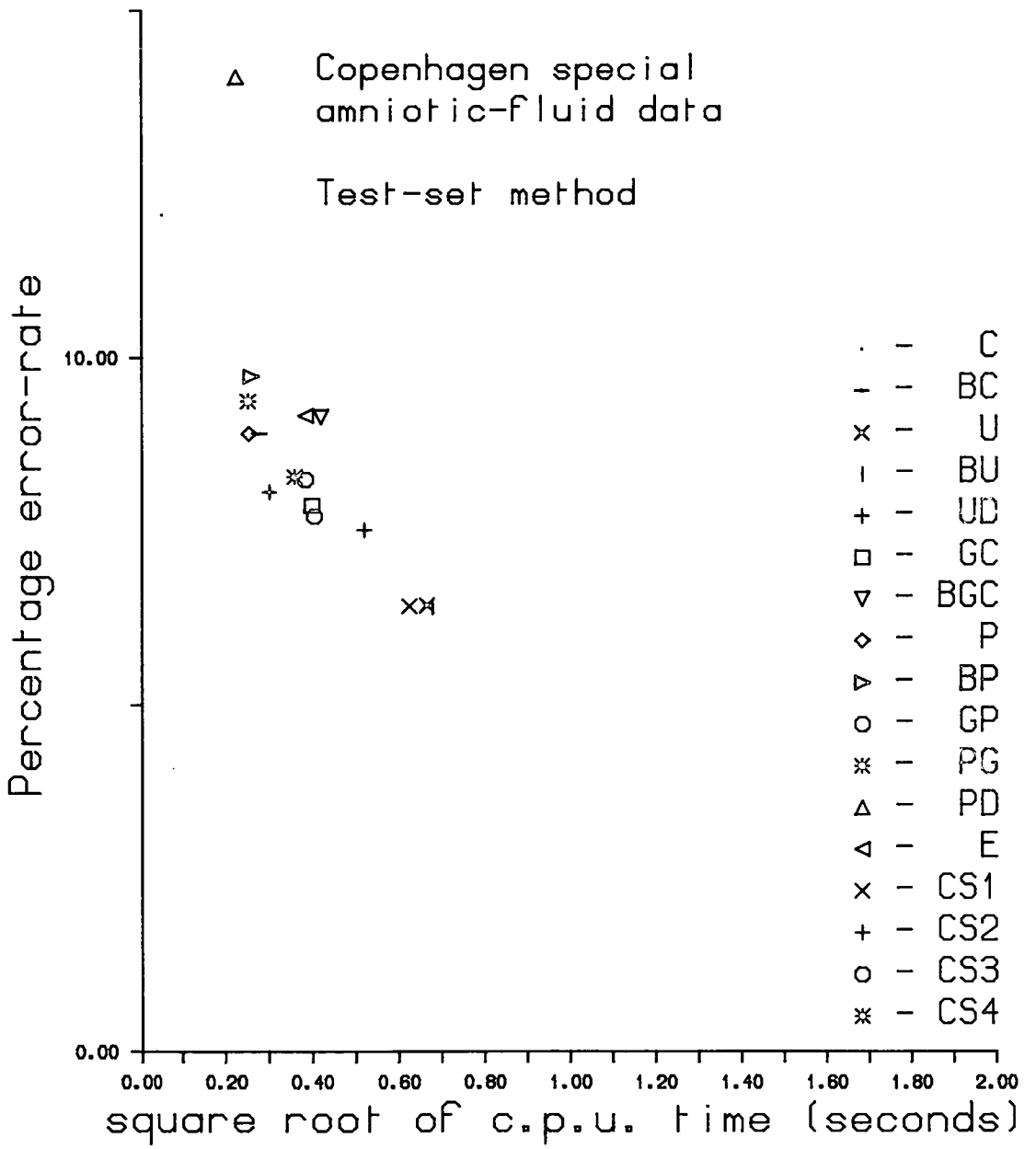
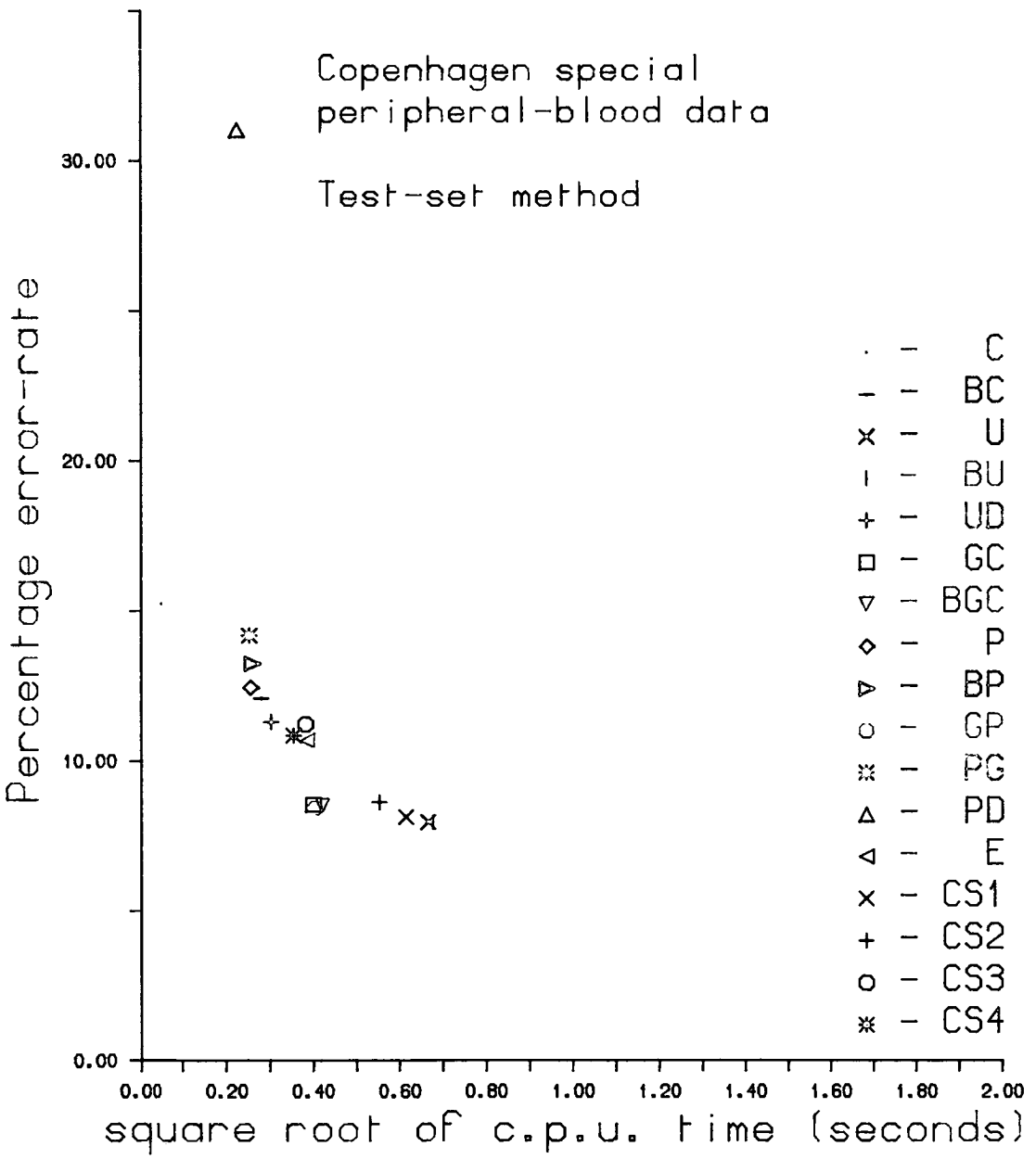

Copenhagen special
amniotic-fluid data
Test-set method
Percentage error-rate
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)
. - C
- - BC
× - U
| - BU
+ - UD
□ - GC
▽ - BGC
◇ - P
▷ - BP
○ - GP
※ - PG
△ - PD
◁ - E
× - CS1
+ - CS2
○ - CS3
※ - CS4

Figure 6.6


Copenhagen special
peripheral-blood data
Test-set method
Percentage error-rate
30.00
20.00
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)
. - C
- - BC
x - U
| - BU
+ - UD
□ - GC
▽ - BGC
◇ - P
▷ - BP
○ - GP
※ - PG
△ - PD
◁ - E
× - CS1
+ - CS2
○ - CS3
※ - CS4

0.0001 for the peripheral-blood data gives a candidate procedure.

The comparison with procedures UD and U from chapter 5 shows that for the Edinburgh, Copenhagen and Philadelphia data sets the estimated percentage error-rates for the sets of covariance selection models CS1 and CS2 are smaller than those for these two procedures. For the special Copenhagen data sets only the set of covariance selection models CS1 for the amniotic-fluid data gives an estimated percentage error-rate as low as procedure U. Both sets of models for both these data sets give lower estimated percentage error-rates than procedure UD.

The comparison with procedure GC shows that only for the Copenhagen data set does one of the set of covariance selection models give a lower estimated percentage error-rate than procedure GC.

Chapter 7
Some two-stage procedures for the calculation of discriminant scores in the automated allocation of human chromosomes.

7.1 Introduction

In chapters 5 and 6 methods of combining class information on variability and covariance selection models were proposed as ways of reducing the numbers of parameters and the numbers of calculations necessary to allocate a chromosome for multivariate Normal discrimination. Another way that the numbers of calculations may be reduced is by the adoption of a multi-stage prcocedure for the allocation of a chromosome to a class. We may define a strictly increasing sequence $E_1,E_2,...,E_m$ of subsets of the p features with E_1 non-empty and hence $m \leqq p$. Such a sequence is implied by an ordering of the features and a strictly increasing sequence defining the number of features included in each subset. The calculation of discriminant scores then proceeds in up to m stages, the features in E_l being used at the lth stage to eliminate some classes from further consideration. These are intended to be the least probable classes given the features observed and discriminant scores are not calculated for them at later stages. For each chromosome the procedure continues either until only one class remains or until the mth and final stage at which the chromosome is allocated to one of the remaining classes. In practice a two-stage procedure is likely to provide computational savings for little additional complexity. This can be seen from Figure 7.1 which is a confusion matrix for procedure U for the Edinburgh data set for the single feature normalised size using the leave-one-cell out error rate estimation method. This figure shows that using just this feature only a minority of the 23 other classes are likely to be confused with the true class of any particular chromosome. The theory of discriminant analysis (Anderson, 1984, page 225) indicates that if the distributions in each class are known such a two-stage procedure would be likely to be sub-optimal and would, therefore, lead to an increase in error rate. However, when the distributions are to be estimated, such a procedure need not lead to an increase in error rate and in this application the trade-off between error rate and allocation time is important rather than error rate alone.

Figure 7.1

Confusion matrix for procedure U from chapter 5 for Edinburgh data set for single feature normalised size. (Columns are predicted classes 1-24, rows are actual classes 1-24.)

(X signifies non-zero entry)

X	X	X	0	0	0	X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X	X	X	X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	X	X	X	X	X	0	0	0	0	0	X	0	0	0	0	0	0	0	0	0	0	0	0
0	X	X	X	X	X	X	0	0	0	0	X	0	0	0	0	0	0	0	0	0	0	0	0
0	0	X	X	X	X	X	0	0	0	0	X	0	0	0	0	0	0	0	0	0	0	0	0
0	0	X	X	X	X	X	X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	X	X	X	X	X	0	0	X	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	X	X	X	X	X	0	X	X	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	X	X	X	X	X	0	X	X	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	X	X	X	X	X	0	X	X	0	0	X	0	0	0	0	0	0	0	0
0	0	0	0	0	0	X	X	X	X	0	X	X	0	0	0	0	X	0	0	0	0	0	0
0	0	0	0	0	0	0	X	X	X	0	X	X	0	0	X	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	X	X	X	X	X	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	X	X	X	X	X	0	X	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	X	X	X	X	X	X	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	X	X	X	X	X	0	X	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	X	X	X	X	X	X	X	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	X	X	X	X	X	0	X	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	X	X	X	X	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	X	0	X	X	0	X	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	X	X	X	X	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	X	X	X	X	0	0
0	0	0	0	X	X	X	X	X	0	0	X	0	0	X	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	X	0	X	0	X	X	X	X	0	0

7.2 Criteria for elimination of classes at first stage.

At least two criteria for the elimination of candidate classes at the first stage for a given subset of features may be suggested.

7.2.1 Estimated posterior probability.

Classes may be eliminated from consideration if their estimated posterior probability is less than a given value γ at the first stage, i.e., class i is eliminated if

$$p(c_i|\underline{x}_k) < \gamma$$

where $p(c_i|\underline{x}_k)$ is the estimated posterior probability of class i given the observed vector of k features $\underline{x}_k$ and γ is in (0,1). This criterion is a natural choice because the one-stage allocation of a chromosome is based on discriminant scores which allocate each chromosome to the class with the biggest estimated posterior probability.

7.2.2 Ratio of estimated posterior probability to maximum estimated posterior probability.

Classes are eliminated at the first stage if

$$p(c_i|\underline{x}_k)\{\max_h p(c_h|\underline{x}_k)\}^{-1} < \delta$$

for some value δ in (0,1). This criterion may be useful if it is apparent from a subset of features that one candidate class is many times more probable than any of the remaining classes.

7.3 Obtaining values for the criteria.

For both criteria for a given set of features at the first stage values of γ and δ may be found by considering the estimated error-rate and c.p.u. time for allocation for a range of values.

In theory a 'best' two-stage procedure for a given ordering of the features

may be found by searching over values of γ and δ for different numbers of features at the first stage. In practice a search over a limited grid of values may be sufficient to give a reasonable combination of criterion value and number of features at the first stage.

It should be noted that the first criterion allows the possibility that all classes may be eliminated after the first stage unless $\gamma \leqq c^{-1}$, where c is the number of classes. This outcome may be dealt with by allocating the chromosome to a 'reject' class and leaving it to be allocated by the human operator or by calculating all the discriminant scores at the second stage to allocate the chromosome, or by relaxing the criterion in such cases. Re-arrangement of the allocations made by these two-stage procedures to achieve a normal complement within a cell is still possible by assigning a probability of zero to a class eliminated at the first stage.

For a given subset of features increasing values of γ and δ will lead to more classes being eliminated and hence to decreased allocation times.

Generally it should be noted that the two criteria are equivalent for the two-class discrimination problem if $\gamma \leqq \frac{1}{2}$ and δ equals $\gamma(1 - \gamma)^{-1}$.

7.4 Four procedures.

As in chapter 5 we may consider combining class information on variability at the second stage when all available features are used, to reduce the number of parameters, calculation time and possibly the percentage error-rate. Such assumptions could also be made at the first stage but for a small number of features this might be expected to be unnecessary. Four procedures may be defined by using each criterion with procedure U at the first stage and procedure U or procedure GC at the second stage.

7.5 Application to Edinburgh, Copenhagen and Philadelphia data sets.

Two of the four procedures were applied to each of the three data sets using at the first stage features selected by the MSEPCOR method for the full data sets displayed in Table 5.3 . For the Edinburgh and Philadelphia data sets the two two-stage procedures used were those which estimate a common covariance matrix per Denver group at the second stage. For the Copenhagen

data set the two two-stage procedures used were those which estimate an unrelated covariance matrix for each class at the second stage. These choices were made because of the relative performance of procedures GC and U for these data sets reported in chapter 5. As in the previous two chapters the normalisation of the data to account for between-cell variation was that currently used and described in chapter 3.

For both criteria, 1, 3, 6, 9, 12, 16, 20 and 24 features were used at the first stage for the following values of γ: 0.01, 0.05, 0.10, 0.20, 0.30, 0.40 and 0.50 . The same values of δ were used with the addition of the values 0.60, 0.70 and 0.80. Values greater than 0.50 for γ correspond to using just the subset of features to make an allocation if the estimated posterior probability for one class is greater than or equal to γ .

For the first criterion, if no candidate classes were left after the first stage the chromosome was considered wrongly allocated.

Percentage error-rates were estimated by the leave-one-out method described in chapter 5. The percentage error-rates for each class were weighted by the prior probabilities for each class to give an overall estimated percentage error-rate. As in chapters 4, 5 and 6 prior probabilities corresponding to all cells from males were used for the Edinburgh data set and prior probabilities corresponding to equal numbers of male and female cells were used for the other two data sets. Also as in chapters 4, 5 and 6 no rearrangement of the allocations to satisfy a normal karyotype was done.

7.6 Results.

Estimated percentage error-rates were tabulated against the expected c.p.u. time required to allocate 46 chromosomes in a normal cell and also an empirical upper bound for the c.p.u. time taken to allocate the chromosomes in a normal cell. This was done for each of the chosen subsets of features at the first stage and criterion value. The expected c.p.u. time was calculated as

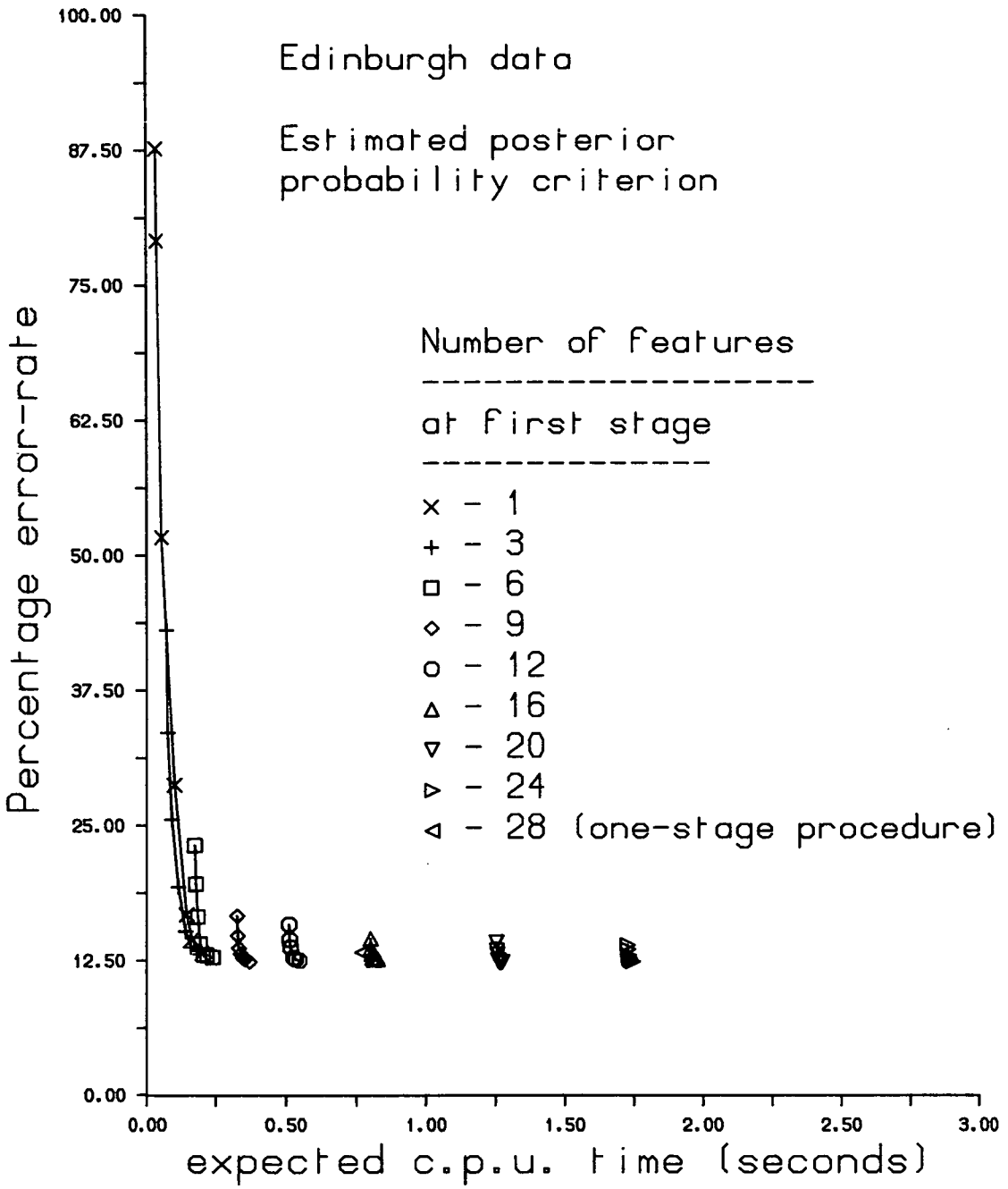
c.p.u. time to calculate all estimated posterior probabilities or ratios of estimated posterior probabilities to maximum estimated posterior probability for the subset of features for 46 chromosomes

+

Σ_i 46 * P_i * average number of classes left for a class i chromosome * average c.p.u. time for one discriminant score

where P_i is the prior probability of class i . For the Edinburgh data, estimated percentage error-rate has been plotted against expected c.p.u. time in Figure 7.2 as an example. The empirical upper bound for the c.p.u. time was calculated as

Figure 7.2


Edinburgh data
Estimated posterior probability criterion
Number of features
at first stage
× - 1
+ - 3
□ - 6
◇ - 9
o - 12
△ - 16
▽ - 20
▷ - 24
◁ - 28 (one-stage procedure)
Percentage error-rate
100.00
87.50
75.00
62.50
50.00
37.50
25.00
12.50
0.00
0.00
0.50
1.00
1.50
2.00
2.50
3.00
expected c.p.u. time (seconds)

c.p.u. time to calculate all estimated posterior probabilities or ratios of estimated posterior probabilities to maximum estimated posterior probability for the subset of features for 46 chromosomes

+

$\Sigma_{i=1}^{1=22}$ 2 * maximum number of classes left for a class i chromosome * average c.p.u. time for one discriminant score

+

max[2 * maximum number of classes left for a class 23 chromosome * average c.p.u. time for one discriminant score, {(maximum number of classes left for a class 23 chromosome + maximum number of classes left for a class 24 chromosome) * average c.p.u. time for one discriminant score}]

for the Copenhagen and Philadelphia data sets. For the Edinburgh data set which contains only cells from males the possibilty of two class 23 chromosomes was ignored. The discriminant scores were calculated as described in chapter 5 and the estimated posterior probabilities derived from these discriminant scores. All c.p.u. timings are the average of ten times recorded on the Edinburgh University NAS computer using the same operands for programs written in Fortran 77 with double-precision arithmetic. The c.p.u. time taken to set a 'flag' showing whether a discriminant score is to be calculated for a class at the second stage was assumed to be negligible. Also assumed negligible is the time taken to check the value of a 'flag' at the second stage.

7.6.1 Edinburgh data.

Tables 7.1 and 7.2 show that marked proportional savings in expected c.p.u. time may be achieved for no increase in the final estimated percentage error-rate compared with the one-stage procedure, GC, for both criteria. Indeed there is some evidence that a lower estimated percentage error-rate is achieved for some feature subsets. It may be conjectured that this is sometimes because of the estimation of fewer parameters at the first stage compared with the one-stage procedure. In some cases where the expected c.p.u. time is less than for the one-stage procedure and the percentage error-rate is also less than or equal to that of the one-stage procedure the empirical upper bound is less than the expected c.p.u. time for the one-stage procedure.

7.6.2 Copenhagen data.

Tables 7.3 and 7.4 show that an estimated percentage error-rate as low as for the one-stage procedure, U, may be achieved along with a reduction in expected c.p.u. time for both criteria. In all of these cases the corresponding result for the empirical upper bound is less than the expected c.p.u time for the one-stage procedure, U.

7.6.3 Philadelphia data.

Table 7.6 shows that an estimated percentage error-rate as low as or lower than the one-stage procedure, GC, may be achieved with marked proportional savings in expected c.p.u. time. However, none of the corresponding values for the empirical upper bound is less than the expected c.p.u. time for the one-stage procedure, GC.

7.7 Discussion

The results for the three data sets show that it is possible to get estimated percentage error-rates as low as or lower than that for the one-stage procedures used for a considerable reduction in expected allocation time. The maximum proportional savings in expected allocation time for estimated percentage error-rates as low as the one-stage procedures are 0.75, 0.39 and 0.43 for the Edinburgh, Copenhagen and Philadelphia data sets respectively. These results are achieved for 3, 20 and 1 feature at the first stage for the value 0.01 for the first criterion, 0.01 for the second criterion and 0.01 for the second criterion. The corresponding empirical upper bounds are less than the

Table 7.1

Estimated percentage error-rates, expected c.p.u. times and empirical upper bounds for c.p.u. time for estimated posterior probability criterion for Edinburgh data set.

Unrelated covariance matrices estimated at first stage, common covariance matrix per Denver group estimated at second stage. (Estimated percentage error-rate, followed by expected c.p.u. time and empirical upper bound for c.p.u. time.)

Number of features at first stage	Estimated posterior probability 0.01	0.05	0.10
1	13.1,0.20,0.47	14.4,0.16,0.38	16.7,0.15,0.31
3	13.0,0.19,0.41	13.7,0.16,0.34	15.2,0.14,0.31
6	12.7,0.24,0.45	13.0,0.22,0.39	13.0,0.21,0.35
9	12.3,0.37,0.57	12.6,0.35,0.51	12.7,0.35,0.49
12	12.4,0.55,0.73	12.6,0.54,0.68	12.7,0.53,0.66
16	12.5,0.83,0.99	12.6,0.82,0.96	12.5,0.82,0.93
20	12.3,1.28,1.43	12.2,1.27,1.39	12.3,1.26,1.38
24	12.3,1.75,1.89	12.4,1.74,1.86	12.5,1.74,1.84

Number of features at first stage	Estimated posterior probability 0.20	0.30	0.40
1	28.7,0.11,0.23	51.7,0.06,0.15	79.1,0.04,0.05
3	19.3,0.12,0.24	25.5,0.10,0.20	33.6,0.08,0.17
6	14.0,0.19,0.32	16.5,0.19,0.29	19.5,0.18,0.26
9	13.1,0.34,0.46	13.6,0.33,0.42	14.7,0.33,0.40
12	13.3,0.52,0.63	13.7,0.52,0.60	14.4,0.51,0.58
16	12.7,0.81,0.92	13.1,0.81,0.88	13.6,0.80,0.87
20	12.5,1.26,1.36	13.0,1.26,1.33	13.4,1.26,1.32
24	12.6,1.73,1.82	12.9,1.73,1.81	13.5,1.73,1.78

Table 7.1 (continued)

Estimated percentage error-rates, expected c.p.u. times and empirical upper bounds for c.p.u. time for estimated posterior probability criterion for Edinburgh data set.

Unrelated covariance matrices estimated at first stage, common covariance matrix per Denver group estimated at second stage. (Estimated percentage error-rate, followed by expected c.p.u. time and empirical upper bound for c.p.u. time.)

Number of features at first stage	Estimated posterior probability 0.50
1	87.7,0.04,0.04
3	43.0,0.08,0.08
6	23.1,0.18,0.18
9	16.6,0.33,0.33
12	15.8,0.51,0.51
16	14.5,0.80,0.80
20	14.1,1.25,1.25
24	13.9,1.73,1.73

Result for one-stage procedure, GC

13.2,0.77,0.77

Table 7.2

Estimated percentage error-rates, expected c.p.u. times and empirical upper bounds for c.p.u. time for ratio of estimated posterior probability to maximum estimated posterior probability criterion for Edinburgh data set.

Unrelated covariance matrices estimated at first stage, common covariance matrix per Denver group estimated at second stage. (Estimated percentage error-rate, followed by expected c.p.u. time and empirical upper bound for c.p.u. time.)

Number of features at first stage	Ratio of estimated posterior probability to maximum estimated posterior probability: 0.01	0.05	0.10
1	13.0,0.23,0.50	13.3,0.19,0.43	14.1,0.18,0.40
3	13.0,0.21,0.44	13.4,0.17,0.38	14.1,0.16,0.35
6	12.8,0.25,0.48	12.9,0.23,0.41	13.0,0.22,0.40
9	12.3,0.38,0.60	12.4,0.37,0.54	12.5,0.36,0.51
12	12.4,0.57,0.76	12.6,0.56,0.72	12.6,0.55,0.70
16	12.6,0.83,1.00	12.6,0.82,0.97	12.5,0.82,0.95
20	12.2,1.33,1.49	12.1,1.32,1.45	12.3,1.31,1.43
24	12.3,1.79,1.94	12.4,1.78,1.91	12.5,1.78,1.89

Number of features at first stage	Ratio of estimated posterior probability to maximum estimated posterior probability: 0.20	0.30	0.40
1	15.6,0.16,0.36	17.8,0.14,0.32	21.2,0.13,0.30
3	16.0,0.14,0.31	18.1,0.13,0.30	20.3,0.12,0.27
6	13.5,0.21,0.36	14.2,0.21,0.34	15.3,0.20,0.33
9	12.9,0.35,0.50	13.0,0.35,0.48	13.4,0.35,0.47
12	12.9,0.54,0.67	13.4,0.54,0.66	13.6,0.54,0.64
16	12.6,0.82,0.94	12.8,0.81,0.92	12.8,0.81,0.91
20	12.5,1.31,1.42	12.5,1.31,1.41	12.8,1.31,1.40
24	12.6,1.78,1.88	12.8,1.78,1.86	12.9,1.77,1.86

Table 7.2 (continued)

Estimated percentage error-rates, expected c.p.u. times and empirical upper bounds for c.p.u. time for ratio of estimated posterior probability to maximum estimated posterior probability criterion for Edinburgh data set.

Unrelated covariance matrices estimated at first stage, common covariance matrix per Denver group estimated at second stage. (Estimated percentage error-rate, followed by expected c.p.u. time and empirical upper bound for c.p.u. time.)

Number of features at first stage	Ratio of estimated posterior probability to maximum estimated posterior probability: 0.50	0.60	0.70
1	24.5,0.12,0.29	28.4,0.11,0.27	32.9,0.10,0.25
3	23.0,0.11,0.25	25.2,0.10,0.23	27.2,0.10,0.21
6	16.3,0.20,0.31	17.3,0.20,0.31	18.5,0.19,0.29
9	13.5,0.34,0.45	14.1,0.34,0.43	14.4,0.34,0.42
12	13.6,0.54,0.64	13.8,0.54,0.62	14.2,0.53,0.60
16	12.9,0.81,0.90	13.1,0.81,0.89	13.3,0.81,0.88
20	13.0,1.31,1.39	13.2,1.31,1.38	13.4,1.31,1.37
24	13.1,1.77,1.85	13.4,1.77,1.84	13.5,1.77,1.83

Number of features at first stage	Ratio of estimated posterior probability to maximum estimated posterior probability: 0.80
1	40.3,0.08,0.23
3	29.3,0.09,0.19
6	19.1,0.19,0.28
9	14.7,0.34,0.40
12	14.5,0.53,0.59
16	13.6,0.81,0.86
20	13.4,1.30,1.36
24	13.6,1.77,1.82

Result for one-stage procedure, GC

13.2,0.77,0.77

Table 7.3

Estimated percentage error-rates, expected c.p.u. times and empirical upper bounds for c.p.u. time for estimated posterior probability criterion for Copenhagen data set.

Unrelated covariance matrices estimated at each stage. (Estimated percentage error-rate, followed by expected c.p.u. time and empirical upper bound for c.p.u. time.)

Number of features at first stage	Estimated posterior probability 0.01	0.05	0.10
1	4.8,0.59,1.55	6.8,0.46,1.13	11.3,0.37,0.97
3	4.9,0.35,1.14	5.9,0.27,0.85	7.0,0.22,0.70
6	4.4,0.26,0.86	5.2,0.22,0.72	5.7,0.21,0.62
9	4.4,0.38,0.88	4.7,0.35,0.75	4.9,0.34,0.68
12	4.0,0.54,1.03	4.2,0.53,0.90	4.4,0.52,0.84
16	4.0,0.82,1.26	4.2,0.81,1.14	4.2,0.81,1.07
20	4.0,1.27,1.65	4.0,1.26,1.57	4.1,1.26,1.53
24	3.9,1.74,2.11	3.9,1.73,2.02	4.0,1.73,1.99

Number of features at first stage	Estimated posterior probability 0.20	0.30	0.40
1	33.7,0.21,0.57	61.5,0.08,0.33	87.0,0.04,0.10
3	11.1,0.17,0.55	16.4,0.12,0.41	24.2,0.09,0.31
6	6.7,0.19,0.52	7.7,0.19,0.45	8.9,0.18,0.36
9	5.4,0.34,0.63	5.8,0.33,0.56	6.2,0.33,0.51
12	4.5,0.52,0.75	4.8,0.52,0.70	5.0,0.51,0.70
16	4.3,0.81,0.99	4.5,0.81,0.97	4.5,0.80,0.96
20	4.4,1.26,1.47	4.4,1.26,1.45	4.5,1.26,1.40
24	3.9,1.73,1.93	3.9,1.73,1.89	3.9,1.73,1.86

Table 7.3 (continued)

Estimated percentage error-rates, expected c.p.u. times and empirical upper bounds for c.p.u. time for estimated posterior probability criterion for Copenhagen data set.

Unrelated covariance matrices estimated at each stage. (Estimated percentage error, followed by expected c.p.u. time and empirical upper bound for c.p.u. time.)

Number of features at first stage	Estimated posterior probability 0.50
1	90.9,0.04,0.04
3	33.5,0.08,0.08
6	10.2,0.18,0.18
9	6.9,0.33,0.33
12	5.2,0.51,0.51
16	4.7,0.80,0.80
20	4.5,1.25,1.25
24	4.0,1.73,1.73

Result for one-stage procedure, U

3.9,2.18,2.18

Table 7.4

Estimated percentage error-rates, expected c.p.u. times and empirical upper bounds for c.p.u. time for ratio of estimated posterior probability to maximum estimated posterior probability criterion for Copenhagen data set.

Unrelated covariance matrices estimated at each stage. (Estimated percentage error-rate, followed by expected c.p.u. time and empirical upper bound for c.p.u. time.)

Number of features at first stage	Ratio of estimated posterior probability to maximum estimated posterior probability: 0.01	0.05	0.10
1	4.6,0.67,1.73	5.1,0.56,1.45	5.9,0.50,1.28
3	4.7,0.37,1.26	5.6,0.29,0.98	6.1,0.26,0.87
6	4.4,0.27,0.91	5.1,0.24,0.77	5.5,0.22,0.68
9	4.3,0.39,0.92	4.6,0.37,0.80	4.8,0.36,0.74
12	4.0,0.56,1.06	4.2,0.55,0.95	4.3,0.54,0.90
16	4.0,0.83,1.27	4.1,0.82,1.18	4.2,0.82,1.10
20	3.9,1.32,1.72	4.0,1.31,1.63	4.1,1.31,1.61
24	3.9,1.78,2.17	3.9,1.78,2.08	4.0,1.78,2.05

Number of features at first stage	Ratio of estimated posterior probability to maximum estimated posterior probability: 0.20	0.30	0.40
1	7.9,0.43,1.11	10.6,0.39,1.03	13.4,0.35,0.98
3	7.6,0.22,0.74	9.3,0.19,0.68	11.3,0.17,0.63
6	6.3,0.21,0.61	6.7,0.20,0.57	7.1,0.20,0.52
9	5.2,0.35,0.69	5.4,0.35,0.65	5.6,0.35,0.61
12	4.4,0.54,0.84	4.5,0.54,0.79	4.7,0.54,0.75
16	4.3,0.81,1.05	4.3,0.81,1.01	4.4,0.81,0.99
20	4.3,1.31,1.56	4.3,1.31,1.53	4.4,1.31,1.51
24	3.9,1.77,2.02	3.9,1.77,1.97	3.9,1.77,1.97

Table 7.4 (continued)

Estimated percentage error-rates, expected c.p.u. times and empirical upper bounds for c.p.u. time for ratio of estimated posterior probability to maximum estimated posterior probability criterion for Copenhagen data set.

Unrelated covariance matrices estimated at each stage. (Estimated percentage error-rate, followed by expected c.p.u. time and empirical upper bound for c.p.u. time.)

Number of features at first stage	Ratio of estimated posterior probability to maximum estimated posterior probability: 0.50	0.60	0.70
1	17.6,0.31,0.86	23.0,0.27,0.75	29.3,0.23,0.72
3	13.4,0.15,0.55	15.3,0.13,0.51	17.4,0.12,0.47
6	7.5,0.20,0.50	8.1,0.19,0.44	8.6,0.19,0.40
9	5.8,0.34,0.59	6.0,0.34,0.56	6.1,0.34,0.54
12	4.8,0.53,0.73	4.9,0.53,0.72	5.0,0.53,0.72
16	4.5,0.81,0.98	4.5,0.81,0.97	4.5,0.81,0.97
20	4.4,1.31,1.50	4.5,1.30,1.50	4.5,1.30,1.47
24	3.9,1.77,1.93	3.9,1.77,1.93	3.9,1.77,1.91

Number of features at first stage	Ratio of estimated posterior probability to maximum estimated posterior probability: 0.80
1	36.6,0.19,0.70
3	19.7,0.10,0.42
6	9.1,0.19,0.37
9	6.3,0.34,0.51
12	5.1,0.53,0.69
16	4.6,0.81,0.95
20	4.5,1.31,1.42
24	3.9,1.77,1.88

Result for one-stage procedure, U

3.9,2.18,2.18

Table 7.5

Estimated percentage error-rates, expected c.p.u. times and empirical upper bounds for c.p.u. time for estimated posterior probability criterion for Philadelphia data set.

Unrelated covariance matrices estimated at first stage, common covariance matrix per Denver group estimated at second stage. (Estimated percentage error-rate, followed by expected c.p.u. time and empirical upper bound for c.p.u. time.)

Number of features at first stage	Estimated posterior probability 0.01	0.05	0.10
1	17.8,0.33,0.79	21.9,0.23,0.56	34.6,0.16,0.41
3	17.7,0.28,0.73	20.2,0.20,0.49	25.3,0.16,0.36
6	17.8,0.29,0.66	18.6,0.25,0.49	19.9,0.23,0.40
9	17.7,0.42,0.73	18.2,0.38,0.61	18.7,0.37,0.54
12	17.7,0.59,0.89	18.4,0.56,0.77	19.0,0.54,0.71
16	17.6,0.86,1.11	17.9,0.83,1.02	18.5,0.83,0.97
20	17.5,1.29,1.53	17.8,1.28,1.46	18.0,1.27,1.43
24	17.5,1.76,1.98	17.4,1.75,1.92	17.8,1.74,1.90

Number of features at first stage	Estimated posterior probability 0.20	0.30	0.40
1	75.2,0.06,0.16	91.1,0.04,0.06	93.4,0.04,0.05
3	39.4,0.11,0.24	53.8,0.09,0.17	67.6,0.08,0.13
6	23.4,0.20,0.33	28.7,0.19,0.28	35.0,0.18,0.24
9	21.2,0.35,0.48	24.5,0.34,0.43	28.9,0.33,0.40
12	20.4,0.53,0.65	22.6,0.52,0.61	25.4,0.51,0.58
16	19.1,0.82,0.93	19.8,0.81,0.91	21.2,0.81,0.87
20	18.6,1.26,1.38	19.0,1.26,1.35	19.8,1.26,1.32
24	18.3,1.74,1.85	18.6,1.73,1.83	19.4,1.73,1.80

Table 7.5 (continued)

Estimated percentage error-rates, expected c.p.u. times and empirical upper bound for c.p.u. time for estimated posterior probability criterion for Philadelphia data set.

Unrelated covariance matrices estimated at first stage, common covariance matrix per Denver group estimated at second stage. (Estimated percentage error-rate, followed by expected c.p.u. time and empirical upper bound for c.p.u. time.)

Number of features at first stage	Estimated posterior probability 0.50
1	96.4,0.04,0.04
3	77.4,0.08,0.08
6	42.9,0.18,0.18
9	35.3,0.33,0.33
12	29.5,0.51,0.51
16	23.2,0.80,0.80
20	21.3,1.25,1.25
24	20.7,1.73,1.73

Result for one-stage procedure, GC

17.6,0.70,0.70

Table 7.6

Estimated percentage error-rates, expected c.p.u. times and empirical upper bounds for c.p.u. time for ratio of estimated posterior probability to maximum estimated posterior probability criterion for Philadelphia data set.

Unrelated covariance matrices estimated at first stage, common covariance matrix per Denver group estimated at second stage. (Estimated percentage error-rate, followed by expected c.p.u. time and empirical upper bound for c.p.u. time.)

Number of features at first stage	Ratio of estimated posterior probability to maximum estimated posterior probability: 0.01	0.05	0.10
1	17.5,0.40,0.99	17.9,0.33,0.80	18.3,0.29,0.75
3	17.7,0.34,0.86	18.1,0.26,0.71	19.0,0.22,0.63
6	17.7,0.32,0.73	18.3,0.27,0.60	18.8,0.25,0.54
9	17.6,0.44,0.80	17.8,0.41,0.68	18.3,0.39,0.64
12	17.7,0.61,0.95	18.2,0.58,0.85	18.6,0.57,0.81
16	17.6,0.86,1.15	17.8,0.84,1.06	18.2,0.83,1.02
20	17.5,1.35,1.60	17.6,1.33,1.53	17.9,1.32,1.51
24	17.5,1.81,2.05	17.3,1.79,1.99	17.8,1.79,1.96

Number of features at first stage	Ratio of estimated posterior probability to maximum estimated posterior probability: 0.20	0.30	0.40
1	20.7,0.25,0.64	23.0,0.22,0.59	26.8,0.20,0.52
3	21.7,0.19,0.52	25.2,0.17,0.44	29.1,0.15,0.40
6	20.1,0.23,0.47	21.2,0.22,0.42	22.8,0.21,0.39
9	19.0,0.37,0.58	20.2,0.37,0.54	21.5,0.36,0.51
12	19.3,0.56,0.75	20.1,0.55,0.72	20.9,0.55,0.70
16	18.7,0.83,0.98	19.0,0.82,0.96	19.2,0.82,0.94
20	18.3,1.32,1.48	18.4,1.31,1.45	18.6,1.31,1.43
24	17.9,1.78,1.94	18.4,1.78,1.91	18.6,1.78,1.89

Table 7.6 (continued)

Estimated percentage error-rates, expected c.p.u. times and empirical upper bounds for c.p.u. time for ratio of estimated posterior probability to maximum estimated posterior probability criterion for Philadelphia data set.

Unrelated covariance matrices estimated at first stage, common covariance matrix per Denver group estimated at second stage. (Estimated percentage error-rate, followed by expected c.p.u. time and empirical upper bound for c.p.u. time.)

Number of features at first stage	Ratio of estimated posterior probability to maximum estimated posterior probability: 0.50	0.60	0.70
1	31.1,0.18,0.45	36.2,0.16,0.40	43.0,0.14,0.38
3	32.9,0.14,0.35	36.9,0.12,0.30	40.9,0.11,0.26
6	24.6,0.21,0.37	26.2,0.20,0.35	27.6,0.20,0.32
9	22.6,0.35,0.50	23.8,0.35,0.47	24.7,0.35,0.46
12	21.7,0.54,0.68	22.7,0.54,0.66	23.5,0.54,0.64
16	19.8,0.82,0.93	20.2,0.81,0.92	20.5,0.81,0.90
20	18.8,1.31,1.42	19.2,1.31,1.41	19.4,1.31,1.39
24	18.7,1.78,1.88	19.0,1.78,1.87	19.0,1.77,1.86

Number of features at first stage	Ratio of estimated posterior probability to maximum estimated posterior probability: 0.80
1	49.4,0.12,0.31
3	45.0,0.10,0.24
6	29.0,0.19,0.30
9	26.0,0.34,0.44
12	23.7,0.54,0.63
16	20.9,0.81,0.89
20	19.7,1.31,1.37
24	19.2,1.77,1.84

Result for one-stage procedure, GC

17.6,0.70,0.70

expected allocation time for the one-stage procedures for the Edinburgh and Copenhagen data sets but not for the Philadelphia data set. It is conjectured that this reflects the quality of the data which was much worse for the Philadelphia data set.

The Edinburgh and Philadelphia data sets both show that the use of one feature at the first stage can give large savings in the expected allocation time for 46 chromosomes in a normal cell and an estimated percentage error-rate lower than or as low as the one-stage procedure used. For each of these data sets the first feature chosen by the MSEPCOR feature selection procedure is either a measure of chromosome size or is related to chromosome size. The first feature for the Copenhagen data set is also a measure of chromosome size. To see if this one feature could give an expected allocation time and estimated percentage error-rate less than the one-stage procedure, U, results for the following further values of the two criteria were obtained: 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8} and 10^{-9}. These further results are given in Tables 7.7 and 7.8 . The results show that an estimated percetage error-rate as low as the one-stage procedure, U, is not obtainable for any of these values. However, on occasions, the estimated percentage error-rate is very close to that of the one-stage procedure and the expected allocation time is less than that of the one-stage procedure. This indicates that the use of a single size feature in the first stage of these two-stage procedures may be generally useful.

Table 7.7

Estimated percentage error-rates, expected c.p.u. times and empirical upper bounds for c.p.u. time for estimated posterior probability criterion for Copenhagen data set.

Unrelated covariance matrices estimated at each stage. (Estimated percentage error-rate, followed by expected c.p.u. time and empirical upper bound for c.p.u. time.)

	Estimated posterior probability		
Number of features at first stage	10^{-4}	10^{-5}	10^{-6}
1	4.3,0.73,1.85	4.1,0.86,2.04	4.0,0.98,2.32

	Estimated posterior probability		
Number of features at first stage	10^{-7}	10^{-8}	10^{-9}
1	4.0,1.07,2.57	4.0,1.15,2.70	4.0,1.22,2.90

Result for one-stage procedure, U

3.9,2.18,2.18

Table 7.8

Estimated percentage error-rates, expected c.p.u. times and empirical upper bounds for c.p.u. time for ratio of estimated posterior probability to maximum estimated posterior probability criterion for Copenhagen data set.

Unrelated covariance matrices estimated at each stage. (Estimated percentage error-rates, followed by expected c.p.u. time and empirical upper bound for c.p.u. time.)

Number of features at first stage	Ratio of estimated posterior probability to maximum estimated posterior probability: 10^{-4}	10^{-5}	10^{-6}
1	4.2,0.81,1.98	4.1,0.93,2.25	4.0,1.03,2.50

Number of features at first stage	Ratio of estimated posterior probability to maximum estimated posterior probability: 10^{-7}	10^{-8}	10^{-9}
1	4.0,1.11,2.65	4.0,1.18,2.76	4.0,1.25,2.93

Result for one-stage procedure, U

3.9,2.18,2.18

Chapter 8
The application of three non-parametric methods and a semi-parametric method to the automated allocation of human chromosomes.

8.1 Introduction.

In this chapter the automated allocation of human chromosomes by versions of three non-parametric methods and a semi-parametric method is described. These four methods are:

1. Classification trees.
2. Nearest neighbour discrimination.
3. Kernel density discrimination.
4. Logistic discrimination.

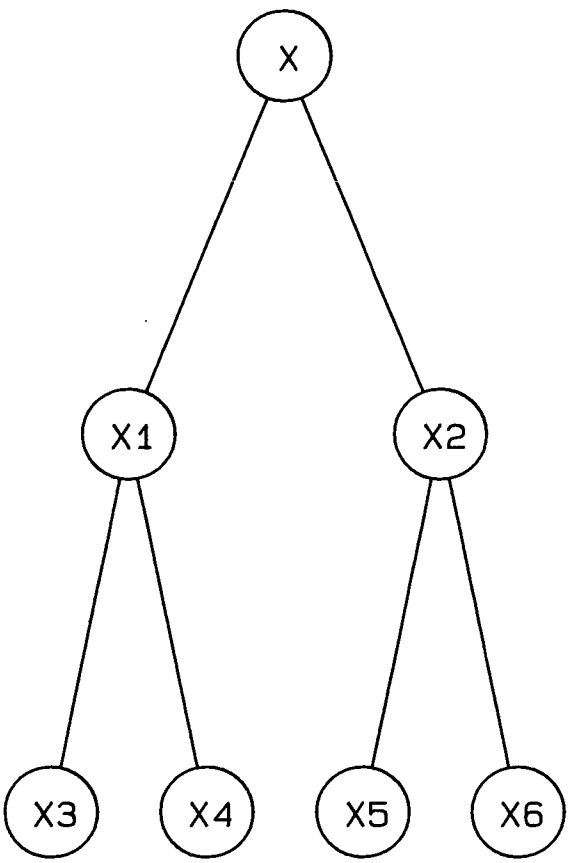
The methods are first described and then versions of three of them are applied to the five data sets used in chapters 5 and 6.

8.2 Classification trees.

A discrimination method described by Breiman et al (1984) is to use a so-called binary tree classifier. This is constructed by repeated splits of subsets of training data into two descendant subsets, beginning with all the training data. The two descendant subsets of each subset are disjoint and their union is equal to the subset. The classifier is called a tree classifier because it may be pictured as in Figure 8.1 . This figure shows an inverted tree with the root at the top. All the training data starts at the root. The training data is then repeatedly split by conditions on the elements of the feature vector $\underline{x}$. In the figure, X refers to all the training data and X1 to X6 to subsets of the training data. The subsets of the training data are referred to as nodes and these are joined by branches. If a node has no further nodes beneath it (descendant nodes) it is referred to as a terminal node. In the figure, nodes X3 , X4 , X5 and X6 are terminal nodes. Each terminal node is given a class label. The tree classifier allocates a new object to a class, given a feature vector $\underline{x}$, as follows: from the definition of the split at the root node it is determined whether the object goes to the left or right descendant node of the root node. The definition of the split at the next node is then used to send the object to the appropriate descendant node of this next node. The process continues

Figure 8.1

Classification tree



until the object reaches a terminal node and it is then allocated to the class corresponding to the class label of the terminal node.

The construction of such a tree using a training set of data requires:

1. A set of questions at each node (usually defined such that a split depends on the value of a single feature).
2. A 'goodness of split' criterion.
3. A rule to stop splitting nodes.
4. A class allocation for each terminal node.

For a quantitative feature, x , the set of questions is

$$x \leqq c_q? \quad q=1,\ldots,n' \leqq (n-1) \tag{8.1}$$

where c_q is defined as mid-way between consecutive distinct values for all n ordered values of the feature. For a categorical feature, x , the set of questions is

$$x \in S? \tag{8.2}$$

where S ranges over all subsets of the possible values for the feature. This gives $2^{(l-1)}$ possible splits for a categorical feature which may take one of l values.

Breiman et al (1984) consider a number of so-called impurity measures for measuring 'goodness of split', at each node, which satisfy the following criteria:

1. The measure has a maximum value if the resubstitution estimates of the probability of a class at a node are equal for all classes.
2. The measure has a minimum value when the resubstitution estimate of the probability of a class at a node equals one.
3. The measure is a symmetric function of the resubstitution estimates of the probability of a class at a node.
4. The measure is a strictly concave function of the resubstitution estimates

of the probability of a class at a node.

The resubstitution estimate of the probability of a class i at a node t is given by

$$p(i|t) = P_i n_i(t) n_i^{-1} p(t)^{-1} \qquad (8.3)$$

where P_i is the prior probability of class i , $n_i(t)$ is the number of class i objects in t in the training set, n_i is the number of class i objects in the training set and

$$p(t) = \Sigma_i P_i n_i(t) n_i^{-1} \quad .$$

Breiman et al (1984) found that the choice of measure did not appear to be crucial to the results obtained if the splitting rule satisfied these criteria. Given a measure of node impurity at a node t , imp(t) , the 'goodness of split' of a split, s, sending proportions p_L and p_R to left and right descendant nodes t_L and t_R is given by

$$\Delta(s,t) = imp(t) - p_L imp(t_L) - p_R imp(t_R) \quad . \qquad (8.4)$$

Breiman et al (1984) suggest maximising $\Delta(s,t)$, over all questions for all features at a node, for the Gini index of node impurity given by

$$imp(t) = \Sigma\Sigma_{i \neq m} p(i|t)p(m|t) \quad , \qquad (8.5)$$

or, when there are more than two classes, maximising, over all questions for all features at a node, the Twoing criterion

$$p_L p_R(4)^{-1}[\Sigma_i\{p(i|t_L)-p(i|t_R)\}]^2 \quad . \tag{8.6}$$

The latter criterion is derived from amalgamating classes so that considered as a two-class problem the biggest decrease in node impurity is obtained (Breiman et al, 1984, pages 104-108).

For a rule to stop splitting nodes Breiman et al (1984) propose growing a tree with too many nodes which will give a bigger estimated error rate, using the test set or cross-validation method, than a tree with fewer nodes. This tree is then pruned back using a minimal cost-complexity pruning algorithm. This algorithm obtains a nested finite sequence of subtrees with progressively fewer terminal nodes. This sequence is obtained by minimising

$$R_\alpha(T) = R(T) + \alpha|T| \tag{8.7}$$

for each α as α is increased from zero, where R(T) is the resubstitution error rate estimate obtained by allocating the objects in the training set. $|T|$ is the number of nodes in the tree and α is positive. The test set or cross-validation method is then used to pick among the decreasing subsequence of trees that with the minimum estimated error-rate (within certain bounds of variability of the estimate).

The class allocation rule for each terminal node used by Breiman et al (1984), when the cost of misallocating a class m object as a class i object is M_{mi} , is to select the class i to minimise

$$\Sigma_m M_{mi} p(m|t) \quad . \tag{8.8}$$

The classification tree method is usually used only for splits based on single features, but Breiman et al (1984) also consider the inclusion of linear combinations of quantitative features and Boolean combinations of discrete features (Breiman et al, 1984, chapter 5). The increased complexity of the

optimisation process, however, means that optimal splits cannot be guaranteed.

8.3 Nearest neighbour discrimination.

The nearest neighbour discrimination method proceeds by allocating a new object to the class which maximises (Hand, 1981b, page 32)

$$P_i k_i n_i^{-1} \tag{8.9}$$

where P_i is the prior probability of class i, k_i is the number of objects from class i amongst the k nearest neighbours in a training data set using some measure of distance between objects and n_i is the number of objects from class i in a training data set. This rule is obtained from taking as an estimate of $f_i(\underline{x})$, the probability density function for class i ,

$$k_i n_i^{-1} \{A(k,\underline{x})\}^{-1} \tag{8.10}$$

where $A(k,\underline{x})$ is the volume of a hypersphere which just encloses the k nearest points in the training set to a vector, $\underline{x}$, of feature values.

8.4 Kernel density discrimination.

The kernel density estimator of a class probability density function may basically be considered to be a sum of bumps placed at the observations. The kernel function determines the shape of the bumps and a window width determines their width. The multivariate kernel density estimator for the ith class with kernel function K and window width h_i is defined by

$$\hat{f}_i(\underline{x}) = n_i^{-1} h_i^{-p} \Sigma_j K\{h_i^{-1}(\underline{x} - \underline{x}_{ij})\} \tag{8.11}$$

for p dimensions where $\underline{x}$ is a vector of feature values, n_i is the number of

observations for class i and $\underline{x}_{ij}$ is the vector of feature values for the jth object in class i . The kernel function for p-dimensional $\underline{x}$ satisfies

$$\int_{R^p} K(\underline{x})\, d\underline{x} = 1 \quad .$$

For discriminant analysis each $\hat{f}_i(\underline{x})$ is used in the usual discriminant rules (Silverman, 1986, pages 121-122).

8.5 Logistic discrimination.

Logistic discrimination may be considered to be a semi-parametric approach to discrimination in that it makes a parametric assumption about the likelihood ratios but not about the probability distribution for each class.

For c classes the likelihood ratios are assumed to satisfy

$$L_{ic}(\underline{x}) = \exp(\alpha_{0i} + \underline{\alpha}_i^T \underline{x}) \quad (i=1,\ldots,c-1) \tag{8.12}$$

where α_{0i} and $\underline{\alpha}_i$ are parameters to be estimated and $\underline{x}$ is a vector of feature values for an object from one of the c populations. This form is implied by an assumption of multivariate Normal distributions with equal covariance matrices and a number of other commonly occurring models (Anderson, 1972). Allocation of an object is then to the class maximising the set of linear logistic discrimination functions

$$l_i(\underline{x}) = \alpha_{0i} + \ln(P_i P_c^{-1}) + \underline{\alpha}_i^T \underline{x} \quad (i=1,\ldots,c) \quad , \tag{8.13}$$

where α_{0c} and the elements of $\underline{\alpha}_c$ are zero and P_i is the prior probability of class i . If α_{oi}^* denotes $\alpha_{0i} + \ln(P_i P_c^{-1})$ and $\underline{x}_{ij}$ is the vector of feature values for the jth object in the ith class then Anderson (1972) has shown that maximum-likelihood estimates of α_{0i}^* and $\underline{\alpha}_i$ (i=1,...,c-1) are obtained by

maximising, with respect to the parameters,

$$\prod_i \prod_j \exp(\alpha_{0i}^{*} + \underline{\alpha}^T \underline{x}_{ij}) [\Sigma_s \exp(\alpha_{0s}^{*} + \underline{\alpha}^T{}_s \underline{x}_{sj})]^{-1} \quad . \tag{8.14}$$

For separate sampling of the c populations the intercepts require the additional adjustment (Albert and Lesaffre, 1986)

$$\hat{\alpha}_{0i} = \hat{\alpha}_{0i}^{*} + \ln(n_c n_i^{-1}) \tag{8.15}$$

where n_i is the number of objects in class i in a training set. The maximisation of (8.14) may be done by the standard Newton-Raphson optimisation procedure.

8.6 Application to five human chromosome data sets.

All of the available features containing no exact linear dependencies were used for the Edinburgh, Copenhagen and Philadelphia data sets except for the classification tree results for the Copenhagen data. Only the first 16 features given by the MSEPCOR feature selection method as described in chapter 5 were used with this method for this data set. This was because of the limited memory capability of the program CART. For the special Copenhagen data sets, the eleven features used were those included in the WDD classifier described by Lundsteen, Gerdes and Maahr (1986) and specified in chapter 3.

The test-set method of error rate estimation was used for all the methods because of the computational time that would be required to use the leave-one-out method. The division of the data sets into two was that described in chapter 3.

For all the data sets the normalisation of measurements for between-cell variation was that currently used and described in chapter 3.

Prior probabilities of 2/46 for chromosome classes 1-22 and 1/46 for chromosome classes 23 and 24 were used for the Edinburgh data set which

has cells only from males. The prior probabilities for classes 23 and 24 were changed to 3/92 and 1/92 for all the other data sets which have cells from both sexes.

The overall estimated percentage error-rate was taken as the weighted average of the individual class percentage error-rates using the specified prior probabilities as the weights.

No re-allocation of chromosomes to satisfy a normal karyotype as described in chapter 2 was performed.

8.6.1 Classification trees.

The program CART (California Statistical Software Inc., 1985) was used to obtain results for the five data sets in two ways. One way excluded the linear combination algorithm and the other used this algorithm with different minimum node sizes for its use. Full details of the linear combination algorithm for quantitative features are given in the appendix to chapter 5 of Breiman et al (1984). A constant for the linear combination algorithm is specified so that not necessarily all features are used in a linear combination. Specifically, for the coefficients for a best linear split found by the algorithm using all the features, each feature is omitted in turn to find the most important (i.e., leading to the smallest decrease in impurity) and least important (i.e., leading to the biggest decrease in impurity) features. This is done by finding the best split of the form

$$\Sigma_{j \neq j'} a_j x_j \leqq c_{j'} \qquad (8.16)$$

where j' is the excluded feature, the a_j are the coefficients for the best linear split using all the features, x_j is the value for the jth feature and $c_{j'}$ is a threshold which is optimised. Defining Δ^* as the reduction in impurity for using all the features in the linear combination, Δ_{min} as the reduction in impurity for the linear combination leaving out the most important feature and Δ_{max} as the reduction in impurity for the linear combination leaving out the least important feature then if

$$\Delta^* - \Delta_{max} < \beta(\Delta^* - \Delta_{min}) \qquad (8.17)$$

where β is a pre-specified constant, the least important feature is deleted. If the least important feature is deleted the most important and least important features are then found again in the same way using the coefficients of the linear combination for all features and excluding the deleted feature. Test (8.17) is then applied again. The process is repeated until no more features are deleted. Finally, the search algorithm is used again to find the best linear split for the features not deleted. The splitting criterion and parameter settings are reproduced in Table 8.1 . The value of 0.2 used for β is the default value in CART (California State Software Inc., 1985). The subsampling option refers to taking a subsample at a node instead of all the data to find the best split for that node. This is done to avoid excessive storage requirements. The tree chosen by the test-set method when 'pruning' was that with the minimum estimated error-rate.

8.6.2 Nearest neighbour discrimination.

Four versions of the estimated Mahalanobis' distance between the feature vector for a new object, $\underline{x}$, and the feature vector for the jth object from class i, $\underline{x}_{ij}$,

$$(\underline{x} - \underline{x}_{ij})^T \hat{\underline{\Sigma}}_i^{-1} (\underline{x} - \underline{x}_{ij}) , \qquad (8.18)$$

were used. The four versions were obtained by letting $\hat{\underline{\Sigma}}_i$ be:

1. Unrelated diagonal.
2. Common for all classes.
3. Common within Denver groups.
4. Unrelated non-diagonal for each class.

The estimators defined in chapter 5 were used. To examine the effect of altering the number of nearest neighbours, results were obtained for 1, 5 and 10 nearest neighbours.

Table 8.1

Parameter settings for CART program.

Gini splitting rule
minimum size of node to continue splitting = 5
maximum size of node without subsampling = 1000
minimum size of node for linear combination option = 20, 50 or 100
constant for deletion of features for linear combination option = 0.2

8.6.3 Kernel density discrimination.

Because of the importance of areas of low density in multivariate density estimation, the adaptive kernel method was used (Silverman, 1986, pages 100–110). This procedure allows the window width to vary so that in areas of low density a broader kernel is used. The initial estimate of $\hat{f}_i(\underline{x})$ was obtained using the Normal kernel in the density estimate

$$\hat{f}_i(\underline{x}) = \det(\hat{\underline{\Sigma}}_i)^{-\frac{1}{2}} n_i^{-1} h_i^{-p} \Sigma_j k[(h_i^{-2}(\underline{x} - \underline{x}_{ij})^T \hat{\underline{\Sigma}}_i^{-1}(\underline{x} - \underline{x}_{ij})] \,, \tag{8.19}$$

where h_i is the smoothing parameter which, using a Normal kernel, minimises the mean integrated square error for a standard Normal distribution (Silverman, 1986, page 87), i.e.,

$$\{(2p+1)n_i/4\}^{-1/(p+4)} \tag{8.20}$$

and $k(\underline{x}^T\underline{x}) = K(\underline{x})$. The multivariate Normal kernel is defined as

$$K(\underline{x}) = (2\pi)^{-\frac{1}{2}p} |\Sigma|^{-\frac{1}{2}} \exp(-\tfrac{1}{2}\underline{x}^T \underline{\Sigma}^{-1}\underline{x}) \quad . \tag{8.21}$$

The same four estimators of $\hat{\underline{\Sigma}}_i$ were used as for the nearest neighbour analysis described above. The density estimate at the second stage was as above with h_i replaced by $h_i\lambda_{ij}$, where

$$\lambda_{ij} = \{\hat{f}_i(\underline{x}_{ij})g^{-1}\}^{-\alpha} \tag{8.22}$$

and

$$\ln(g) = n_i^{-1} \Sigma_j \ln\{\hat{f}_i(\underline{x}_{ij})\} \quad . \tag{8.23}$$

α is a sensitivity parameter. Following Abramson (1982) this parameter was set equal to $\frac{1}{2}$.

The use of the smoothing parameter for a standard multivariate Normal distribution at the first stage may be justified on the grounds that adaptive kernel density estimates have been found not to be sensitive to fine detail of the initial estimate (Breiman, Meisel and Purcell, 1977) and also multivariate Normal methods have been found to give good results for these data sets.

8.6.4 Logistic discrimination

The computer program of Albert and Harris (1987) was used to obtain results. This program was used in case partial separation (Lesaffre and Albert, 1989), geometrically defined as complete separation of clusters of classes where not all clusters contain one class, was detected in the training data set. If partial separation occurs maximum-likelihood estimates of the parameters in (8.12) do not exist. However, defining (8.12) to hold within each cluster containing more than one class then maximum likelihood estimates do exist for each cluster. The parameter estimates for each of these clusters can be derived from the estimates of the parameters when (8.12) is assumed to hold across all classes. The computer program of Albert and Harris (1987) detects the divergence in parameter estimates caused by partial separation, stops the iterative solution and prints out the allocation matrix for objects in the training set. This allocation matrix can be used to identify the precise form of the partial separation.

8.7 Results.

8.7.1 Classification trees.

The estimated percentage error-rates for the classification trees for the five data sets are given in Table 8.2 .

8.7.2 Nearest neighbour discrimination.

The results for nearest neighbour discrimination using the four measures of distance and three numbers of nearest neighbours for the five data sets are given in Table 8.3 .

Table 8.2

Estimated percentage error-rates for classification trees. (Result for classification tree without linear combination option followed by result for classification tree with linear combination option for minimum node sizes of 20, 50 and 100 .)

Edinburgh data set

28.6,23.1,23.1,23.7

Copenhagen data set

12.5,10.8,10.9,10.8

Philadelphia data set

31.7,26.2,26.0,25.9

Copenhagen special amniotic-fluid data set

13.6,10.8,11.0,11.0

Copenhagen special peripheral-blood data set

17.8,14.3,14.5,14.6

Table 8.3

Estimated percentage error-rates for nearest neighbour discrimination procedures.
(Results in order are for 1, 5 and 10 nearest neighbours.)

Edinburgh data set

Estimate of covariance matrix	Estimated percentage error-rate
unrelated diagonal	26.5, 28.7, 28.7
common	32.1, 36.8, 36.9
unrelated	21.4, 22.1, 22.2
common within Denver groups	19.3, 21.3, 21.4

Copenhagen data set

Estimate of covariance matrix	Estimated percentage error-rate
unrelated diagonal	5.7, 6.2, 6.2
common	10.7, 11.3, 11.3
unrelated	6.6, 7.0, 6.7
common within Denver groups	5.2, 6.4, 6.4

Philadelphia data set

Estimate of covariance matrix	Estimated percentage error-rate
unrelated diagonal	27.0, 30.6, 30.6
common	38.7, 43.7, 43.8
unrelated	27.4, 28.2, 28.2
common within Denver groups	24.4, 28.3, 28.4

Copenhagen special amniotic-fluid data set

Estimate of covariance matrix	Estimated percentage error-rate
unrelated diagonal	8.1, 10.4, 10.4
common	16.1, 21.2, 21.2
unrelated	9.5, 12.3, 12.3
common within Denver groups	8.9, 12.0, 12.0

Copenhagen special peripheral-blood data set

Estimate of covariance matrix	Estimated percentage error-rate
unrelated diagonal	16.2, 18.2, 18.3
common	25.5, 30.3, 30.4
unrelated	12.3, 13.4, 14.2
common within Denver groups	11.4, 14.3, 14.3

8.7.3 Kernel density discrimination.

The results for kernel density discrimination using the four density estimates given by (8.19) (with h_i replaced by $h_i\lambda_{ij}$) and the different estimates of $\hat{\underline{\Sigma}}_i$ for the five data sets are given in Table 8.4 .

8.7.4 Logistic discrimination.

The program of Albert and Harris (1987) was used to obtain results using just the first 3 and 6 features chosen by the MSEPCOR feature selection procedure for the Edinburgh data set. A greater number of features was not tried because of the extremely large amount of computer c.p.u. time required to estimate parameters for even these small numbers of features. Partial separation was not detected in the training data set for these features. The estimated percentage error-rates and c.p.u. time required on the Edinburgh University Castle main-frame computer are given in Table 8.5 . Also given in Table 8.5 are the corresponding results for the procedures which assume multivariate Normality and are described in chapter 5.

8.8 Discussion.

The results for the classification trees are worse than those for the procedures which assume multivariate Normality, confirming the result of Shepherd, Piper and Rutovitz (1987). The classification trees used here differ from that used by Shepherd, Piper and Rutovitz (1987) in that:

1. A different measure of impurity is used.
2. An additional linear combination algorithm is optionally used.
3. The minimal cost-complexity algorithm is used to prune too large a tree rather than growing a tree until the estimated error-rate of a test data set increases.
4. Prior probabilities for the twenty-four classes are used.

Allocation time for the chromosomes in a normal cell is not considered here because the estimated percentage error-rates are much worse than the procedures described in chapter 5. It should also be noted that the test data sets were used to 'prune' back the trees as well as to estimate the error rates. This means that these test data sets were not truly independent of the data used to 'grow' the trees.

The results for the nearest neighbour and kernel density discrimination

Table 8.4

Estimated percentage error-rates for kernel density discrimination procedures.

Edinburgh data set

Estimate of covariance matrix	Estimated percentage error-rate
unrelated diagonal	16.9
common	17.5
unrelated	20.2
common within Denver groups	18.2

Copenhagen data set

Estimate of covariance matrix	Estimated percentage error-rate
unrelated diagonal	9.2
common	4.8
unrelated	9.6
common within Denver groups	7.7

Philadelphia data set

Estimate of covariance matrix	Estimated percentage error-rate
unrelated diagonal	27.7
common	22.7
unrelated	27.6
common within Denver groups	23.4

Copenhagen special amniotic-fluid data set

Estimate of covariance matrix	Estimated percentage error-rate
unrelated diagonal	10.7
common	8.0
unrelated	8.1
common within Denver groups	8.6

Copenhagen special peripheral-blood data set

Estimate of covariance matrix	Estimated percentage error-rate
unrelated diagonal	11.7
common	9.3
unrelated	11.6
common within Denver groups	10.9

Table 8.5

Estimated percentage error-rates and c.p.u. time on Edinburgh University main-frame Castle computer for logistic discrimination.

Edinburgh data set

Number of features	Estimated percentage error-rate	c.p.u. time seconds
3	36.3	9490
6	24.8	21014

Corresponding results for procedures based on multivariate Normality described in chapter 5.
(Estimated percentage error-rate for 3 features followed by estimated percentage error-rate for 6 features.)

Procedure	Percentage error-rates
C	38.5, 27.7
BC	38.2, 26.1
U	34.8, 22.3
BU	30.5, 22.3
UD	35.7, 23.3
GC	36.9, 23.3
BGC	43.5, 26.3
P	37.8, 24.8
BP	37.3, 25.4
GP	36.4, 23.9
PG	37.8, 25.6
PD	43.3, 35.4
E	35.7, 23.2

procedures are no better than the results for the methods which assume multivariate Normality. Because of this no attempt has been made here to consider allocation time. If practical application were to be attempted the data sets would have to be condensed (Hand, 1981b, pages 30–31) and/or fast algorithms (Friedman, Bentley and Finkel, 1977) used to find the nearest neighbours of a new feature vector or all the neighbours within a certain distance (for a kernel of finite support). Consideration of the results given here together with those of chapter 5 suggest that it is unlikely that these non-parametric procedures will provide candidate combinations of estimated percentage error-rate and allocation time for the number of features used here. This is because (except for procedure C and the result for one data set for procedure UD) the Estimative multivariate Normal procedures which use corresponding definitions of Mahalanobis distance to the four procedures defined above for nearest neighbour and kernel density discrimination give lower estimated percentage error-rates than these non-parametric procedures. This means that the non-parametric procedures with the same definitions of Mahalanobis distance as procedures UD, U and GC would in general require less computation than each of these corresponding procedures to provide possible candidate combinations of estimated percentage error-rate and allocation time. The kernel density procedure which uses the estimate of a common covariance matrix for all classes gives results no better than procedure GC. Therefore, this procedure would require less computation than procedure GC to give possible candidate combinations of estimated percentage error-rate and allocation time. For all the nearest neighbour and kernel density procedures, this means that only a small proportion of the training sets considered here could be used to allocate a new object if a possible candidate combination of estimated percentage error-rate and allocation time were to result. The same conclusion would apply if a kernel of finite support had been used and given similar results.

The large amount of computer c.p.u. time required to estimate the parameters in (8.12) for even a very small number of features appears to rule the logistic discrimination method out of practical consideration. Whilst the time taken to estimate parameters for a training data set is not usually a consideration in the automated allocation of human chromosomes, such an excessive amount of computer time seems undesirable. For the small number of features used here there is no evidence of better performance than some of

the procedures which assume multivariate Normality.

Chapter 9
Modelling the probabilities of band-transition sequences for the automated allocation of human chromosomes.

9.1 Introduction.

As outlined in chapter 2, so-called band-transition sequences have been proposed as an alternative to the use of weighted sums of density profiles for use in the automated allocation of human chromosomes (Lundsteen et al, 1981). In this chapter the former method is briefly outlined and some non-parametric and parametric models proposed for describing dependence between values of the same feature measured on successive chromosome segments starting from one end of the chromosome and between the values of two features measured for each segment. These models are applied to three human chromosome data sets.

9.2 Band-transition sequences.

As described in chapter 2, Lundsteen and Granum (1979) have suggested the division of a chromosome into 13 segments ($5\frac{1}{2}$ for the short arm of a chromosome and $7\frac{1}{2}$ for the long arm). For each of these segments, if there is a peak density of staining in the segment then its density is recorded together with the difference in density of staining between the peak and its adjacent valley (light band) proceeding from the short arm of the chromosome to the long arm. Peak density is measured on an integer scale of 0-6 and density difference on an integer scale of 0-4 with a 0 being recorded for both features if there is no peak density of staining in a segment. This gives 26 feature values for each chromosome. In addition, a pair of features is used to describe the starting point which is taken as the beginning of the short arm. These features give the value of the first peak and the average value of the "valley" preceding it because there is no true preceding valley. A similar artificial "valley" is considered to follow the final peak. All 28 feature values may be referred to as a band-transition sequence.

9.3 Non-parametric models for the probabilities of band-transition sequences.

Lundsteen et al (1981) have considered the following complete

independence model for the probabilities of the band-transition sequences

$$p(c_i|\underline{x}) \propto \prod_k \prod_l p(x_{kl}|c_i).P_i \qquad (9.1)$$

where $p(c_i|\underline{x})$ is the posterior probability of class i given the band transition values, x_{kl} is the value for the lth feature measured on the kth segment, $p(x_{kl}|c_i)$ is the class-conditional probability for class i and P_i is the prior probability of class i . The outer product is from k = 1 to k = 14 and the inner product is from l = 1 to l = 2. This model takes no account of the orderings of the segments, of dependence between the values of features measured on successive segments or of dependence between the values for the two features measured for each segment.

A model which takes into account the dependence between the value of a feature for a segment and the value for the same feature measured on the previous segment, if there is a previous segment, is

$$p(c_i|\underline{x}) \propto \prod_l p(x_{1l}|c_i).\prod_k p(x_{k+1,l}|c_i\ ,\ x_{k,l}).P_i\ . \qquad (9.2)$$

The outer product is from l = 1 to l = 2 and the inner product is from k = 1 to k = 13. A model which takes account of the bivariate distribution for the values for each segment but ignores dependence between values of the same or different features measured on successive segments is

$$p(c_i|\underline{x}) \propto \prod_k p(\underline{x}_k|c_i).P_i\ , \qquad (9.3)$$

where $\underline{x}_k$ is a 2 vector of feature values for the kth segment and the product is from k = 1 to k = 14. These two models may be combined to give one which takes account of the bivariate distribution for each segment and the dependence between values of the same feature measured on immediately successive segments,

$$p(c_i|\underline{x}) \propto p(\underline{x}_1|c_i).\prod_k p(\underline{x}_{k+1}|c_i\ ,\ \underline{x}_k).P_i\ , \qquad (9.4)$$

where the product is from k = 1 to k = 13. Further models which take into account dependence between the value of a feature for a segment and the values for the same feature for the r_k previous segements may be defined in a similar manner, where r_k = min(k −1, r) and 1 < r < p. However, even with the size of the data sets considered here, sparsity of data rapidly becomes a problem in estimating some of the class-conditional probabilities. The only other model which is considered further below is that which assumes the probability of observing a value for a feature on a segment is dependent on the values for the same feature for the r_k preceeding segments with r = 2 but takes the values for the two features measured for a segment or different features measured on different segments to be independent, i.e.,

$$p(c_i|\underline{x}) \propto \prod_l p(x_{1l}|c_i).p(x_{2l}|c_i\ ,\ x_{1l}).\prod_k p(x_{k+2,l}|c_i,\ x_{k+1,l}\ ,\ x_{kl}).P_i\ . \qquad (9.5)$$

Here the outer product is from l = 1 to l = 2 and the inner product is from k = 1 to k = 12 .

Estimates of the class-conditional probabilities in equations (9.1), (9.2), (9.3), (9.4) and (9.5) can be obtained by using the proportion of observations on chromosomes of class i with the particular values in a training set of data.

9.4 Multivariate Normal models for band-transition sequences.

Given the reductions in percentage error-rates obtained in chapter 5 by combining class information on variability for the features considered there, the same procedures may be considered here for use with the band-transition sequence data. All of the assumptions about the covariance matrices for the 24 classes made in chapter 5 can be tried for the band-transition sequence data. Additionally, the dependence between the values for the features measured in successive segments may be modelled by using so-called

ante-dependence models (Kenward, 1987). These models form a subset of the class of covariance selection models considered in chapter 6 . As in chapters 5 and 6 , these models are considered for application because of the reduction in the number of parameters and hence the sampling variability of the predicted distribution.

An ante-dependence model of order r is one where the kth feature (k > r) of a set of p ordered features, given the preceding r, is independent of all further preceding features. For this structure the inverse of the covariance matrix has zeroes everywhere except on the leading diagonal and the r diagonals above and below. The probability density function can be expressed as the product of p components, i.e., as

$$\prod_k f(x_k|x_{k-r_k},\ldots,x_{k-1}) \tag{9.6}$$

where the product is from k = 1 to k = p and r_k = min (k - 1, r) (Kenward, 1987). Consequently, a discriminant score may be calculated as the sum of p components when logs are taken, prior probabilities are ignored and all misallocation costs are equal.

For the band-transition sequence data we consider, as above for the non-parametric models, that the sequence of pairs of features 'starts' at the end of the short arm of the chromosome and that we have a bivariate distribution for each segment; thus, under an ante-dependence model of order r , we have the product of the 14 bivariate components for each class (dependence on i is suppressed for the remainder of this section)

$$\prod_k f(\underline{x}_k|\underline{x}_{k-r_k},\ldots,\underline{x}_{k-1}) \ . \tag{9.7}$$

For each of these 14 components the conditional mean vector of $\underline{x}_k$ given $\underline{x}_{k-r_k},\ldots,\underline{x}_{k-1}$ is given by

$$\underline{\mu}_k + \underline{\Sigma}_{kr}\,\underline{\Sigma}_{r_k r_k}^{-1}(\underline{x}_{r_k} - \underline{\mu}_{r_k}) \qquad (9.8)$$

where $\underline{\Sigma}_{kr}$ is a 2 by $2r_k$ matrix of covariances between $\underline{x}_k$ and $\underline{x}_{k-r_k},\ldots,\underline{x}_{k-1}$, $\underline{\Sigma}_{r_k r_k}$ is a $2r_k$ by $2r_k$ covariance matrix for the $\underline{x}_{k-r_k},\ldots,\underline{x}_{k-1}$, $\underline{x}_r$ is a $2r_k$ vector of observed feature values and $\underline{\mu}_r$ is a $2r_k$ vector of means (Morrison, 1976, page 92). The conditional covariance matrix is given by

$$\underline{\Sigma}_{kk} - \underline{\Sigma}_{kr_k}\underline{\Sigma}_{r_k r_k}^{-1}\underline{\Sigma}_{kr_k}^{T}\ , \qquad (9.9)$$

where $\underline{\Sigma}_{kk}$ is the covariance matrix for $\underline{x}_k$ (Morrison, 1976, page 92). A discriminant score using just the band-transition data for a given class i may therefore be calculated as

$$\Sigma_k\{-\ln|\underline{S}_{kc}| - (\underline{x}_k - \bar{\underline{x}}_{kc})^T\underline{S}_{kc}^{-1}(\underline{x}_k - \bar{\underline{x}}_{kc})\} + 2\ln(P_i)\ , \qquad (9.10)$$

where the $\bar{\underline{x}}_{kc}$ and $\underline{S}_{kc}$ are estimates of the conditional mean vector and covariance matrix given above for the kth pair of features obtained by using the usual unbiased estimates of $\underline{\mu}_k$, $\underline{\Sigma}_{kr_k}$, $\underline{\Sigma}_{kk}$ and $\underline{\Sigma}_{r_k r_k}$.

The number of parameters to be estimated for each covariance matrix for an ante-dependence structure of order r is, for p bivariate features, $3p + 3(p-1) + \ldots + 3(p-r)$. The number of calculations to calculate one discriminant score using (9.10) is $5p + \Sigma_k 4r_k$ multiplications and $5p + \Sigma_k 2r_k + \Sigma_k 4r_k + 3p + 2(p-1) + 1$ additions and subtractions. As in chapter 5, the Cholesky decomposition of the estimated inverse of each conditional covariance matrix is used and it is assumed that the number of additions required to evaluate

$$\underline{z}^T\underline{L}\underline{L}^T\underline{z}\ ,$$

where $\underline{L}$ is a lower-triangular matrix, is equal to the number of multiplications

required. These numbers of calculations compare with $p(p + 1) + 2p$ multiplications and $p(p + 1) + 4p + 1$ additions and subtractions for procedure U (the estimation of unrelated covariance matrices for each chromosome class) in chapter 5.

9.5 Application to reduced Copenhagen and special Copenhagen data sets.

The reduced Copenhagen data set and special Copenhagen data sets described in chapter 3 were used to obtain results for the models described above.

Percentage error-rates were estimated by the test-set and leave-one-out methods described in chapter 5 except for procedure E (the estimation of common principal components) from chapter 5 for which the leave-one-out method was found to require an excessive amount of computational time. The overall estimated percentage error-rate was calculated as the weighted sum of the individual class percentage error-rates with the weights given by the prior probabilities for each class. The prior probabilities used were those corresponding to equal numbers of cells from males and females. No re-arrangement of the allocations within a cell to satisfy a normal karyotype was performed. The splits of the data into two were those described in chapter 3. The normalisation of the one size feature used in conjunction with the band-transition features was that described in chapter 3 .

9.5.1 Non-parametric models for the probabilities of band-transition sequences.

All five of the non-parametric models for the band-transition sequences described above were used for these data sets. Following Lundsteen et al (1981), in order to avoid zero estimated class-conditional probabilities and to stop underflow on the computer the class-conditional probabilities were multiplied by 100, rounded to the nearest integer and any zeroes replaced by ones. The replacement of zeroes by ones gives lower error-rates. Again following Lundsteen et al (1981), area, area centromeric index and density centromeric index were included in the model by assuming them to be independent with Normal distributions. The posterior probability (except for a proportionality factor) for a band-transition sequence was multiplied by the values for the univariate Normal probability density functions (except for proportionality factors). The univariate Normal p.d.f.s were weighted more

heavily by raising the values for them to the powers 3, 2 and 2 for area, area centromeric index and density centromeric index respectively because this was found by Lundsteen et al (1981) to give smaller estimated error-rates.

9.5.2 Ante-dependence models.

Ante-dependence models of orders 0 to 5 were fitted to the data for each chromosome class. The ante-dependence model of order 0 corresponds to independent bivariate Normal distributions for the features for each segment. Area, area centromeric index and density centromeric index were included in the discrimination procedure by assuming them independent of the band-transition sequences and having independent univariate Normal distributions for comparability with the results for the non-parametric models.

9.5.3 Chapter 5 procedures.

All of the assumptions about class covariance matrices proposed in chapter 5 were also tried for the band-transition sequences together with area, area centromeric index and density centromeric index.

9.5.4 Ante-dependence models and chapter 5 procedures.

Exact linear dependence between features for some classes occurs in the data sets. This is because for some classes no peak density of staining is observed in some segments. Consequently, only zeroes are recorded for the two features in these segments. Because of this exact linear dependence between features for some classes a common small constant was added to the diagonal of the covariance matrix of the band-transition features for each class for the ante-dependence models and the related covariance matrix models of chapter 5. This was done so that the covariance matrix was invertible. The common small constant to be added was determined by a grid search with three internal points using the reduced size Copenhagen data set and test-set error-rate estimation. An initial grid search evaluated estimated percentage error-rate at a number of values for the constant in order to determine an interval containing a local minimiser. The value of the constant chosen was that which corresponded to the middle point when three points of the grid search agreed to one decimal place for the estimated percentage error-rate (the other two points giving bigger estimated percentage error-rates). This process assumes that only a global minimum exists or that any local minima are close to the global minimum, and that agreement to one decimal place of three points of the grid search is a reasonable stopping criterion. The constant

estimated for each procedure was then used in the leave-one-out method for the reduced Copenhagen data set and for both methods for the two special Copenhagen data sets.

9.6 Results

9.6.1 Non-parametric models for the probabilities of band-transition sequences.

Test-set and leave-one-out results for the non-parametric models using the band-transition sequences with zero estimated class-conditional probabilities (when rounded to 2 decimal places) replaced by the value 0.01 are given in Table 9.1 . For comparison, results are given in Table 9.2 for the same models when zero estimated class-conditional probabilities (when rounded to 2 decimal places) were not replaced by the value 0.01 . If all the discriminant scores calculated for a chromosome were equal to zero the chromosome was allocated to a class at random.

9.6.2 Ante-dependence models for band-transition sequences.

Table 9.3 contains the estimated percentage error-rates using the test-set method for the reduced Copenhagen data set; an initial coarse grid of values was used for the constant added to the diagonal of the covariance matrix for each class. The blanks correspond to values which give non-invertible covariance matrices in the procedures when NAG routine F01ACF (Numerical Algorithms Group Limited, 1988) for matrix inversion was used. Table 9.4 gives the values of the constant found by the grid search described in sub-section 9.5.4 and the estimated percentage error-rate for the use of the constant with the leave-one-out method. Table 9.4 also gives the estimated percentage error-rates for the use of these constants for the special Copenhagen data sets. Again blanks occur in this table when covariance matrices were not invertible for some procedures.

9.6.3 Chapter 5 procedures.

Table 9.4 gives the estimated percentage error-rates for the three data sets for the values of the constants added to the diagonal of each covariance matrix.

9.6.4 Chapter 5 procedures applied to WDD features.

For comparison, the results for the chapter 5 procedures applied to the WDD classifier features (defined in chapter 3) are given for the reduced

Table 9.1

Estimated percentage error-rates for the non-parametric models for probabilities of band-transition sequences.
(Zero estimated probabilities for class-conditional probabilities in models when estimated probabilities rounded to 2 decimal places replaced by the value 0.01 . Estimated posterior probabilities for band-transition sequences multiplied by values for univariate Normal p.d.f.s for area, area centromeric index and density centromeric index raised to powers given in text. Result for leave-one-out method followed by result for test-set method.)

Equation in text for model	Data set		
	Reduced Copenhagen	Copenhagen special amniotic fluid	Copenhagen special peripheral blood
(9.1)	5.2,5.6	10.4,11.0	12.8,13.5
(9.2)	4.5,4.9	10.0,10.4	12.9,13.6
(9.3)	5.7,6.8	11.2,11.5	13.3,13.6
(9.4)	5.3,6.0	10.0,10.8	12.2,12.4
(9.5)	5.3,5.6	10.7,11.0	13.6,14.1

Table 9.2

Estimated percentage error-rates for the non-parametric models for probabilities of band-transition sequences.
(Zero estimated probabilities for class-conditional probabilities in models when estimated probabilities rounded to 2 decimal places not replaced by the value 0.01 . Estimated posterior probabilities for band-transition sequences multiplied by values for univariate Normal p.d.f.s for area, area centromeric index and density centromeric index raised to powers given in text. Result for leave-one-out method followed by result for test-set method.)

Equation in text for model	Data set: Reduced Copenhagen	Copenhagen special amniotic fluid	Copenhagen special peripheral blood
(9.1)	9.7,11.1	13.8,15.7	16.4,19.1
(9.2)	21.5,24.9	24.8,27.8	25.8,30.9
(9.3)	15.9,18.5	19.0,21.1	20.6,24.7
(9.4)	35.3,37.2	33.8,36.6	33.8,39.5
(9.5)	36.4,39.0	39.0,41.9	41.3,47.1

Table 9.3

Estimated percentage error-rates for differing values of constant added to the diagonal of the covariance matrix for each chromosome class for the reduced Copenhagen data set and the test-set method.

Procedure	Value of constant							
	10	5	2	1	0.1	0.01	10^{-5}	10^{-9}
C	13.4	13.0	12.1	11.8	11.6	11.7	11.7	11.7
BC	10.2	9.8	8.8	8.5	8.2	8.3	8.4	8.4
U	7.7	6.6	5.4	5.3	6.8	8.8	12.3	17.5
BU	9.1	7.6	6.0	5.8	6.9	8.9	13.6	21.4
UD	9.9	9.7	9.8	9.8	12.2	15.0	18.5	29.3
GC	9.3	8.0	6.9	6.6	7.1	8.3	9.2	9.9
BGC	9.3	7.9	7.0	6.6	7.0	8.2	9.0	10.0
P	11.1	10.5	9.9	9.9	10.0	–	–	–
BP	14.1	13.4	12.4	12.1	–	–	–	–
GP	9.6	8.5	7.6	7.5	–	–	–	–
PG	10.2	9.8	8.8	8.7	–	–	–	–
PD	11.4	13.3	14.6	16.3	16.4	18.7	19.5	19.7
E	9.9	9.2	9.3	9.5	11.7	14.8	18.0	32.2
AD0*	9.6	8.5	8.3	7.9	9.7	11.5	16.6	25.3
AD1*	8.8	7.1	6.4	6.1	7.6	8.9	13.0	22.2
AD2*	8.9	7.0	6.0	5.8	7.5	8.6	12.7	21.2
AD3*	8.8	7.1	6.0	5.9	7.2	8.9	12.9	19.0
AD4*	8.7	7.1	5.9	5.8	7.2	9.2	12.9	19.0
AD5*	8.7	7.1	6.0	5.7	7.3	9.2	12.9	18.9

* ADr denotes ante-dependence model of order r

Table 9.4

Estimated percentage error-rates for constant found by grid search for the reduced Copenhagen data set and the test-set method added to diagonal of covariance matrix for each chromosome class. (Result for leave-one-out method followed by result for test-set method.)

Procedure	Constant	Data set: Reduced Copenhagen amniotic fluid	Data set: Copenhagen special peripheral blood	Data set: Copenhagen special
C	0.190	8.4,11.6	15.6,17.9	18.3,21.2
BC	0.275	8.5, 8.2	15.9,16.1	18.4,19.0
U	1.760	4.9, 5.2	11.4,11.7	13.6,14.4
BU	0.775	4.9, 5.6	11.4,12.1	13.4,14.6
UD	3.970	9.3, 9.6	14.9,15.4	18.1,18.2
GC	0.540	6.2, 6.6	13.4,13.6	14.8,15.6
BGC	0.775	6.1, 6.6	13.6,13.7	15.0,15.7
P	1.250	9.3, 9.9	16.8,16.9	20.0,21.1
BP	0.550	10.9,11.6	_, _	_, _
GP	1.135	6.6, 7.4	_, _	_, _
PG	1.050	8.4, 8.6	16.8,16.9	19.5,17.9
PD	9.850	12.0,11.5	18.8,18.6	23.6,23.4
E	4.300	_, 9.2	_,14.9	_,18.4
AD0*	0.890	7.8, 8.0	14.2,15.1	17.2,17.5
AD1*	1.390	5.7, 6.0	13.0,13.6	16.0,16.2
AD2*	1.380	5.6, 5.8	12.7,13.2	15.6,15.7
AD3*	1.325	5.5, 5.7	12.7,13.3	15.5,15.7
AD4*	1.380	5.5, 5.7	12.7,13.5	15.4,15.7
AD5*	1.255	5.5, 5.7	12.6,13.4	15.4,15.7

* ADr denotes ante-dependence model of order r

Copenhagen data set in Table 9.5 . The results for the same procedures for the WDD features for the other two data sets appear in Figures 9.3 to 9.6 .

9.7 Discussion.

For two of the three data sets, model (9.2) for the probabilities of the band-transition sequences gives the smallest estimated percentage error-rates for the band-transition sequences together with area, area centromeric index and density centromeric index. The differences from the results for the complete independence model for the probabilities of the band-transition sequences, model (9.1), however are small.

The results for the models which assume multivariate Normal distributions for the band-transition sequences are not as good as those for the non-parametric models. In some instances, however, the differences are small. Defining a peak density and valley of staining for each segment with density measured as a continuous feature might, however, be expected to give data which more closely follows the asumption of multivariate Normality. Such a definition would also avoid exact linear dependence amongst the band-transition features.

Table 9.1 when compared with Table 9.5 shows that the results for the use of the band-transition features plus area, area centromeric index and density centromeric index are not as good as some of the results for the chapter 5 procedures applied to the WDD features for the reduced Copenhagen data set. Comparing Table 9.1 with Figures 9.3 to 9.6 shows that the same is true for the special Copenhagen data sets. To see if this is because of the assumption of independence for the features area, area centromeric index and density centromeric index results were obtained for the assumption that these features have a trivariate Normal distribution independent of the band-transition sequence features for each class. The results are given in Tables 9.6 and 9.7 . For the non-parametric models for the probabilities of band-transition sequences a probability of a band-transition sequence was multiplied by the value of the trivariate Normal p.d.f. raised to the power of an estimated weight. This weight was found for each model for the band-transition sequence probabilities by a grid search for the reduced Copenhagen data set using test-set error-rate estimation. For the ante-dependence models a common constant to be added to the diagonal of each covariance matrix was also found

Table 9.5

Estimated percentage error-rates for WDD classifier features for reduced Copenhagen data set.
(Result for leave-one-out method followed by result for test-set method.)

Procedure	Estimated percentage error-rate
C	4.4, 8.9
BC	4.4, 5.0
U	2.5, 3.7
BU	2.5, 3.6
UD	4.9, 5.5
GC	2.9, 3.3
BGC	2.9, 3.3
P	4.8, 5.6
BP	5.2, 6.0
GP	3.0, 3.4
PG	4.9, 5.5
PD	6.3,13.8
E	4.4, 4.9

Table 9.6

Estimated percentage error-rates for the non-parametric models for probabilities of band-transition sequences.
(Zero estimated probabilities for class-conditional probabilities in models when estimated probabilities rounded to 2 decimal places replaced by the value 0.01 . Estimated posterior probabilities for band-transition sequences multiplied by value for trivariate p.d.f. for area, area centromeric index and density centromeric index raised to power given by value of weight below. Result for leave-one-out method followed by result for test-set method.)

Equation in text for model	Weight	Data set: Reduced Copenhagen	Copenhagen special amniotic fluid	Copenhagen special peripheral blood
(9.1)	2.125	4.7,5.3	10.0,10.3	12.6,12.9
(9.2)	2.750	4.3,4.7	9.5, 9.9	11.9,12.7
(9.3)	1.010	5.1,5.6	10.3,11.1	12.5,13.2
(9.4)	2.015	5.0,5.5	9.8,10.3	11.8,12.4
(9.5)	3.660	4.8,5.3	9.5, 9.8	12.0,12.6

Table 9.7

Estimated percentage error-rates for constant found by grid search for the reduced Copenhagen data set and the test-set method added to diagonal of covariance matrix for each chromosome class. (Result for leave-one-out method followed by result for test-set method.)

Procedure	Constant	Data set: Reduced Copenhagen amniotic fluid	Data set: Copenhagen special peripheral blood	Data set: Copenhagen special
AD0*	1.688	7.3, 7.8	13.7,14.3	16.5,16.8
AD1*	1.500	5.5, 5.7	12.5,12.9	15.4,15.6
AD2*	1.370	5.2, 5.5	12.3,12.9	14.9,15.5
AD3*	1.325	5.2, 5.6	12.4,12.9	14.8,15.3
AD4*	1.000	5.1, 5.5	12.4,13.1	14.7,15.4
AD5*	0.875	5.0, 5.5	12.4,13.2	14.8,15.3

* ADr denotes ante-dependence model of order r

in the same way. Each constant and weight used was found by the process described in sub-section 9.5.4 . Tables 9.6 and 9.7 show that the results for the use of the band-transition features plus area, area centromeric index and density centromeric index, although improved, are still not as good as those in Table 9.5 and Figures 9.3 to 9.6 .

To see if the non-parametric models for the band-transition sequences give candidate combinations of estimated percentage error-rate and allocation time for 46 chromosomes in a cell the results in Table 9.6 were plotted on the same figures as the results for the multivariate Normal procedures applied to the WDD features (Figures 9.1 to 9.6). As in chapters 5, 6 and 7 the allocation times were the average of ten c.p.u. times for the same operands for programs written in Fortran 77 using double-precision arithmetic executed on the Edinburgh University NAS computer. A discriminant score for each of the models for the band-transition sequences was calculated by taking natural logs of the estimates of the right-hand side of each of equations (9.1) to (9.5), adding the weight times the discriminant score for the Normal distribution and the natural log of the prior probability for the particular class. The discriminant score for the Normal distribution was calculated as described in chapter 5 and multiplied by a $\frac{1}{2}$. The figures show that none of the models (9.1) to (9.5) for the probabilities of the band-transition sequences multiplied by the value of a trivariate Normal p.d.f. for area, area centromeric index and density centromeric index raised to the power of an estimated weight (labelled as procedures BTS1 to BTS5 respectively) give candidate procedures.

Figure 9.1

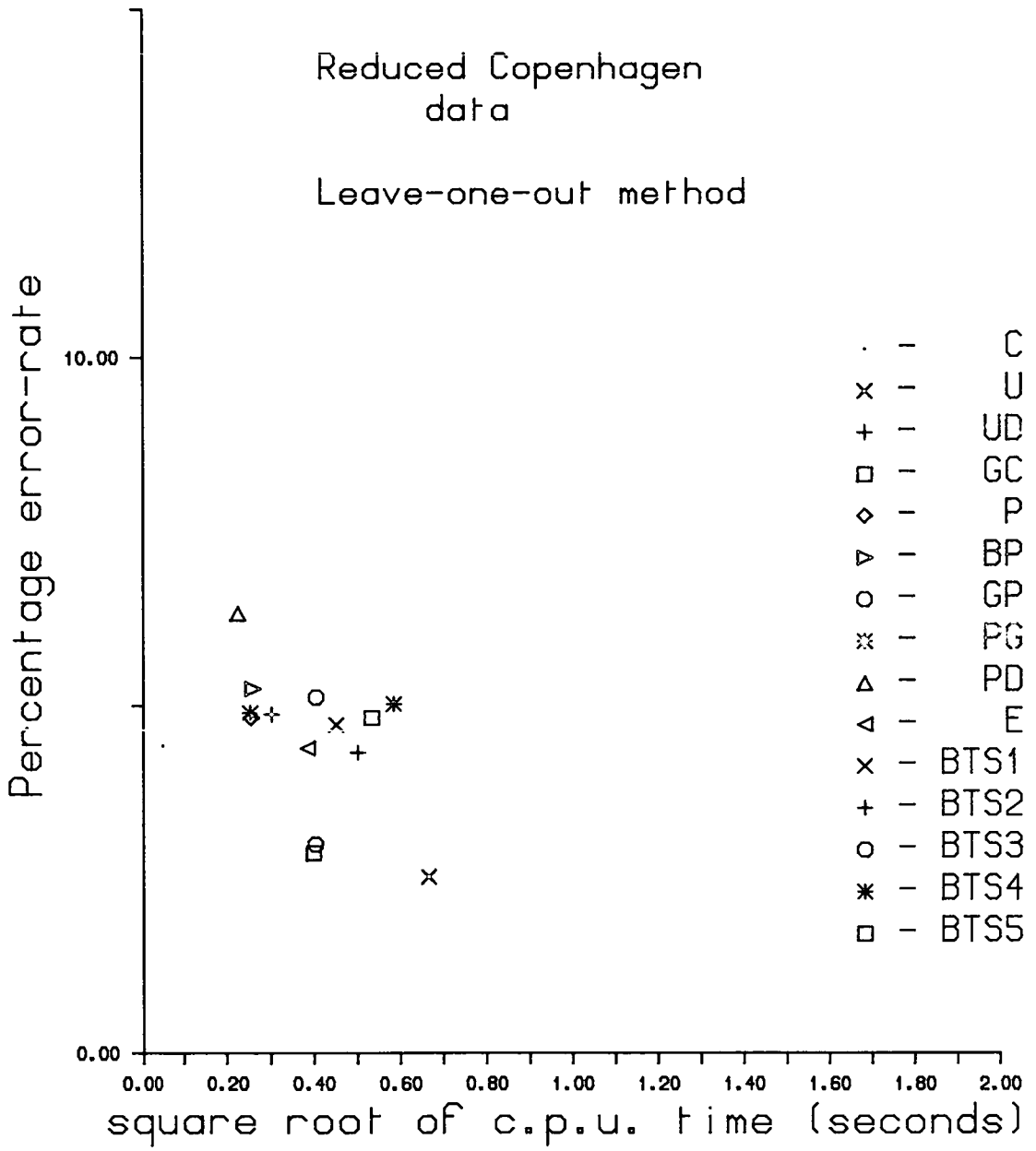

Reduced Copenhagen data
Leave-one-out method
Percentage error-rate
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)
. - C
× - U
+ - UD
□ - GC
◇ - P
▷ - BP
○ - GP
※ - PG
△ - PD
◁ - E
× - BTS1
+ - BTS2
○ - BTS3
✳ - BTS4
□ - BTS5

Figure 9.2

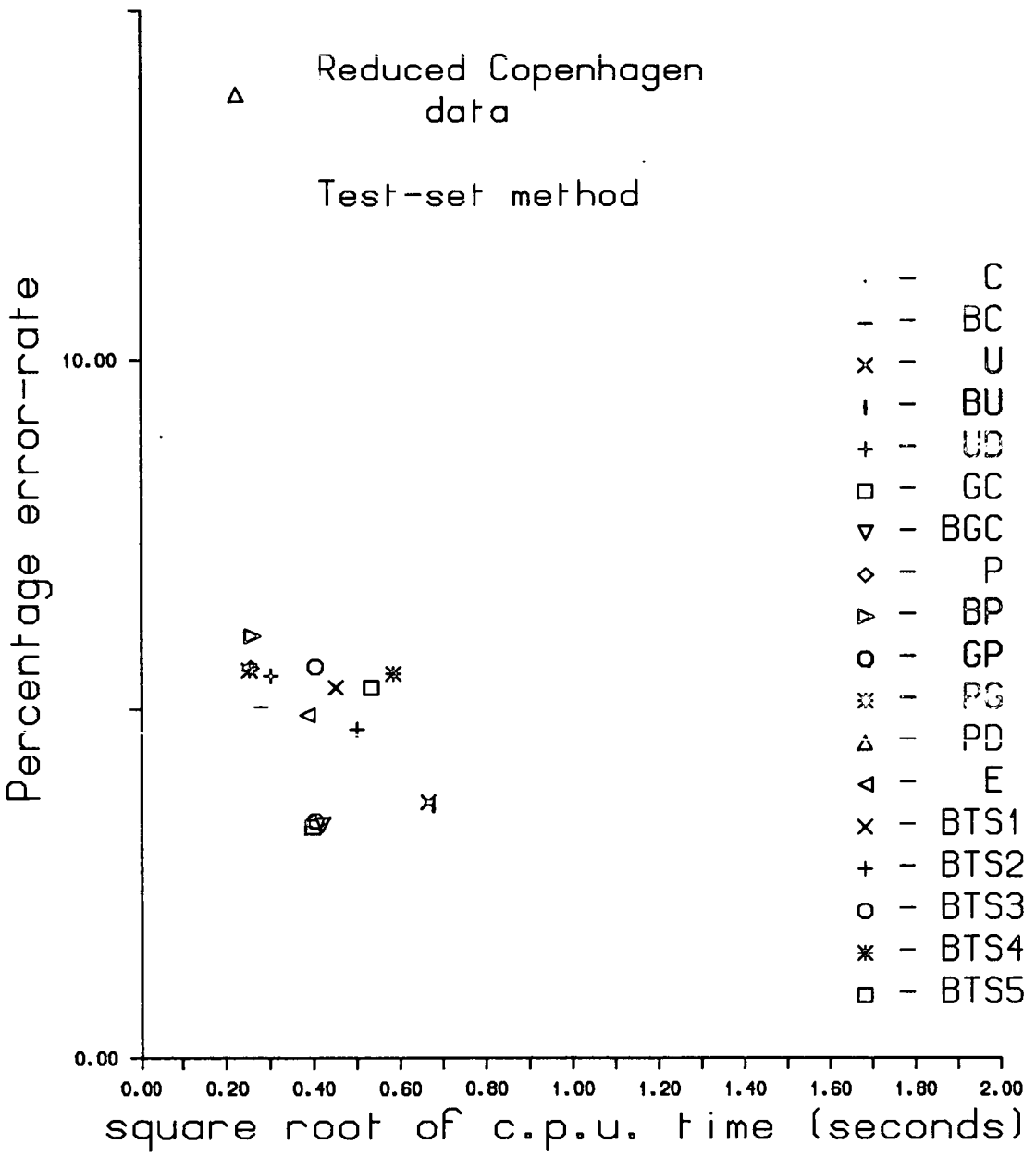

Reduced Copenhagen data
Test-set method
Percentage error-rate
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)
C
BC
U
BU
UD
GC
BGC
P
BP
GP
PG
PD
E
BTS1
BTS2
BTS3
BTS4
BTS5

Figure 9.3

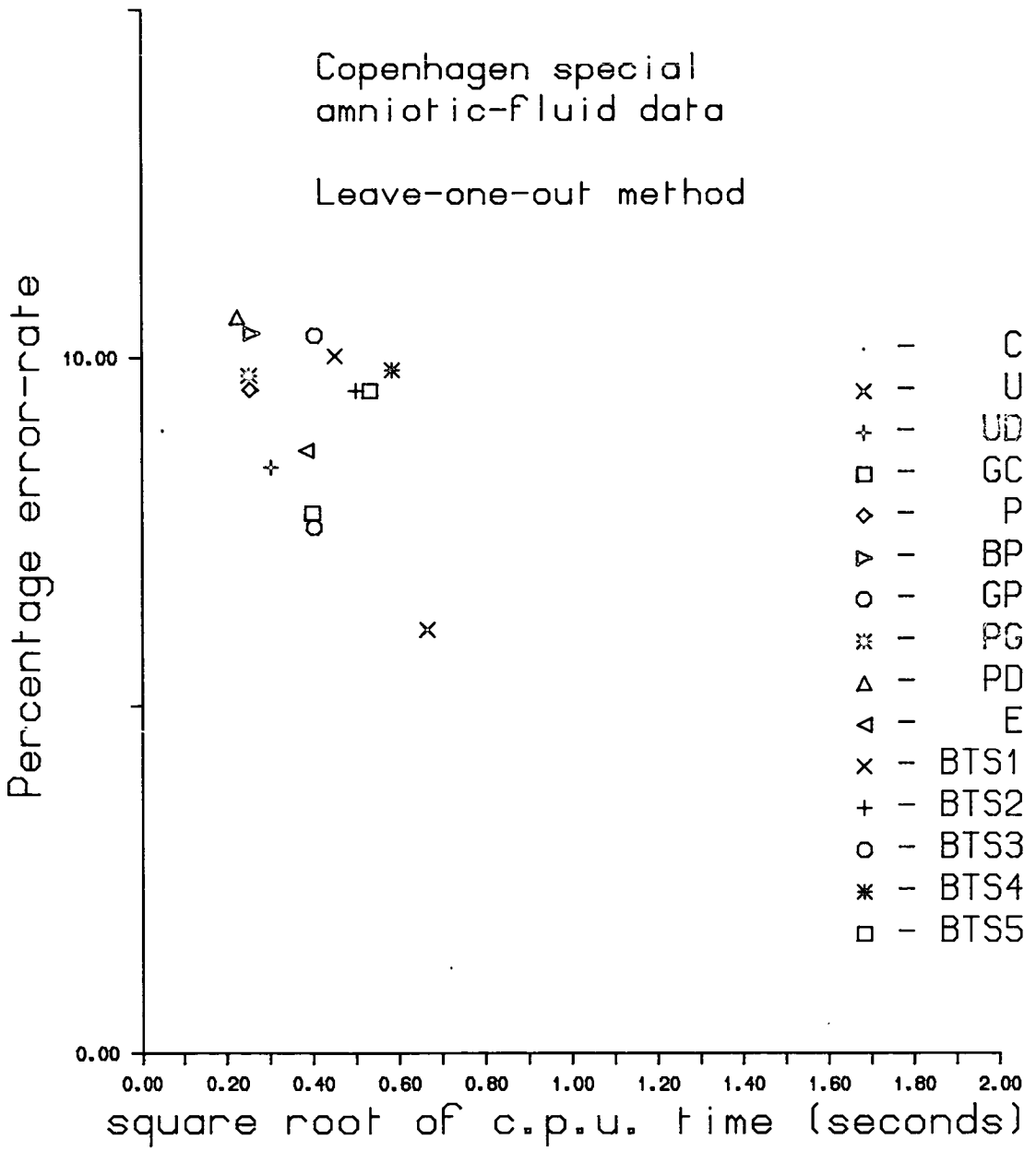

Copenhagen special
amniotic-fluid data
Leave-one-out method
Percentage error-rate
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)
. - C
× - U
+ - UD
□ - GC
◇ - P
▷ - BP
○ - GP
※ - PG
△ - PD
◁ - E
× - BTS1
+ - BTS2
○ - BTS3
✳ - BTS4
□ - BTS5

Figure 9.4

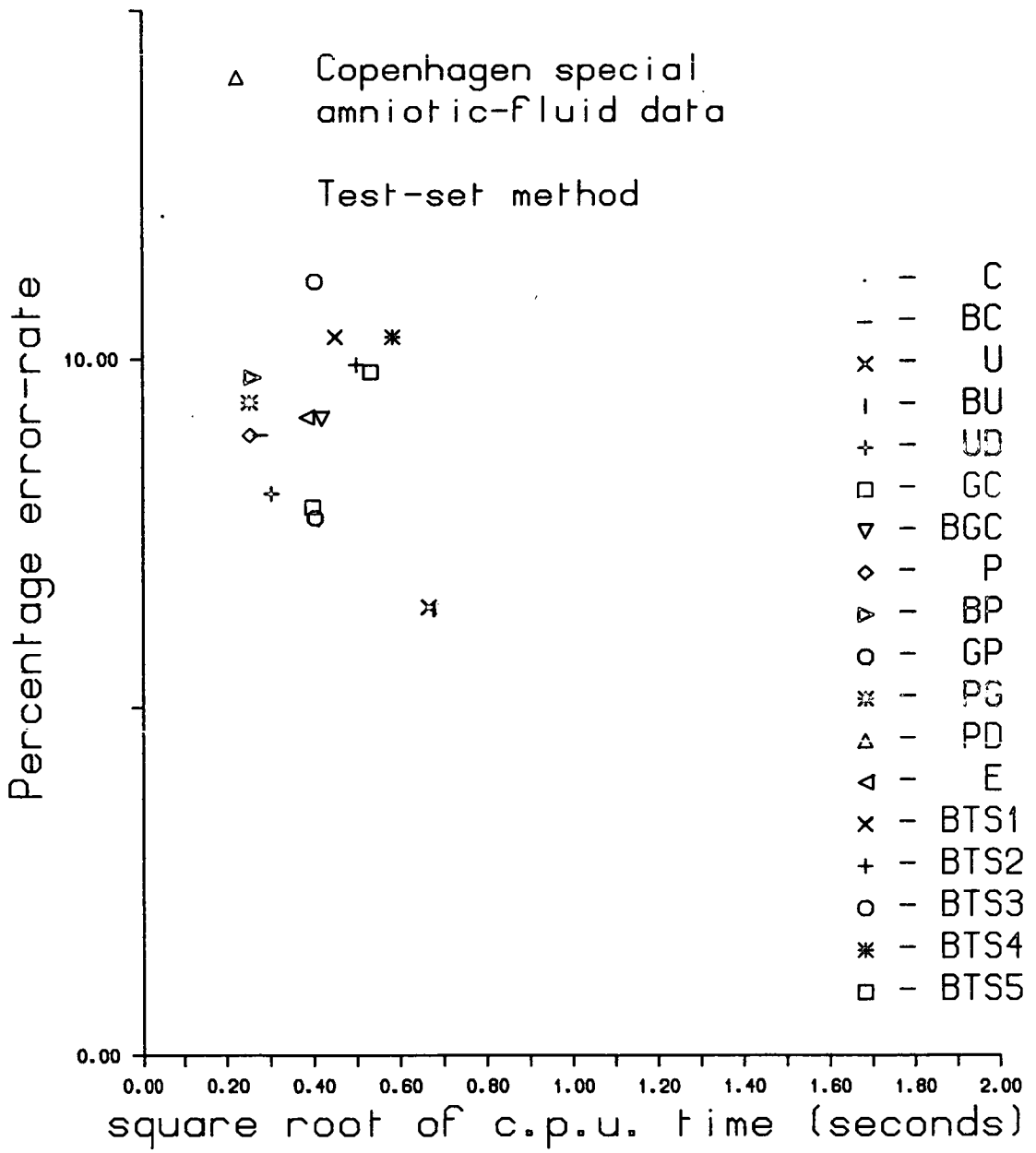

Copenhagen special
amniotic-fluid data
Test-set method
Percentage error-rate
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)
· - C
– - BC
× - U
| - BU
+ - UD
□ - GC
▽ - BGC
◇ - P
▷ - BP
○ - GP
※ - PG
△ - PD
◁ - E
× - BTS1
+ - BTS2
○ - BTS3
✳ - BTS4
□ - BTS5

Figure 9.5

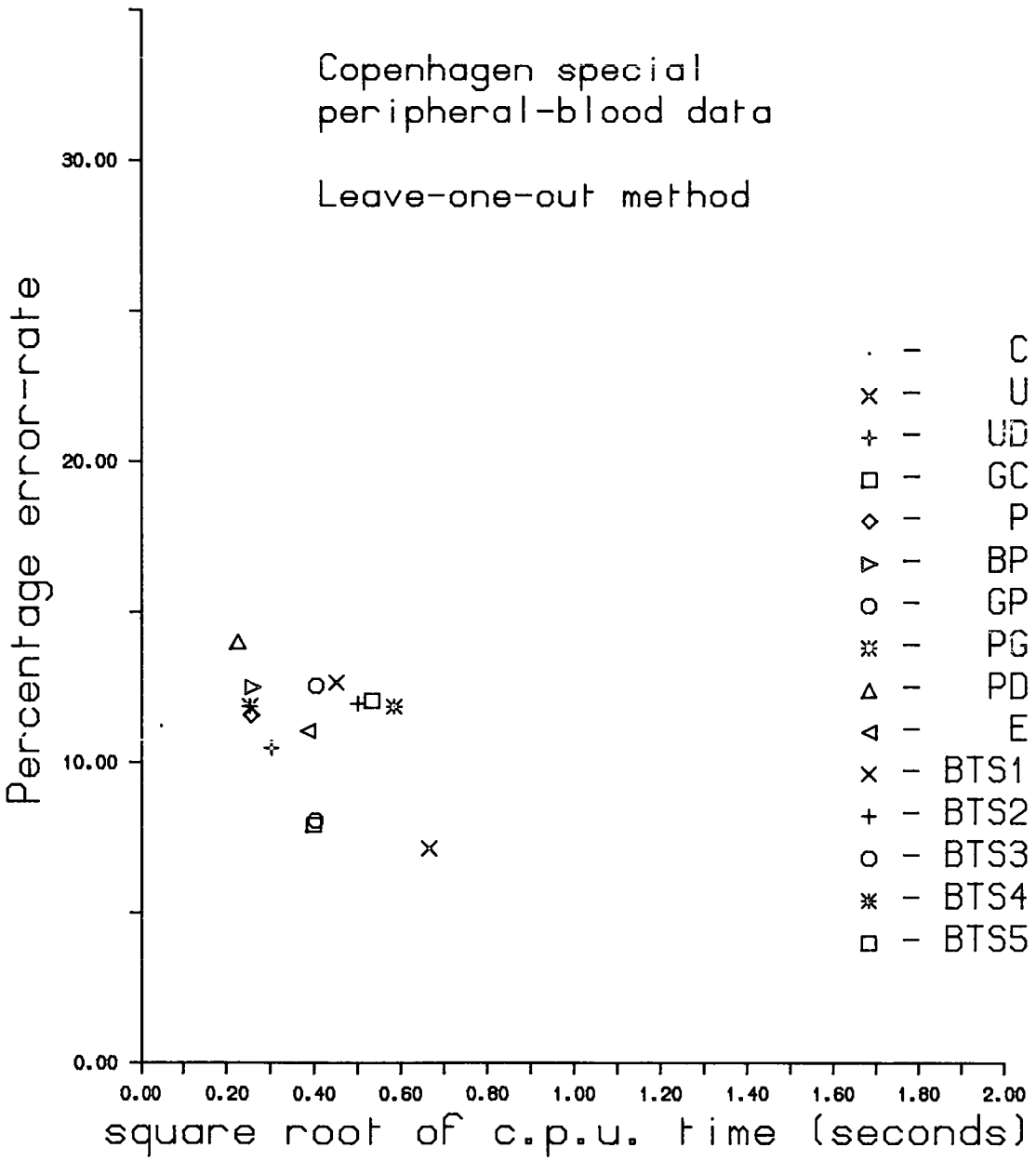
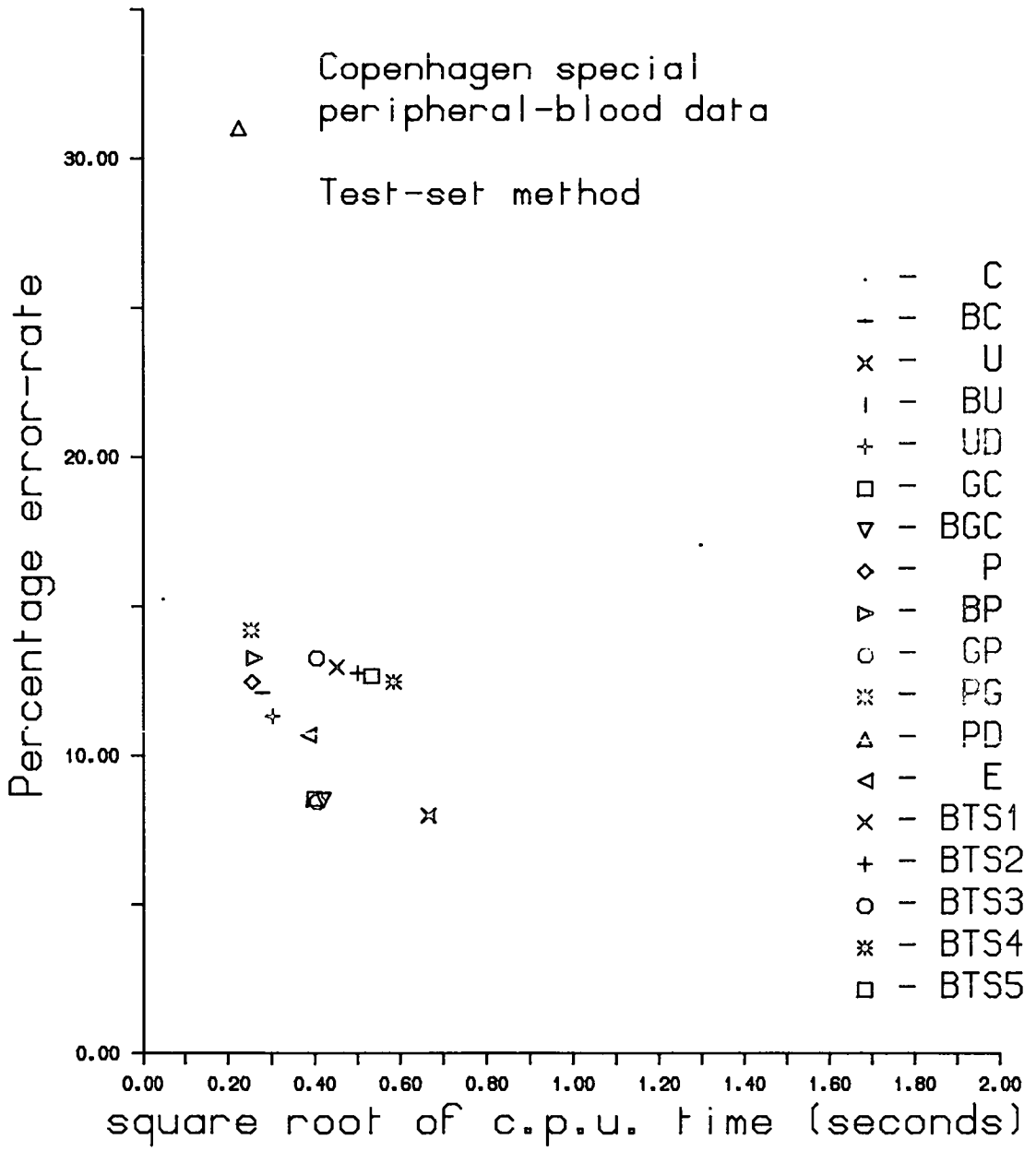

Copenhagen special
peripheral-blood data
Leave-one-out method
Percentage error-rate
30.00
20.00
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)
. - C
× - U
+ - UD
□ - GC
◇ - P
▷ - BP
o - GP
※ - PG
△ - PD
◁ - E
× - BTS1
+ - BTS2
o - BTS3
※ - BTS4
□ - BTS5

Figure 9.6


Copenhagen special
peripheral-blood data
Test-set method
Percentage error-rate
30.00
20.00
10.00
0.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
square root of c.p.u. time (seconds)
. - C
– - BC
× - U
| - BU
+ - UD
□ - GC
▽ - BGC
◇ - P
▷ - BP
○ - GP
※ - PG
△ - PD
◁ - E
× - BTS1
+ - BTS2
○ - BTS3
※ - BTS4
□ - BTS5

Chapter 10
The automated allocation of cervical smear specimens.

10.1 Introduction.

As described in chapter 3 , cervical smears are classified by cytologists on an ordinal scale of increasing risk of malignancy. For most of the specimens in the data set described in chapter 3 there are also a constant number of feature values obtained from the operation of an object discrimination procedure and the intervention of an operator. It is apparent from inspection of different cytologists' assessments of the same specimens that cytologists may differ in their allocations. Consequently, it is preferable to look for a discrimination method that explicitly allows for this uncertainty. The approach adopted here is to derive a consensus probability of a cervical smear specimen being abnormal using the method described by Dawid and Skene (1979). Multiple regression equations are then used to try to predict the logit transformations of the consensus probabilities for the specimens. A first multiple regression equation, which uses as predictors just features measured automatically on a specimen, is used if the probability derived from the predicted logit is below or above certain threshold values. If the probability derived from the predicted logit from this first equation lies between these threshold values a second multiple regression equation is used. This second equation makes use of the features derived from the intervention of an operator as well as the features obtained automatically. Decision rules based on the probabilities derived from the predicted logit transformations are used to automatically allocate a specimen as normal or abnormal. Previously, linear discriminant functions have been calculated for the two classes 'normal' and 'CIN1-3 or invasive' (see Table 3.5) derived from reference diagnoses which regard the cytologists' scores as continuous features (Carothers, 1988).

10.2 A consensus probability of a cervical smear specimen being abnormal.

To make use of all four cytologists' allocations a consensus probabilty of a cervical smear being abnormal was derived using the method described by Dawid and Skene (1979). This method is to attempt to maximise the likelihood

$$\prod_i \{\Sigma_j p_j \prod_k \prod_l (\pi_{jl}^{(k)})^{n_{il}^{(k)}}\} \qquad (10.1)$$

where p_j is the probability that a cervical smear drawn at random has true category j , $\pi_{jl}^{(k)}$ is the probability that cytologist k will allocate a smear to category l given j is the true category and $n_{il}^{(k)}$ is one if cytologist k allocates the ith specimen to category l and zero otherwise. The product for i is from i = 1 to i = I , for k is from k = 1 to k = K and for l is from l = 1 to l = J . The summation for j is from j = 1 to j = J . The likelihood is derived from assuming a multinomial distribution for the numbers of allocations to each category for each cytologist if the true category takes a particular value. If q is the true category for specimen i the likelihood for cytologist k is

$$\prod_l (\pi_{ql}^{(k)})^{n_{il}^{(k)}} . \qquad (10.2)$$

Assuming independent allocations by the cytologists, the likelihood for the allocations of the specimen is

$$\prod_k \prod_l (\pi_{ql}^{(k)})^{n_{il}^{(k)}} . \qquad (10.3)$$

If the assumption that the true category of specimen i is known is dropped then the probability of the data for specimen i is

$$\Sigma_j p_j \prod_k \prod_l (\pi_{jl}^{(k)})^{n_{il}^{(k)}} . \qquad (10.4)$$

The EM algorithm (Dempster, Laird and Rubin, 1977) may be applied to find maximum likelihood estimates of the p_j and $\pi_{jl}^{(k)}$ in (10.1). The name of this algorithm derives from the two steps in the algorithm, the first of which is an Expectation of missing data given current estimates of the parameters, and the second of which is a Maximisation of the likelihood given current estimates of the missing data. This algorithm can be used because if the true category for

each smear were known the likelihood for the full data would be

$$\prod_i \prod_j \{p_j \prod_k \prod_l (\pi_{jl}^{(k)})^{n_{il}^{(k)}}\}^{T_{ij}}, \tag{10.5}$$

where if q is the true category $T_{ij} = 1$ if $j = q$ and 0 otherwise. The steps of the EM algorithm are to take initial estimates of the 'missing data', i.e. the T_{ij}, then to use the maximum-likelihood estimators

$$\hat{\pi}_{jl}^{(k)} = \Sigma_i T_{ij} n_{il}^{(k)} (\Sigma_l \Sigma_i T_{ij} n_{il}^{(k)})^{-1} \quad (j=1,...,J,\ l=1,...,J,\ k=1,...,K) \tag{10.6}$$

and

$$\hat{p}_j = \Sigma_i T_{ij} (I)^{-1} \quad (j=1,...,J) \tag{10.7}$$

and finally to calculate

$$p(T_{ij}=1|\text{data}) = \prod_k \prod_l (\pi_{jl}^{(k)})^{n_{il}^{(k)}}\ p_j \{\Sigma_q \Pi_k \Pi_l (\pi_{ql}^{(k)})^{n_{il}^{(k)}} p_q\}^{-1} \tag{10.8}$$

where $p(T_{ij}=1|\text{data})$ is the probability that $T_{ij} = 1$ given the data. These steps are repeated until convergence is achieved. As pointed out by Dawid and Skene (1979), the EM algorithm only guarantees a local maximum. However, they found that the initial estimates

$$\hat{T}_{ij} = \Sigma_k n_{il}^{(k)} (\Sigma_k \Sigma_l n_{il}^{(k)})^{-1} \tag{10.9}$$

gave good results in practice. Each $\hat{T}_{ij}$ for the application considered here corresponds to the estimated consensus probability of a cervical smear

belonging to category j . Uebersax and Grove (1990) have pointed out that in the general case, when there are I objects, K assessors and J categories, a necessary condition for identifiability of the parameters is that there are three assessors for two classes. For three classes they note that a necessary condition is that there are five assessors. For the data here, from four cytologists, this means that at most two classes can be assumed for identifiability of the parameters to be possible. More generally, Goodman (1974) notes that model identifiability requires the rank of the matrix of derivatives of pattern probabilities with respect to 'a basic set' of model parameters to be equal to the number of columns when each row of the matrix corresponds to the derivatives for a particular pattern probability. By pattern probabilities is meant the probabilities associated with each possible pattern of assessor allocations for an object. By 'basic set' is meant a set of the smallest number of parameters from which values of the remaining parameters can be calculated. The derivatives of the pattern probabilities with respect to the parameters are

$$\prod_k \prod_l (\pi_{jl}^{(k)})^{n_{il}^{(k)}} \quad (j=1,\ldots,J) \tag{10.10}$$

for the p_j and

$$p_j \prod_k \prod_l (\pi_{jl}^{(k)})^{n_{il}^{(k)}} \quad (j=1,\ldots,J,\ l=1,\ldots J) \tag{10.11}$$

for the $\pi_{jl}^{(k)}$ where the product excludes the term involving $\pi_{jl}^{(k)}$ and (10.11) goes to zero if the power $n_{il}^{(k)}$ for $\pi_{jl}^{(k)}$ is zero. The $n_{il}^{(k)}$ are zero or one according to the particular pattern probability.

10.3 A multiple regression model for probabilistic assessments.

Given probabilistic assessments for each of J possible categories for an object, Aitchison and Begg (1976) propose the logit transformation

$$v_{ij} = \ln \{p_{ij}p_{i1}^{-1}\} \quad (i=1,\ldots,I,\ j=2,\ldots,J) \quad , \tag{10.12}$$

where p_{ij} is the probability of the ith object belonging to the jth category, as being a useful way of transforming the data to be on the scale $-\infty$ to ∞ . They then assume a multivariate Normal distribution for the vector $(\underline{v},\underline{x})$ where $\underline{v}$ is the vector of logit transformations and $\underline{x}$ is a vector of feature values. Use of the vague Normal-Wishart prior (Aitchison and Dunsmore, 1975, page 21) for $(\underline{\mu},\underline{\Sigma})$, the mean and covariance matrix of $(\underline{v},\underline{x})$, then gives the Bayesian predictive conditional density function of $\underline{v}$ given $\underline{x}$

$$\begin{aligned} &St[n-1,\ \bar{\underline{v}} + \underline{S}_{vx}\underline{S}_{xx}^{-1}(\underline{x} - \bar{\underline{x}}),\ \{(n+1)n^{-1}\} \\ &\quad x\quad (n-1)^{-1}(\underline{S}_{vv} - \underline{S}_{vx}\underline{S}_{xx}^{-1}\underline{S}_{xv})\{1 + (n-1)^{-1}Q(\underline{x})\}] \end{aligned} \tag{10.13}$$

where St is the generalized Student distribution (Aitchison and Begg, 1976), n is the number of observations,

$$\underline{S}_{vv} = \Sigma_i(\underline{v}_i - \bar{\underline{v}})(\underline{v}_i - \bar{\underline{v}})^T \quad , \tag{10.14}$$

$$\underline{S}_{xx} = \Sigma_i(\underline{x}_i - \bar{\underline{x}})(\underline{x}_i - \bar{\underline{x}})^T \quad , \tag{10.15}$$

$$\underline{S}_{vx} = \Sigma_i(\underline{x}_i - \bar{\underline{x}})(\underline{v}_i - \bar{\underline{v}})^T \tag{10.16}$$

and

$$Q(\underline{x})=(\underline{x} - \bar{\underline{x}})^T\{(1 + n^{-1})\underline{S}_{xx}(n-1)^{-1}\}^{-1}(\underline{x} - \bar{\underline{x}}) \quad . \tag{10.17}$$

Aitchison and Begg (1976) show that the distribution of $\underline{v}$ given $\underline{x}$ may be reduced to a single value p_{i2} for the case of two classes by using an approximation to the integration required to obtain the mean of the corresponding distribution of p_{i2} . This approximation is

$$\Phi[b\{2.942 + kc(k - 2)^{-1}\}^{-\frac{1}{2}}] \ , \tag{10.18}$$

where Φ is the standard Normal distribution function, for the density function St(k, b, c).

The method derived by Aitchison and Begg (1976) requires, however, multivariate Normality of the feature vectors which is not satisfied for the cervical smear data because of extreme skewness for some of the features in the vector $\underline{x}$. A Bayesian predictive approach may still be used, however, for the conditional distribution of $\underline{v}$ given $\underline{x}$ by fitting a regression model for the prediction of $\underline{v}$ from $\underline{x}$. The Bayesian predictive approach is to obtain the predictive conditional density $p(\underline{v}|\underline{x})$ from

$$\int_{\underline{\Theta}} p(\underline{v}|\underline{\Theta})\ p(\underline{\Theta}|\underline{x}\ ,\ z)d\underline{\Theta} \ , \tag{10.19}$$

where $\underline{\Theta}$ is a vector of parameters and z is given data. The $p(\underline{\Theta}|\underline{x},\ z)$ is a conditional posterior density function for $\underline{\Theta}$ based on a prior distribution for $\underline{\Theta}$, given data z and a feature vector $\underline{x}$. Use of the vague Normal-Wishart prior for the conditional mean and covariance matrix of $\underline{v}$ given $\underline{x}$ and the multivariate Normal p.d.f. for $p(\underline{v}|\underline{\Theta})$ gives equation (10.13) with (n - 1) replaced by the number of degrees of freedom for the regression model and the full feature vector $\underline{x}$ replaced by the feature vector of the features included as predictors in the model.

10.4 Sequential use of multiple regression equations for the allocation of cervical smear specimens.

The explanatory features for specimen allocation become available in two stages. The first stage features are those which are available from the operation of the object discrimination procedure with no intervention from an operator. Those at the second stage are explanatory features that result from the operator's intervention. The aim is to achieve a suitable trade-off between machine error-rates and the need for operator intervention. Here, this is explored by allowing a range of thresholds for the probabilities derived from the logit transformations predicted by a multiple regression equation using only the first stage features. These thresholds are lower and upper bounds for the probability of a specimen being abnormal. Below the lower probability and above the upper probability allocation of a specimen is made without the intervention of an operator. If the probability is between these thresholds the second-stage features are obtained and a second multiple regression equation based on both first and second stage features is used to predict the logit transformation for a specimen. The trade-off between false-positive and false-negative error-rates is studied for a number of decision rules based on the probabilities derived from the predicted logit transformations.

10.5 Multiple regression equations.

The cervical smear specimens were divided at random into two parts of equal size and regression equations for the logit transformation of the consensus probability of a smear being abnormal regressed on the feature values derived from the first part. For the training data, features resulting from operator intervention were available for all specimens. Features were included in the regression equations if the estimated residual mean square error was reduced. Squared and cross-product terms for such features were then added if the residual mean square error was further reduced. The inclusion of terms was done by forward selection. Plots of the residuals against the fitted values and each explanatory feature showed no obvious departures from the model assumptions.

10.6 Criterion for allocation of a smear and error rate estimation.

At the first stage a smear from the second part of the data was allocated to the normal class if the probability of its being abnormal derived from the predicted logit transformation was below the lower threshold and to the abnormal class if this probability lay above the upper threshold. If the additional explanatory features were required to allocate a specimen the decision rule at the second stage was used to allocate the specimen.

Each error rate was estimated for each combination of first-stage thresholds and second-stage allocation rule using the estimated consensus probabilities in expression (10.6) with $n_{il}^{(k)}$ replaced by the allocation described in the paragraph above. This is the maximum-likelihood estimator when the consensus probabilities are known.

10.7 Results

10.7.1 The consensus probabilities.

306 cervical smear specimens for which all four cytologists gave an allocation and feature values were available were used. To obtain a two-category classification for the cytologists' allocations the normal classification was coded as 0 and other classifications were coded as 1 (see Table 3.5). The starting values recommended by Dawid and Skene (1979) for the consensus probabilities were used. A number of other arbitrary starting values were also tried but these gave either the same or a smaller value of the likelihood (10.1). The results show that the consensus probability is almost one for a category agreed on by a majority of the cytologists. Specimens for which there were two cytologists recording the verdict 'normal' and two recording 'abnormal' show a more even split in the consensus probabilities between the two classes. Evaluation of the matrix of derivatives of pattern probabilities with respect to 'a basic set' of the estimated parameters shows that the rank is equal to the number of columns when each row of the matrix corresponds to the derivatives for a particular pattern probability. The parameters are, therefore, identifiable (Goodman, 1974). Estimated cytologist error-rates suggest that some of the cytologists have very high error rates (Table 10.1).

Table 10.1

Estimated cytologist error-rates under the model corresponding to the likelihood in (10.1) .

Cytologist	False-positive	False-negative
1	0.06	0.35
2	0.06	0.06
3	0.79	0.06
4	0.20	0.08

10.7.2 Estimated error-rates for the use of the multiple regression equations.

The estimated error rates for all combinations of the following lower and upper probabilities at the first stage are plotted, joined by splines, in Figures 10.1-10.9:

lower probabilities

0.0001 0.001 0.01 0.05 0.1 0.2

upper probabilities

0.6 0.7 0.8 0.9 0.95 0.99 .

These extreme values of the lower threshold were used because of the importance of avoiding false-negative results for the allocation of cervical smear specimens. The proportion of specimens allocated at the first stage for each pair of thresholds is given in Table 10.2 . At the second stage, specimens not already allocated were allocated to the normal class if their probability of being abnormal derived from the predicted logit transformation was below a certain threshold and to the abnormal class otherwise. The probabilities used as thresholds at the second stage were

0.001, 0.01, 0.05, 0.10, 0.15
0.20 , 0.25, 0.30, 0.35, 0.40
0.50 , 0.55, 0.60, 0.65, 0.70
0.75 , 0.80, 0.85, 0.90, 0.95 .

10.8 Discussion.

Figures 10.1-10.9 show that automated allocation with the possibilty of operator intervention can give error-rates similar to the error-rates of the third cytologist in Table 10.1 . It would need to be ensured, however, that the level

Table 10.2

Proportion of specimens allocated to either class without operator intervention for given lower and upper probabilities at the first stage below and above which an allocation is made.

Proportion	Lower probability	Upper probability
0.373	0.0001	0.6
0.268	0.0001	0.7
0.183	0.0001	0.8
0.105	0.0001	0.9
0.065	0.0001	0.95
0.007	0.0001	0.99
0.373	0.001	0.6
0.268	0.001	0.7
0.183	0.001	0.8
0.105	0.001	0.9
0.065	0.001	0.95
0.007	0.001	0.99
0.373	0.01	0.6
0.268	0.01	0.7
0.183	0.01	0.8
0.105	0.01	0.9
0.065	0.01	0.95
0.007	0.01	0.99
0.379	0.05	0.6
0.275	0.05	0.7
0.190	0.05	0.8
0.111	0.05	0.9
0.072	0.05	0.95
0.013	0.05	0.99
0.418	0.10	0.6
0.314	0.10	0.7
0.229	0.10	0.8
0.150	0.10	0.9
0.111	0.10	0.95
0.052	0.10	0.99
0.529	0.20	0.6
0.425	0.20	0.7
0.340	0.20	0.8
0.261	0.20	0.9
0.222	0.20	0.95
0.163	0.20	0.99

Figure 10.1
Estimated false-positive and false-negative error-rates

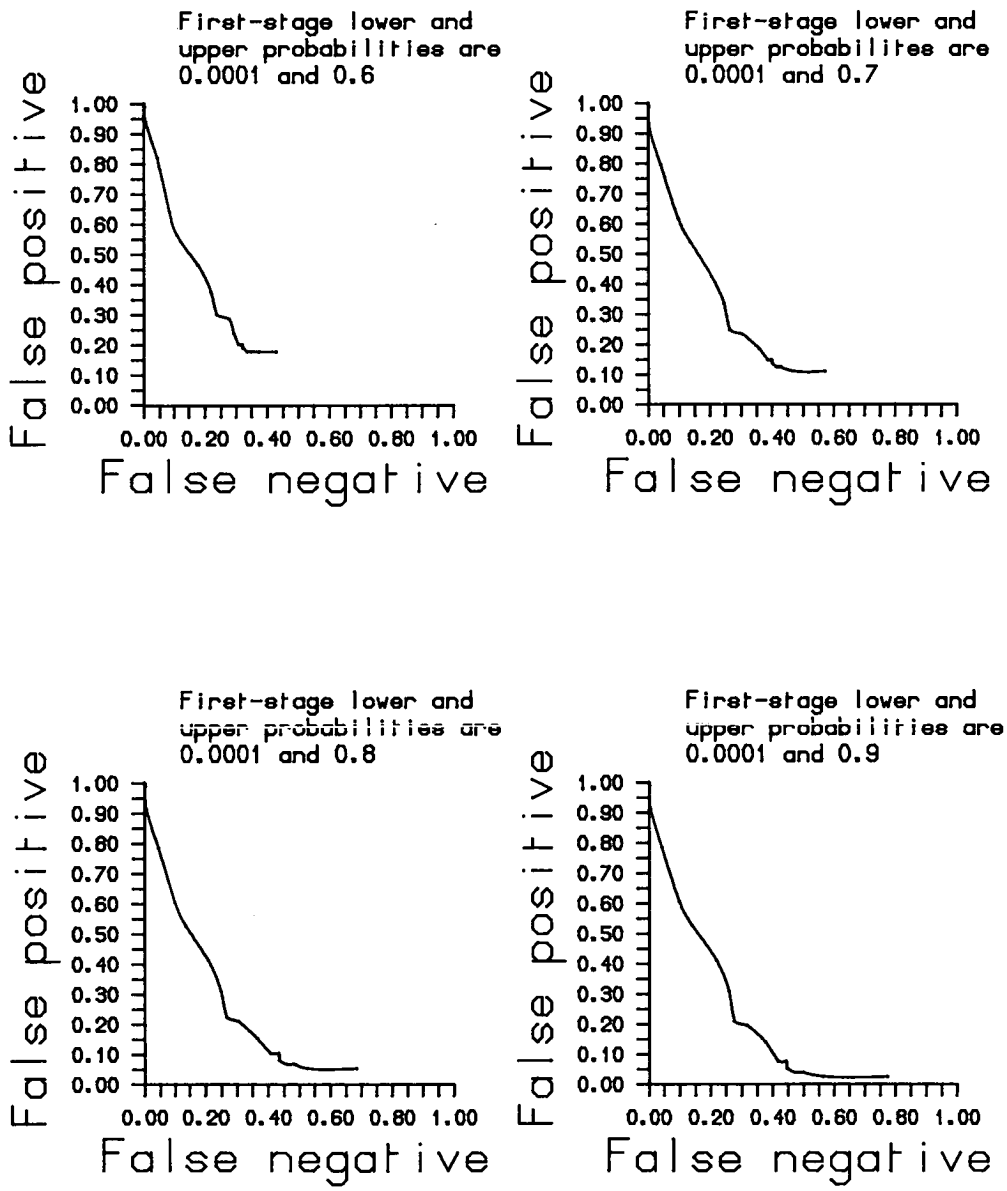


Figure 10.2
Estimated false-positive and false-negative error-rates

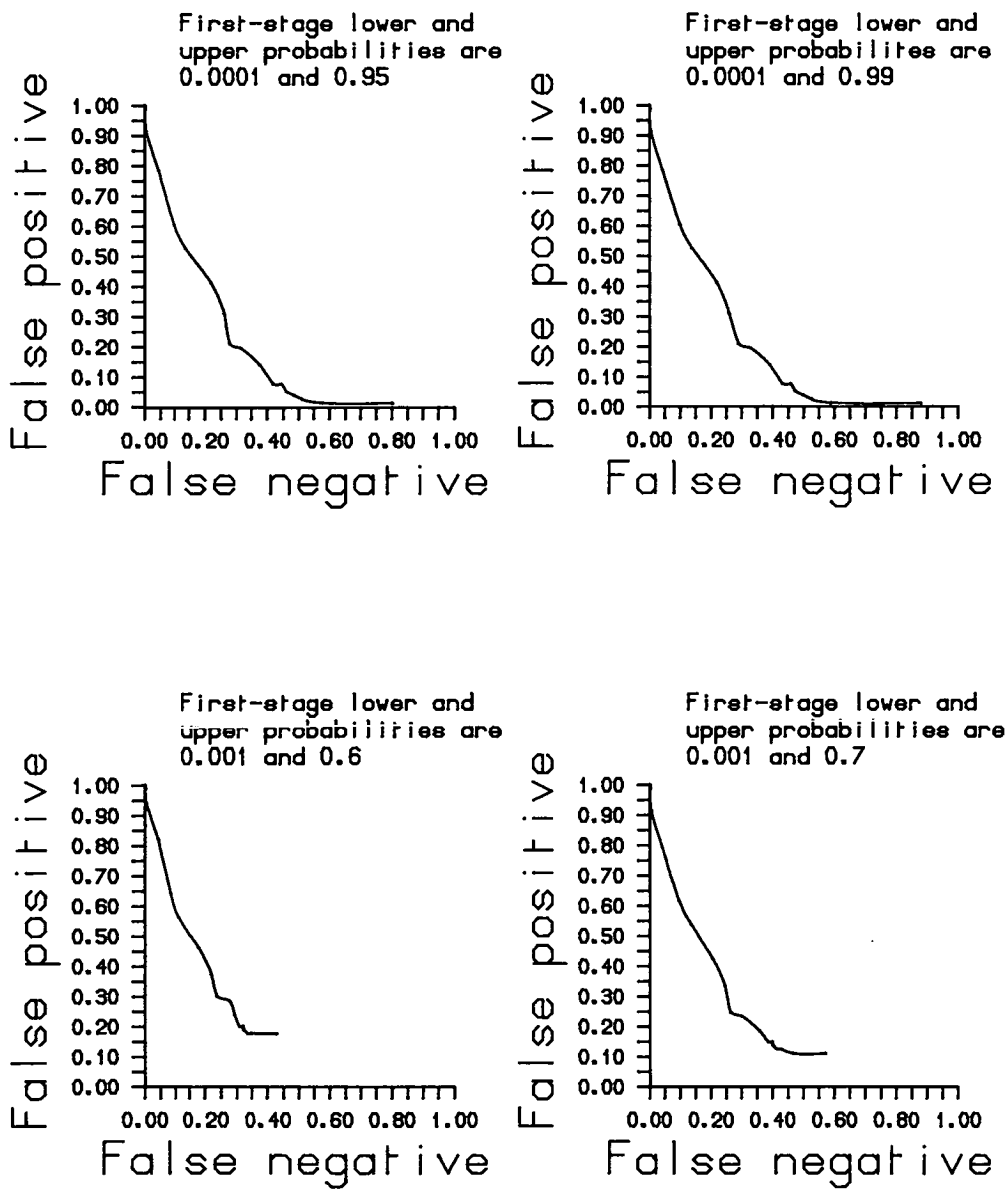


Figure 10.3
Estimated false-positive and false-negative error-rates

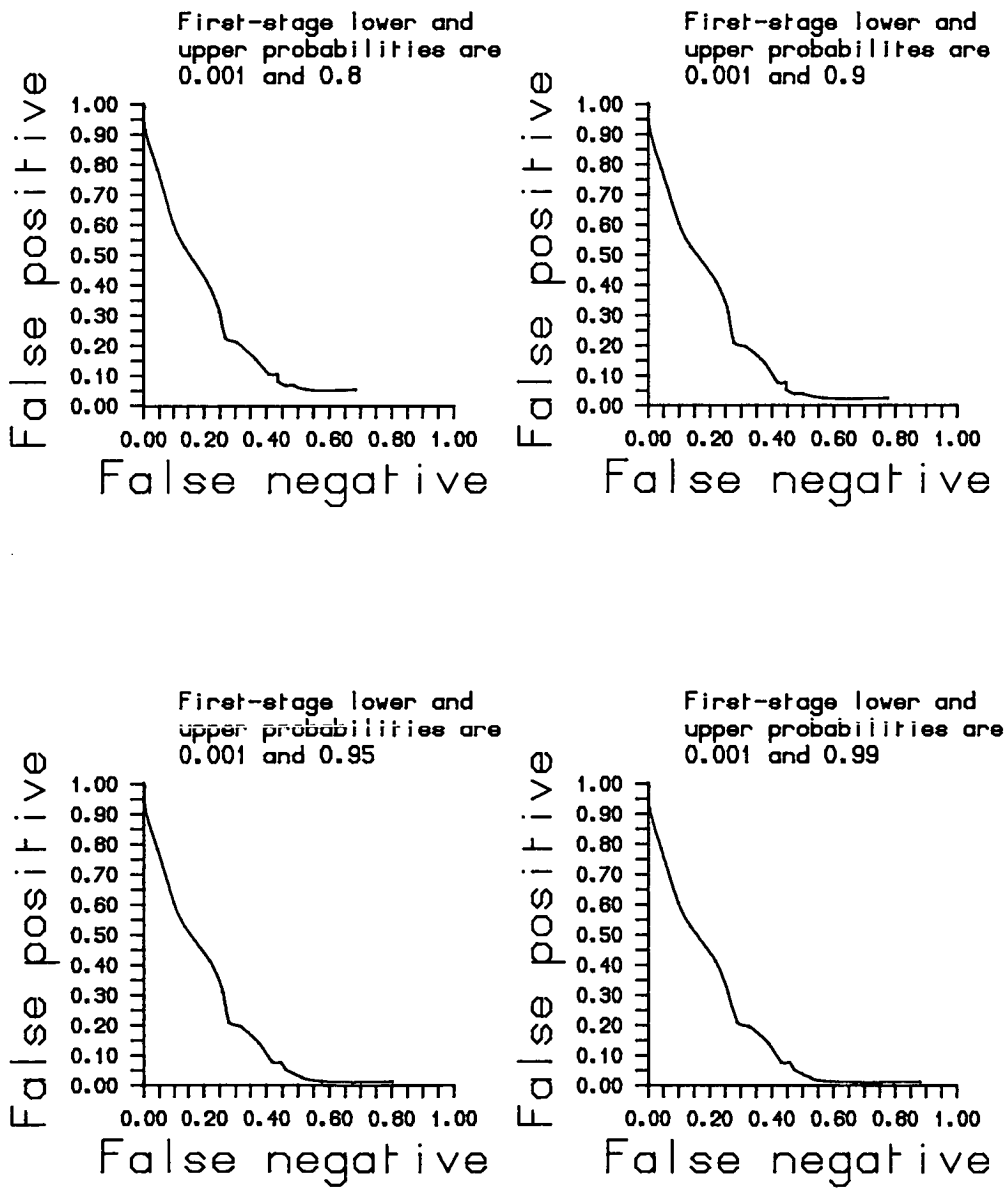


Figure 10.4
Estimated false-positive and false-negative error-rates

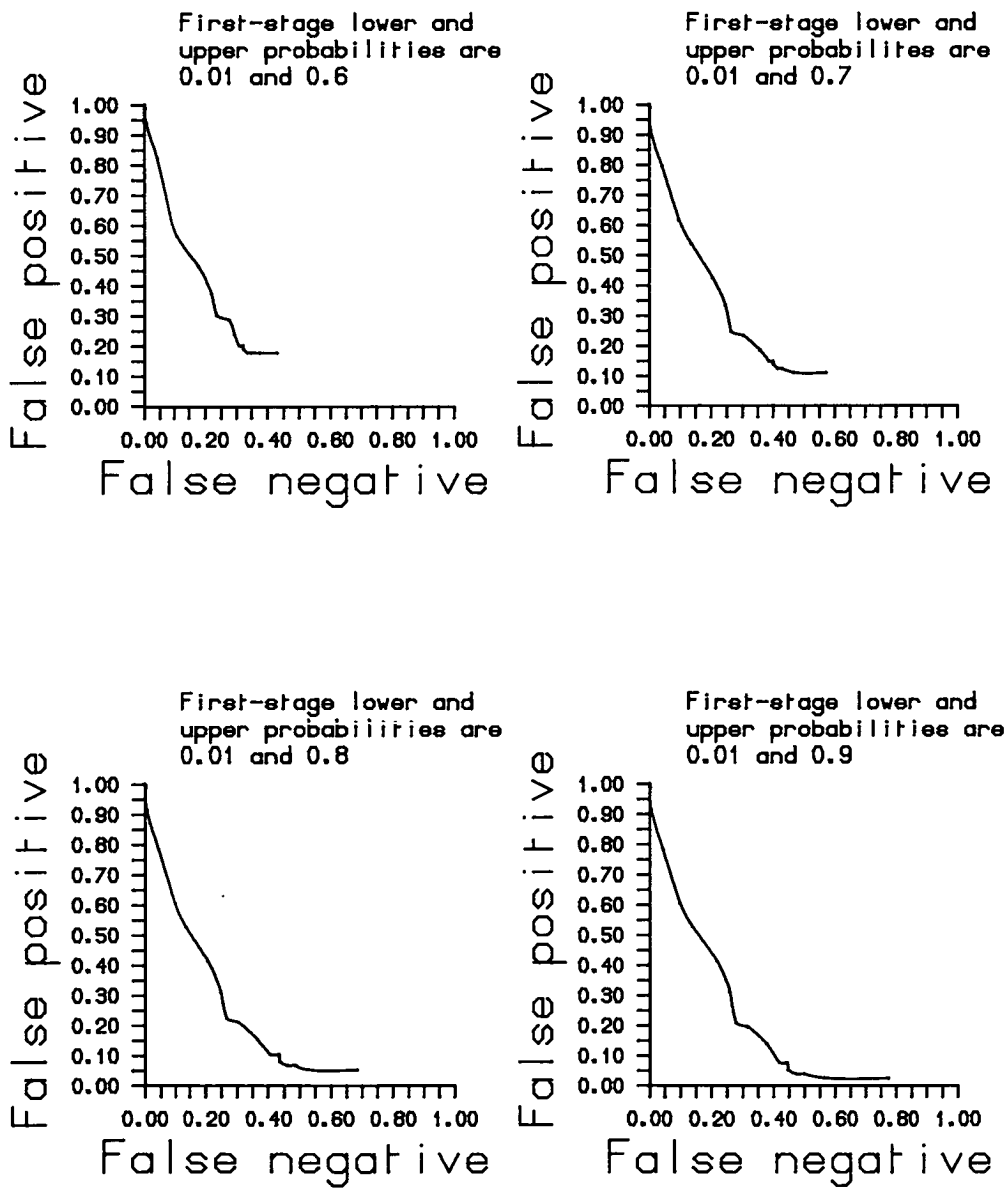


Figure 10.5
Estimated false-positive and false-negative error-rates

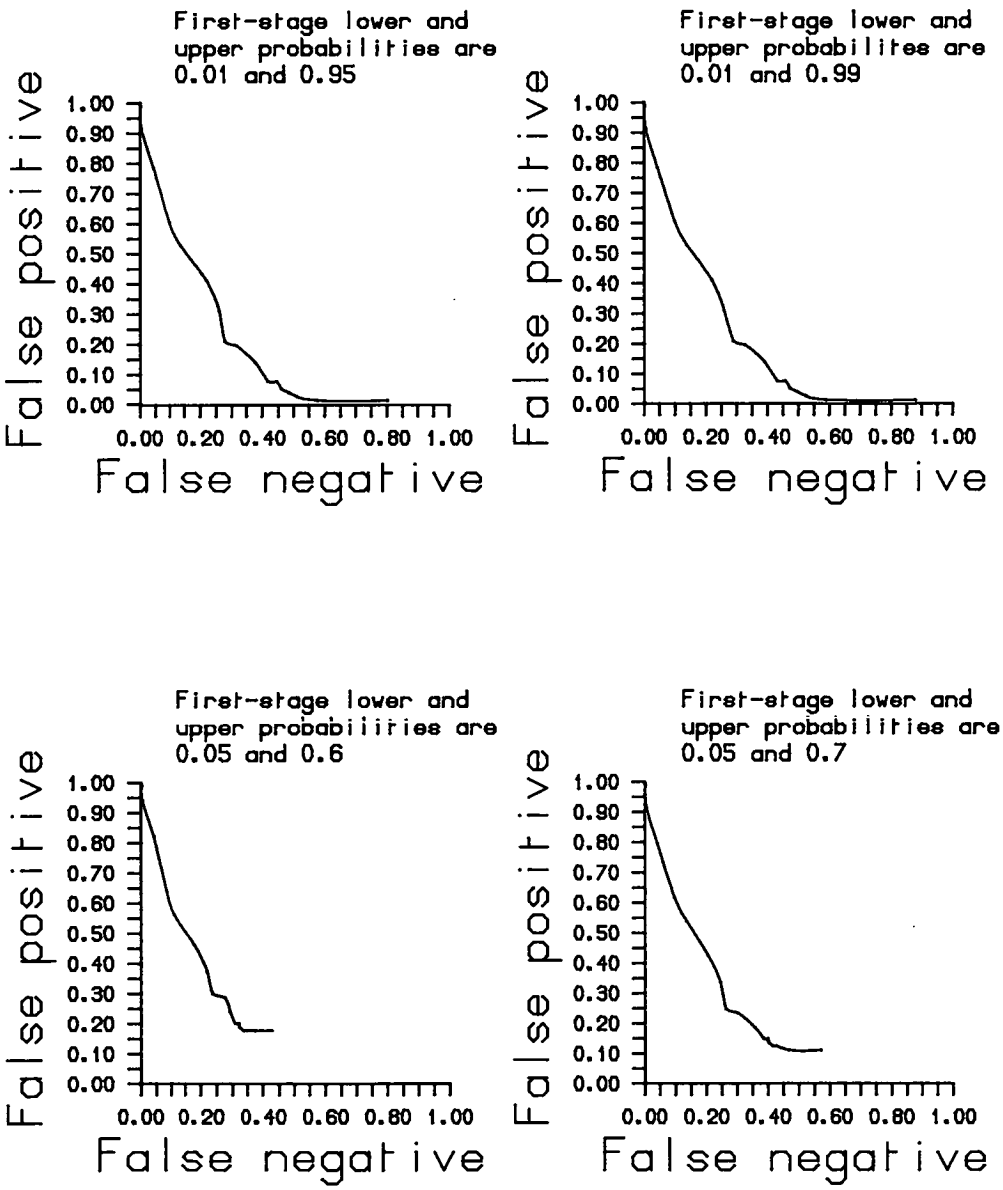


Figure 10.6
Estimated false-positive and false-negative error-rates

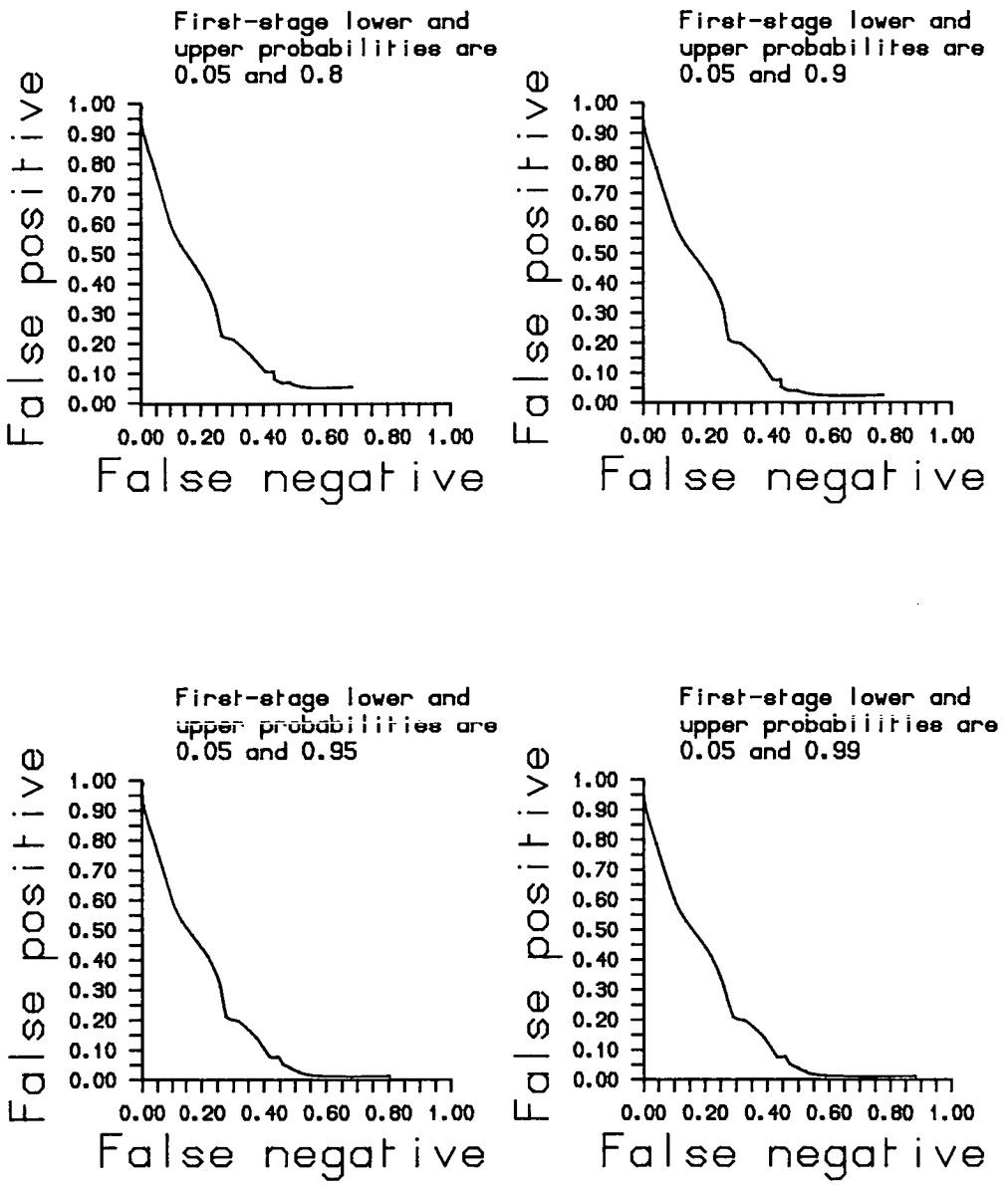


Figure 10.7
Estimated false-positive and false-negative error-rates

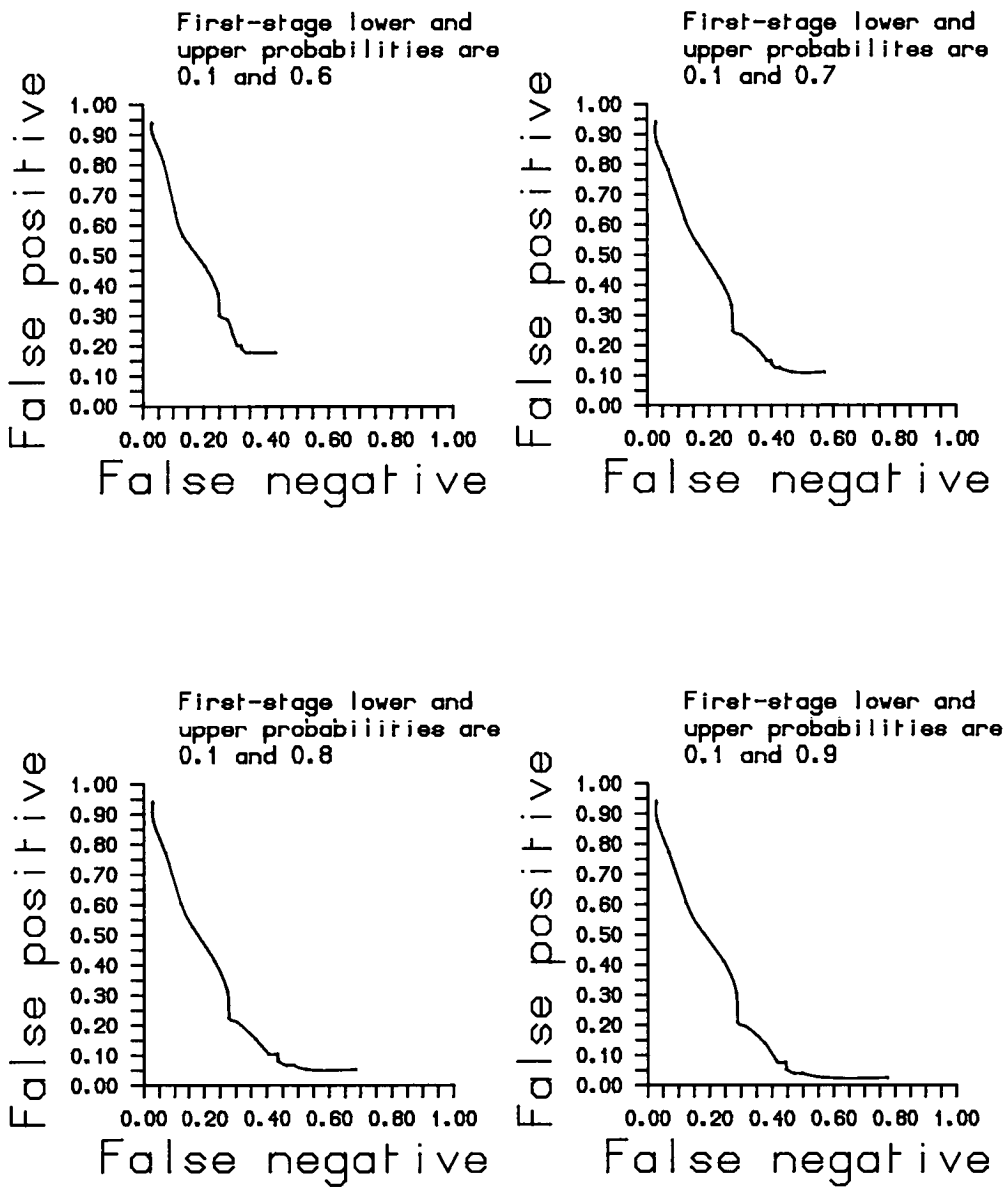


Figure 10.8
Estimated false-positive and false-negative error-rates

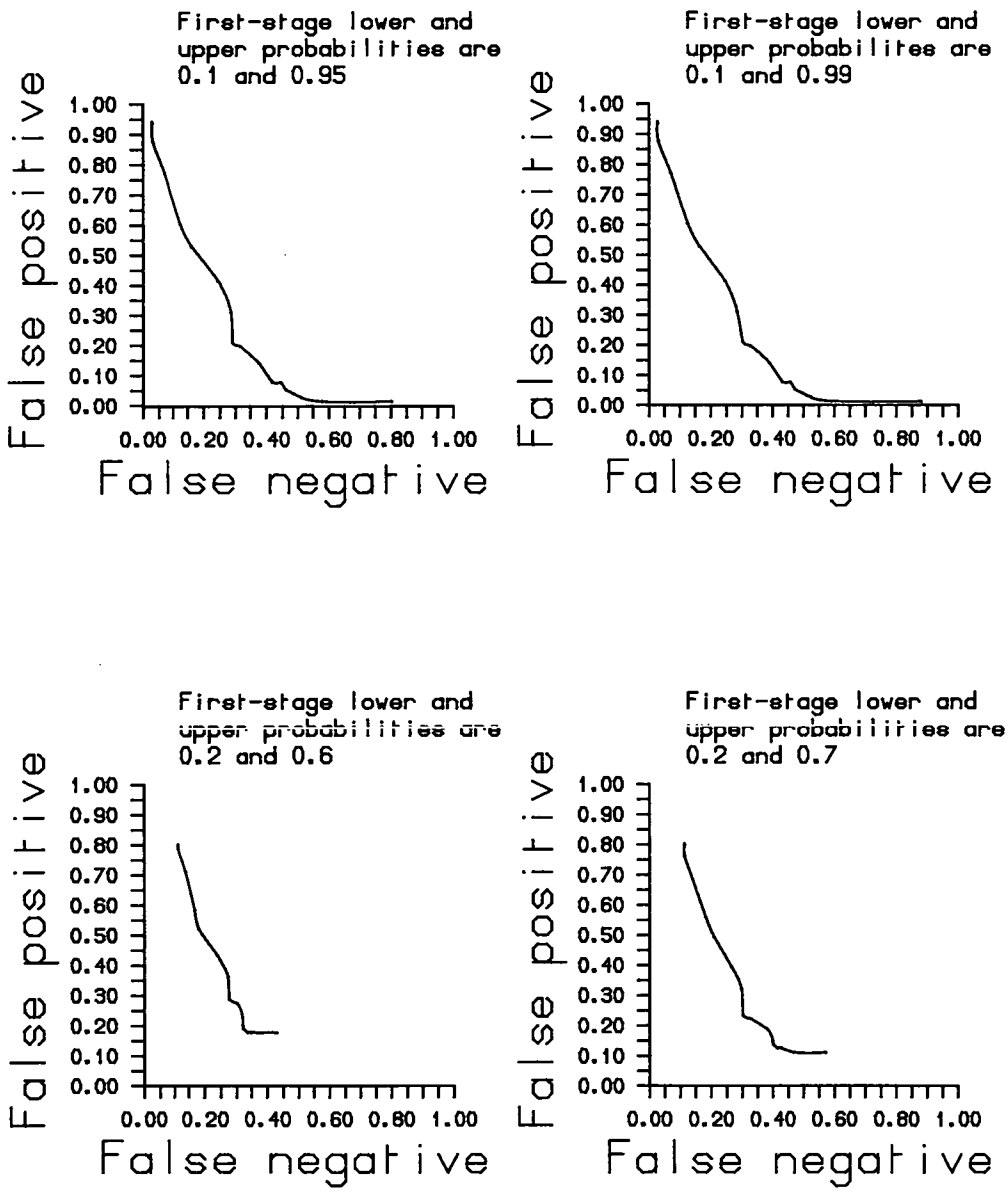
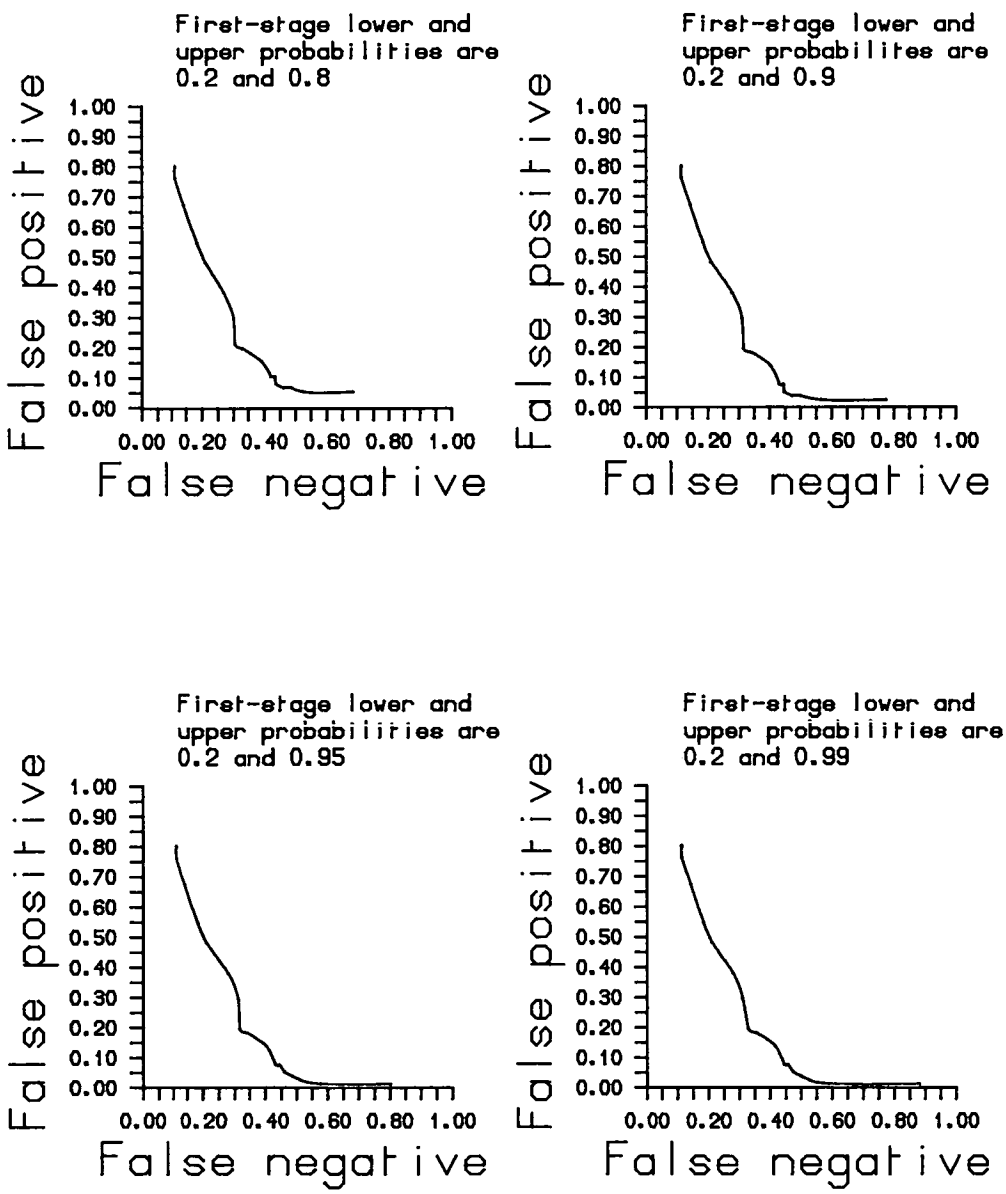


Figure 10.9
Estimated false-positive and false-negative error-rates



of operator intervention was acceptable in terms of time. The results indicate that to achieve the false-negative error rates of the first, second and fourth cytologists a bigger false-positive error-rate than that obtained by each of these cytologists would need to be acceptable. It should be noted for these results that the assumption is made that Dawid and Skene's model is an adequate one for modelling cytologists' assessments.

The division of the specimens into normal or abnormal classes is probably not entirely satisfactory for an assessment of automated allocation (with operator intervention). It is likely to be of interest to repeat this analysis when the parameters for three categories are identifiable in the model corresponding to the likelihood (10.1). Then the specimens could be classified as normal, slightly malignant and severely malignant. The error rates for severely malignant specimens given by this 'automated' allocation, which are likely to be extremely important, could then be estimated.

Chapter 11
Sequential use of features for multivariate discrimination.

11.1 Introduction

When each feature has an associated cost, sequential use of the features in two or more stages with allocation rules at each stage may give a lower cost discrimination procedure than the use of all features at one time. The cost associated with a feature may be made up of a measurement cost and/or the cost of calculating discriminant scores if these are calculated immediately after a value for the feature is obtained. As described in chapter 2, the allocation of individual objects in a cervical smear is an application in which sequential discrimination is used to reduce the time taken on feature measurement. In this chapter the sequential measurement of features is mainly considered for the simplified problem of two known multivariate Normal populations with equal covariance matrices when the only cost associated with a feature is measurement cost.

The sequential measurement of features for discrimination has been studied by Fu (1968), Zielezny and Dunn (1975) and Hora (1980) when feature measurement cost and misallocation cost are commensurable and there is no cost in calculating discriminant scores. Fu (1968) looked at the use of a modified sequential probability ratio test for a fixed order of feature measurement and dynamic programming solutions for fixed and varying orders of feature measurement when all parameters are known. Zielezny and Dunn (1975) considered a two-stage procedure for two multivariate Normal populations with equal covariance matrices in which "cheap" features are measured first and the remaining "expensive" features are then measured if the cost of collection is less than the reduction in expected misallocation cost. They found that for known parameters the two-stage procedure was always at least as "cheap" as using just the first subset of features or all the features. For unknown parameters the two-stage procedure was also best for a wide range of conditions. Hora (1980) has shown that for two known multivariate Normal populations with equal covariance matrices Fu's dynamic programming solution for a fixed order of feature measurement is computationally reasonable because successive posterior probabilities follow a Markov process.

In this chapter the computation required for Fu's dynamic programming solution for an optimal varying order of feature measurement is briefly reviewed. An alternative approach to finding an optimal varying order of feature measurement, when the feature order is free to vary, is to have a fixed order of feature measurement which has minimum cost amongst all fixed orders. This approach may require less computation to obtain a solution. Solutions are considered for two known multivariate Normal distributions with equal covariance matrices for Fu's criterion and a new criterion for allocation of objects, when less than the full number of features have been measured. The new criterion is related to the error rate achieveable when all features are used to allocate objects. Both criteria assume that the only cost associated with a feature is measurement cost. Even this approach can require substantial computation to obtain a solution, so a sub-optimal approach to obtaining a fixed order of feature measurement for each of these two criteria is examined. The evaluation of an optimal fixed order of feature measurement is also considered when the cost of the calculation of the discriminant scores after a feature has been measured is included in the measurement cost of a feature. In this case the two criteria for early allocation need to be re-defined and it may not be optimal to calculate the discriminant scores after each feature. It may also be true that a one-stage procedure gives a lower cost discrimination procedure than a sequential procedure. An empirical approach using density estimates of discriminant scores for the new criterion is suggested for observations from any two distributions when the parameters of the distributions are unknown and the only cost associated with a feature is measurement cost. Finally, this empirical approach is used for the sequential discrimination between artefacts and cells for cervical smears.

11.2 An optimal variable order of feature measurement when the only cost associated with a feature is measurement cost.

To write down Fu's solution the following notation is defined:

1. $\underline{x}_k$ is a vector of k feature values.
2. $\rho(\underline{x}_k|F_{tk})$ is the minimum expected risk having obtained $\underline{x}_k$ where the particular order, F_{tk} , of k of the features is used.
3. F_k is the set of features remaining to be measured.
4. $c(\underline{x}_k|F_{tk})$ is the cost of the next feature, f_{tk+1} , when F_{tk} is selected.
5. $F(x_{k+1}; f_{tk+1}|\underline{x}_k; F_{tk})$ is the conditional distribution function of x_{k+1} when f_{tk+1}

is selected given the measurements $\underline{x}_k$ for the sequence F_{tk} .

6. $R(\underline{x}_k; d_i|F_{tk})$ is the risk of allocating to class i on the basis of $\underline{x}_k$ when F_{tk} is selected.

Fu (1968), page 80, shows that the basic functional equation for a maximum of p features is

$$\rho(\underline{x}_k|F_{tk}),\ k=1,\ldots,(p-1) =$$

$$\text{Min}\begin{cases}\text{Continue:} & \text{Min}_{j_{tk+1}\ \varepsilon\ F_k}\{\, c(\underline{x}_k|F_{tk}) + \int \rho(x_1,\ldots,x_k,x_{k+1}|F_{tk},f_{tk+1})\ dF(x_{k+1};f_{tk+1}|\underline{x}_k;F_{tk})\}\\ \text{Stop:} & \text{Min}_i\ R(\underline{x}_k;d_i|F_{tk})\ ,\ i=1,2\ .\end{cases} \tag{11.1}$$

This may be solved by working backwards from

$$\rho(\underline{x}_p) = \text{Min}_i\ R(\underline{x}_p;d_i|F_{tp})\ . \tag{11.2}$$

If we consider the discrete case with each feature taking one of I possible values then this means that the number of risk functions to be evaluated is

$$2\{I^p + \binom{p}{p-1} I^{(p-1)}+\ldots+\binom{p}{1} I\ \} \tag{11.3}$$

unless a simplifying assumption such as independence of the features can be made. It can be seen from this that the required computation may be excessive, e.g., for I=20 and p=8 the number of risk functions is approximately $7.6*10^{10}$.

11.3 Two criteria for the allocation of an object with a fixed order of feature measurement when k out of p features have been measured and the only cost associated with a feature is measurement cost.

11.3.1 Fu's criterion

With c_{k+1} defined as the cost of the (k + 1)th measurement, the basic functional equation (11.1) simplifies for a fixed order of feature measurement to

$$\rho(\underline{x}_k)\ k=1,\ldots,(p-1)$$

$$= \text{Min} \begin{cases} \text{Continue: } c_{k+1} + \int \rho(\underline{x}_{k+1})\, dF(x_{k+1}|\underline{x}_k) \\ \text{Stop: Min } R(\underline{x}_k;d_i)\ ,\ i=1,2\ . \end{cases} \qquad (11.4)$$

The optimal stopping rule may then be determined by working backwards starting from the given risk function at the last stage using dynamic programming.

11.3.2 Obtaining values for Fu's criterion for two known multivariate Normal populations with equal covariance matrices.

Hora (1980) has shown that it is computationally reasonable to obtain values for Fu's criterion in the case of two known multivariate Normal populations with equal covariance matrices. If $\underline{\Sigma}$ is non-singular, $f_i(\underline{x}_k)$ is the p.d.f. for class i and π_i its prior probability then

$$z_k = \ln\{f_1(\underline{x}_k)f_2(\underline{x}_k)^{-1}\} + \ln\{\pi_1\pi_2^{-1}\} \qquad (11.5)$$

is a linear function of $\underline{x}_k$ so that its distribution is Normal. If Q_k denotes

$$(\underline{\mu}_{k1} - \underline{\mu}_{k2})^T\underline{\Sigma}_k^{-1}(\underline{\mu}_{k1} - \underline{\mu}_{k2}) \qquad (11.6)$$

then z_k has variance Q_k in each population and expectations $\frac{1}{2}Q_k + \ln(\pi_1\pi_2^{-1})$ and $-\frac{1}{2}Q_k + \ln(\pi_1\pi_2^{-1})$ in populations 1 and 2 respectively. Clearly, the values of

$z_1,\ldots,z_p$ are unchanged if $x_2,\ldots,x_p$ are replaced by the sequence of residuals

$$r_k = x_k - E(x_k|\underline{x}_{k-1}) \quad (k=2,\ldots,p) \tag{11.7}$$

which are linear in $\underline{x}_k$ and independent of each other and of x_1; the corresponding covariance matrix is diagonal with kth diagonal element

$$\mathrm{Var}(x_k|\underline{x}_{k-1}) \quad (k=2,\ldots,p) \quad . \tag{11.8}$$

Each difference $z_k - z_{k-1}$ is therefore a linear function of r_k and hence they are independent. Thus for any $l > k$ we may write

$$z_l = z_k + e_{k+1} + \ldots + e_l \tag{11.9}$$

where $e_{k+1},\ldots,e_l$ are independent of each other and of z_k , and e_l has the distributions

$$N\{\tfrac{1}{2}(Q_l - Q_{l-1}),\ Q_l - Q_{l-1}\} \tag{11.10}$$

and

$$N\{-\tfrac{1}{2}(Q_l - Q_{l-1}),\ Q_l - Q_{l-1}\} \tag{11.11}$$

in populations 1 and 2. It follows that the distribution of $z_{k+1},\ldots,z_p$ given z_k is multivariate Normal with

$$E(z_l|\underline{z}_k, \text{ pop. } 1) = z_k + \tfrac{1}{2}(Q_l - Q_k) \tag{11.12}$$

$$E(z_l|\underline{z}_k, \text{ pop. } 2) = z_k - \tfrac{1}{2}(Q_l - Q_k) \tag{11.13}$$

$$\text{Var}(z_l|\underline{z}_k) = Q_l - Q_k \tag{11.14}$$

$$\text{Cov}(z_l, z_m| z_k) = Q_{\min(l, m)} - Q_k \quad (l,m > k) \ . \tag{11.15}$$

Thus the distribution of $z_{k+1},\ldots,z_p$ depends on $\underline{x}_k$ only through z_k and since z_m $(m=k+1,\ldots,p)$ is a known monotone transformation of the posterior probability for class 1 at the mth stage there is Markovian dependence between successive posterior probabilities.

To write down the risk of obtaining measurement x_{k+1} the following notation is defined:

1. $g_i(\underline{z})$ is the p.d.f. of $\underline{z}$ for the ith class.
2. z_{lk} and z_{uk} are the values of z_k below and above which objects not already allocated at the kth stage are allocated to population 2 and population 1 respectively.
3. $\underline{L}_{m1} = (z_{l,k+1},\ldots z_{l,m-1},z_{u,m})$, $\underline{L}_{m2} = (z_{i,k+1},\ldots z_{i,m-1},-\infty)$,
4. $\underline{U}_{m1} = (z_{u,k+1},\ldots,z_{u,m-1},\infty)$, $\underline{U}_{m2} = (z_{u,k+1},\ldots,z_{u,m-1},z_{l,m})$,
5. p_{ik} is the posterior probability of class i after k measurements.
6. $M_{ii'}$ is the cost of misallocating an object from population i as being from population i' .
7. $C_{k+1,m} = \Sigma_{l=k+1}^{l=m} c_l$.
8. R_i , i=1,2 is the risk of deciding population i .

$$9.\ I_{him} = \int_{\underline{L}_{mh}}^{\underline{U}_{mh}} g(\underline{z}|\underline{\theta}_i)d\underline{z} \ .$$

I_{him} is the probability of allocating to population h at the mth stage when i is the true population.

The risk of obtaining measurement x_{k+1} can then be expressed as

$$R_3 = \Sigma_{m=k+1}^{m=p}[p_{1k}\{(M_{12}+C_{k+1,m})I_{21m}+C_{k+1,m}I_{11m}\} + p_{2k}\{(M_{21}+C_{k+1,m})I_{12m}+C_{k+1,m}I_{22m})\}] \ . \qquad (11.16)$$

When d_3 is defined as the decision to obtain the value for the next feature an explicit expression may be obtained. This is done by working backwards defining sets S_{ij} for i=1,2,3 and j=1,...,(p − 1) such that d_i has the smallest expected loss if $p_{ij} \in S_{ij}$. Since $R_3(p_{1j})$ is a concave function of p_{1j} (De Groot, 1970, pages 125–127) there exists a γ_j and a λ_j such that d_1 is optimal if $p_{1j} \geqq \gamma_j$ and d_2 is optimal if $p_{1j} \leqq \lambda_j$ and another measurement should be obtained if $\lambda_j < p_{1j} < \gamma_j$. Values of γ_j and λ_j are found by setting

$$R_1(\gamma_j) = R_3(\gamma_j) \qquad (11.17)$$

and

$$R_2(\lambda_j) = R_3(\lambda_j) \ . \qquad (11.18)$$

11.3.3 A new criterion.

We consider that we wish to discriminate between objects from two populations with known p.d.f.s given by $f_i(\underline{x}_p)$ with prior probabilities π_i .

The criterion for early allocation of an object assumes that measurement costs and misallocation costs are not commensurable and is related to the final error rate achieveable using all features in a one-stage procedure. The criterion is to allocate an object when one feature has been measured if

$$z_1 \leqq z_{l,1}$$

or if

$$z_1 \geqq z_{u,1}$$

where z_{l1} and z_{u1} satisfy the equations

$$\int_{A}^{\infty}\int_{-\infty}^{z_{l,1}} g_1(z_1,z_p)dz_1dz_p - \int_{-\infty}^{A}\int_{z_{u,1}}^{\infty} g_1(z_1,z_p)dz_1dz_p = \alpha_1 \qquad (11.19)$$

and

$$\int_{-\infty}^{A}\int_{z_{u,1}}^{\infty} g_2(z_1,z_p)dz_1dz_p - \int_{A}^{\infty}\int_{-\infty}^{z_{l,1}} g_2(z_1,z_p)dz_1dz_p = \beta_1 \qquad (11.20)$$

and A denotes the value of z_p above which a new object is allocated to

population 1. The optimal rule for the one-stage procedure (Krzanowski, 1988, page 336) is to allocate to population 1 if

$$z_p > M_{21}M_{12}^{-1} . \tag{11.21}$$

At each of the subsequent (p - 2) intermediate stages, when another feature has been measured and $k < p$ $\{k=2,...,(p - 1)\}$ features have been measured in total, an object is allocated if

$$z_k \leqq z_{l,k}$$

or if

$$z_k \geqq z_{u,k}$$

where $z_{l,k}$ and $z_{u,k}$ satisfy the equations

$$\int_{A}^{\infty} \int_{-\infty}^{z_{l,k}} \int_{z_{l,k-1}}^{z_{u,k-1}} \cdots \int_{z_{l,1}}^{z_{u,1}} g_1(z_1 \ldots z_k, z_p) dz_1 \ldots dz_k, dz_p$$

$$- \int_{-\infty}^{A} \int_{z_{u,k}}^{\infty} \int_{z_{l,k-1}}^{z_{u,k-1}} \cdots \int_{z_{l,1}}^{z_{u,1}} g_1(z_1 \ldots z_k, z_p) dz_1 \ldots dz_k, dz_p = \alpha_k \tag{11.22}$$

and

$$\int_{-\infty}^{A}\int_{z_{u,k}}^{\infty}\int_{z_{l,k-1}}^{z_{u,k-1}}\cdots\int_{z_{l,1}}^{z_{u,1}} g_2(z_1\ldots z_k,z_p)dz_1\ldots dz_k,dz_p$$

$$-\int_{A}^{\infty}\int_{-\infty}^{z_{l,k}}\int_{z_{l,k-1}}^{z_{u,k-1}}\cdots\int_{z_{l,1}}^{z_{u,1}} g_2(z_1\ldots z_k,z_p)dz_1\ldots dz_k,dz_p = \beta_k \quad . \qquad (11.23)$$

Objects are allocated to population 1 if $z_k \geqq z_{u,k}$, to population 2 if $z_k \leqq z_{l,k}$ $\{k=1,\ldots,(p-1)\}$, otherwise another feature is measured. The α_k and β_k are the additional errors in allocation made at each stage compared with the one-stage procedure. These could be allowed to vary with k but to avoid excessive complexity we consider here the procedure where we let $\alpha_k = \alpha(p-1)^{-1}$ and $\beta_k = \beta(p-1)^{-1}$, where α and β are the total additional errors for each class compared with the one-stage procedure. The division of the additional error between the k stages to give a minimum cost procedure for a given feature ordering is considered further below.

11.3.4 Obtaining values for the new criterion for two known multivariate Normal populations with equal covariance matrices.

The joint distributions of z_1,z_p and $z_1,\ldots,z_k,z_p$ $\{2 \leqq k \leqq (p-1)\}$ are multivariate Normal for each population as can be seen using the results in sub-section 11.3.2 above. Equations (11.19), (11.20), (11.22) and (11.23) thus involve integrating over regions of multivariate Normal densities. The $z_{l,k}$ and $z_{u,k}$ obtained may be simply transformed back to posterior probability ratios and thence to posterior probabilities if these are required at each stage. It is proposed that the equations may be successively solved by a numerical search procedure.

11.4 An optimal fixed order of feature measurement when the only cost associated with a feature is measurement cost.

In principle we may define an optimal fixed order of feature measurement for either Fu's criterion or the criterion introduced in sub-section 11.3.3 , when

the feature order is free to vary with each object. From the derived $z_{l,k}$ and $z_{u,k}$ for k=1,...,(p - 1) for a given order of feature measurement we may calculate the total proportion of observations that may be allocated after each feature, except the last, has been measured. Defining P_{ik} , k=1,...,(p - 1) , to be the proportion of population i that may be allocated after the kth feature has been measured, this is equal to

$$\int_{z_{u,k}}^{\infty}\int_{z_{l,k-1}}^{z_{u,k-1}}\cdots\int_{z_{l,1}}^{z_{u,1}} f_i(\underline{z}_k)d\underline{z}_k + \int_{-\infty}^{z_{l,k}}\int_{z_{l,k-1}}^{z_{u,k-1}}\cdots\int_{z_{l,1}}^{z_{u,1}} f_i(\underline{z}_k)d\underline{z}_k \quad . \tag{11.24}$$

Hence for the criterion introduced in sub-section 11.3.3 the total cost of a fixed order of feature measurement is proportional to

$$\Sigma_i \Sigma_{k=1}^{k=p} \pi_i P_{ik} C_{1k} \, . \tag{11.25}$$

where P_{ip} is the proportion of population i still not allocated after the values of the (p - 1)th feature have been obtained. For Fu's criterion we must take account of misallocation costs as well so the total cost for a given order of feature measurement is proportional to

$$\Sigma_i \Sigma_{k=1}^{k=p} \pi_i P_{ik} C_{1k} + \Sigma_{k=1}^{k=p} \Pi_1 \delta_k M_{12} + \Sigma_{k=1}^{k=p} \Pi_2 \varepsilon_k M_{21} \tag{11.26}$$

where δ_k and ε_k are the errors in allocation for objects from population 1 and 2 made at the kth stage. The cost of using an optimal fixed order of feature measurement can be expected to be greater than that of using an optimal varying order. It may, however, be computationally easier to work out the optimal fixed order.

11.5 Computation of an optimal fixed order of feature measurement when the only cost associated with a feature is measurement cost.

For the criterion introduced in sub-section 11.3.3 we can use the branch and bound algorithm to attempt to reduce the required number of calculations (Hand, 1981a). We can build an inverted tree with as many roots as there are features. For a node at level k in this tree, where level k is the number of levels down from the roots including the roots, the first k features are in the same order for the nodes below. Given a cost for a particular order we can then work down the tree calculating the cost of the discrimination so far for the first k features and as soon as total cost exceeds the current minimum cost no examination of descendent nodes is necessary. For this criterion the values of the $z_{l,k}$ and $z_{u,k}$ need not be calculated until they are required. The branch and bound strategy is not possible for Fu's criterion because the limits of integration differ for every order of all the features. Further all the $z_{l,k}$ and $z_{u,k}$ must be calculated before the evaluation of an optimal fixed order can start.

11.6 A sub-optimal fixed order of feature measurement when the only cost associated with a feature is measurement cost.

For a large number of features an optimal fixed order of feature measurement may be too computationally expensive to calculate. We can consider instead stepwise approaches to obtaining a fixed order of feature measurement. For the criterion introduced in sub-section 11.3.3 we may take the feature to be measured at the kth stage from those not measured so far as that which maximises

$$c_k^{-1}\Sigma_i \pi_i P_{ik} \quad . \tag{11.27}$$

This criterion may be interpreted as choosing as the feature to be measured next that which gives maximum additional allocation per unit cost. To take account of misallocation costs we may take the next feature to be measured as that which minimises expression (11.26) where the kth feature is regarded as the last.

11.7 Artificial examples

To illustrate the theory described for two known multivariate Normal distributions with equal covariance matrices the following artificial example is used. The population location vectors take the values

$$\underline{\mu}_1^T = [0, 0, 0]$$

and

$$\underline{\mu}_2^T = [0.25, 0.50, 0.75]$$

and the common covariance matrix is taken to be

$$\underline{\Sigma} = \begin{bmatrix} 1 & 0.1 & 0.5 \\ 0.1 & 1 & 0.9 \\ 0.5 & 0.9 & 1 \end{bmatrix}.$$

The measurement costs of each feature are taken to be $c_1 = 1$, $c_2 = 2$ and $c_3 = 3$. Both misallocation costs are taken as equal to 20 and the prior probabilities are assumed equal. It is assumed that there is no cost for the calculation of the discriminant scores. For the criterion introduced in sub-section 11.3.3 an additional misallocation error of 0.01 for each population is regarded as allowable. The stopping rule for Fu's criterion was that both sides of (11.17) and (11.18) agreed to 2 decimal places at each stage. All integrations were done to 2 decimal place accuracy. Equations (11.19) and (11.20) and (11.22) and (11.23) were solved iteratively. This was done by setting to zero the right-hand side of each pair of equations to obtain values of $z_{l,1}$ and $z_{u,1}$ and $z_{l,k}$ and $z_{u,k}$ and then iteratively altering $z_{l,1}$ and $z_{u,1}$ and $z_{l,k}$ and $z_{u,k}$ to satisfy the equations. This method converged for this example. Convergence was defined to be agreement to 4 decimal places with both intended additional errors. All integrations were done to 4 decimal place

accuracy.

The six possible orders of the three features and their associated costs for the two criteria are given in Table 11.1 . The values for the two criteria are not comparable because for Fu's criterion measurement and misallocation costs are assumed to be commensurable whilst for the criterion given in sub-section 11.3.3 this is not the case. For both criteria the ordering of the three features 312 gives the lowest cost discrimination procedure. The 'usual' stepwise forward selection procedure referred to in Table 11.1 is that given by choosing the feature to be measured next as that which maximises the Mahalanobis distance between the mean vectors. The sub-optimal fixed orders obtained are given in Table 11.2 . The branch and bound solution for the optimal fixed order for the criterion given in sub-section 11.3.3 is shown in Figure 11.1 .

The branch and bound solution in Figure 11.1 shows that taking the order of the features given by the sub-optimal procedure as a starting point no reduction in computation may be achieved. This is because only the nodes at level 3 in the tree give discrimination costs greater than the cost of the ordering given by the sub-optimal procedure.

11.8 A minimum cost version of criterion 11.3.3 for a fixed order of feature measurement.

Instead of deciding for ease of computation to divide the additional allowable error per class evenly between the (p - 1) intermediate stages for the criterion given in sub-section 11.3.3 , we may seek the division of the additional error which minimises (11.25) for a particular feature order. This means finding the values of α_k and β_k, k=1,...,(p - 1) which minimise (11.25) subject to $\Sigma_k\alpha_k = \alpha$ and $\Sigma_k\beta_k = \beta$ for each feature order. The lowest cost fixed order of feature measurement would then be the lowest cost achieveable for the given allowable additional errors for each class.

11.9 Additional error versus cost of discrimination for an optimal fixed order of feature measurement when the only cost associated with a feature is measurement cost.

If it is difficult to specify a maximum additional allowable error for each class it may be preferable instead to calculate the cost of discrimination for an

Table 11.1

Cost of sequential discrimination for each order of feature measurement for example in section 11.7 .
(Cost is proportional to number given in table.)

Ordering	Fu's criterion	Section 11.3.3 criterion
123	9.70	5.62
132	9.71	5.56
213	9.58	5.57
231	9.64	5.58
312	9.53	5.43
321	9.55	5.47

Feature order given by usual forward stepwise selection procedure.

321

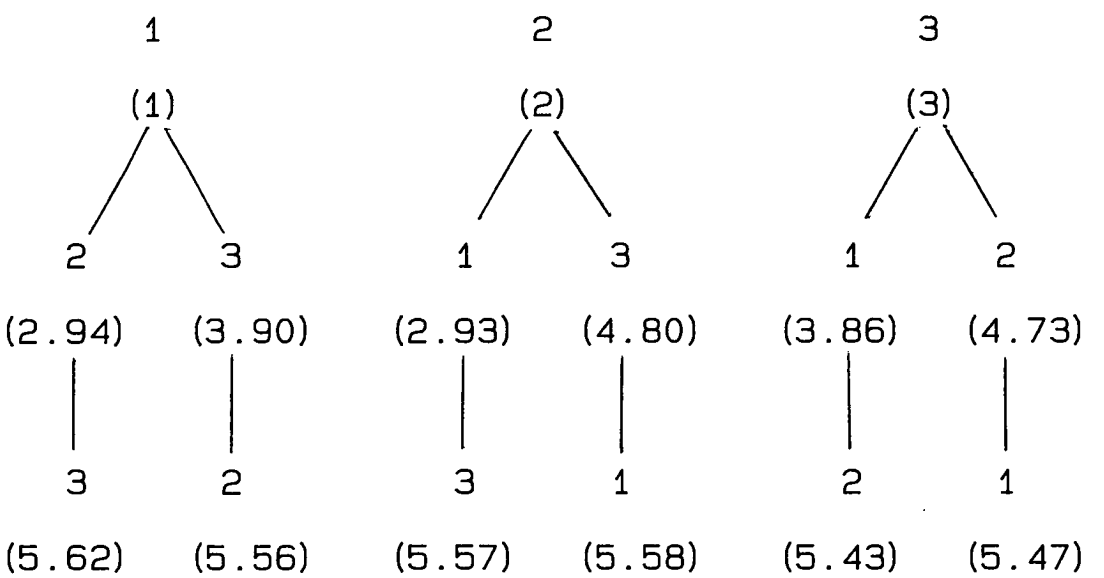
Table 11.2

Orders given by sub-optimal procedures for example given in section 11.7 .

Fu's criterion	Sub-section 11.3.3 criterion
123	312
132	
Usual forward stepwise selection procedure	
321	

Figure 11.1

Branch and bound solution to finding an optimal fixed order of feature measurement for the example given in section 11.7 and the criterion in sub-section 11.3.3 .
(Feature measured at kth stage given at level k with lines joining features measured so far. Discrimination cost so far proportional to number in brackets.)



optimal fixed order of feature measurement for various specified additional errors for each class. A 'best' choice might be then made from amongst these alternatives.

11.10 Including the cost of calculating discriminant scores in the cost associated with a feature for an optimal fixed order of feature measurement.

If the cost associated with a feature is to include the cost of calculating the discriminant scores (assumed commensurable with feature measurement cost) and these are calculated immediately after it has been measured, then it may not be optimal to calculate discriminant scores after every feature measurement. Consequently, for each fixed order of feature measurement there is the choice of calculating or not calculating the discriminant scores after each feature is measured. This gives $2^p - 1$ possibilities for each fixed order of feature if discriminant scores are to be calculated at least once. The criteria for early allocation of an object described in section 11.3 may be re-defined for each of these possibilities so that discriminant scores are calculated only after certain features have been measured. If it is assumed that at least some objects will still not be allocated when the values of the last but one feature have been obtained for the modification of the criterion introduced in sub-section 11.3.3 then discriminant scores must be calculated after all feature values are obtained for these objects. This reduces the $2^p - 1$ possibilities described above to 2^{p-1}. One of these possibilities corresponds to the usual one-stage discrimination procedure. For the modification of Fu's criterion, p of the $2^p - 1$ possibilities correspond to one-stage discrimination procedures using a subset of the features or all the features. When discriminant scores are calculated after a feature is measured the cost is included in the cost associated with that feature, otherwise it is omitted from the cost associated with a feature. Values of $z_{l,k}$ and $z_{u,k}$ for the case of two known multivariate Normal populations with equal covariance matrices can be derived in a similar manner to that described in sub-sections 11.3.2 and 11.3.4 for the sequential procedures with each fixed order of feature measurement. The proportion P_{ik} in expressions (11.25), (11.26) and (11.27) is defined as identically equal to zero if discriminant scores are not calculated after the kth feature is measured. For the modification of the criterion introduced in sub-section 11.3.3 the additional error might be divided equally amongst the

number of intermediate stages at which discriminant scores are calculated.

11.11 An empirical approach for the criterion in sub-section 11.3.3 .

A drawback to the above theory is that the distribution parameters are assumed to be known. For sample sizes large relative to p this need not matter. However, an alternative approach given a training data set and a test data set is to estimate the distributions of discriminant scores for those objects in the test set which would be correctly allocated and incorrectly allocated with all p features when only $k < p$ features are measured, for each class. For $k = 1$ all the objects in the test set that would be correctly or incorrectly allocated with p features are used but for $k = 2,\ldots,(p - 1)$ objects in the test set that would be allocated with less than k features need to be omitted. Numerical integration of the distributions will then provide suitable estimates of $z_{l,1}$ and $z_{u,1}$ and $z_{l,k}$ and $z_{u,k}$ in equations (11.19) and (11.20) and equations (11.22) and (11.23). Additionally, the distribution of discriminant scores using all features of objects still not allocated when the values of the last but one feature have been obtained needs to be estimated for each class. The estimated cost of a particular feature order may then be obtained by integrating over the two distributions obtained at each stage, except the last, to estimate the proportion of each population which can be allocated at each stage for each class. An optimal or sub-optimal feature order may be calculated as described above.

11.12 Sequential discrimination between artefacts and cells for cervical smears.

To illustrate the empirical approach of estimating directly the distributions of z_k for objects that would and would not be correctly allocated using all p features not yet allocated at stage k for each class, using training and test data sets, the linear discriminant function derived from assumptions of multivariate Normality and equal covariance matrices is used sequentially to distinguish between artefacts and cells in cervical smear specimens. The error rates obtained are worse than for the "box" discrimination method with joint ad-hoc setting of upper and lower limits for each feature mentioned in chapter 2. However, this example illustrates the control of the additional error introduced by sequential discrimination.

For the data set used here, described in chapter 3, features were measured in a fixed order of 3 groups of features. It is assumed that the cost of calculating discriminant scores is unimportant relative to feature measurement costs so that the only cost associated with a feature is measurement cost. The decision rule used when all features have been measured was to allocate an object as an artefact if its estimated posterior probability of being an artefact was greater than 0.01 . This gives a reasonable point on the plot of artefact errors versus cell errors. The data set of 92 specimens was divided into two parts of equal numbers of specimens so that the two distributions of discriminant scores at each stage except the last and the distribution of scores after all features have been measured for objects not previously allocated could be estimated for each class using the second part of the data set. It should be noted that each stage corresponds to obtaining the values for a group of features rather than for one feature at a time as described in section 11.3 . The distributions of discriminant scores were estimated using kernel density estimation with a Normal kernel and an adaptive smoothing parameter. The calculation of the bandwidth factors was as described in chapter 8 with the initial smoothing parameter given by the robust estimate (Silverman, 1986, page 48)

$$h = 0.9*An^{-1/5} \qquad (11.28)$$

where A = min(standard deviation, interquartile range/1.34) and n is the sample size. The iterative method of solution described in section 11.7 for the solution of equations (11.19) and (11.20) and (11.22) and (11.23) was also used here for their empirical equivalents and again converged. The number of points used in each numerical integration was the same and was increased until the results gave the intended increase in error compared with the one-stage discrimination procedure.

Table 11.3 shows the proportions of cells and artefacts in the test set that may be allocated at the intermediate stages and the final stage for an intended additional error of 0.1% for each population when the additional error is divided evenly between the two intermediate stages. As discussed above an optimal division of the additional error between the two stages could be sought if this

Table 11.3

Proportion of artefacts and cells that may be allocated after each group of features has been measured using sequential approach described in section 11.12 .

Group	Artefacts	Cells
1	0.741	0.690
2	0.208	0.235
3	0.051	0.075

Error rates for sequential approach.

Artefacts	Cells
0.267	0.690

Error rates for use of all features at once in linear discriminant function.

Artefacts	Cells
0.266	0.689

was thought worthwhile.

Chapter 12
Conclusions.

12.1 Introduction.

In this chapter the results of the work presented in this thesis are summarised and suggestions made for possible future work.

12.2 Review of results for the automated allocation of human chromosomes.

Three new procedures for modelling between-cell variation were presented in chapter 4. These were:

1. The transformation of each feature to marginal Normality when cell and class effects were allowed for in a linear model on the transformed scale. The cell effect was then removed on the transformed scale before the use of a discrimination method based on multivariate Normality.
2. The regression of size-related features on an index of size for the cell, within each chromosome class.
3. The division of cells into three classes according to the degree of contraction of the chromosomes with different sets of discriminant functions for each type of cell.

The estimated error-rates obtained under these three new procedures for three data sets, using Estimative multivariate Normal discrimination with a common covariance matrix per Denver group, showed that there is no consistent evidence of better performance than under the current normalisation. There was also evidence from two of the data sets that the division through by the within-cell standard deviation, currently done for some features, increased the error-rate.

In chapter 5 six methods of combining class information on variability in multivariate Normal discrimination were considered for the automated allocation of human chromosomes. Compared with the use of unrelated covariance matrices, these six methods have the advantages of reducing the number of parameters in the predicted densities and the number of calculations required to allocate the chromosomes in a cell for the numbers of features considered. The results obtained for five human chromosome data sets

showed that the trade-off between computational time and estimated error-rate can be improved compared with that obtained by the estimation of unrelated covariance matrices. For two of the five data sets some of these methods of combining class information gave the lowest estimated error-rates. It is conjectured that this is because the bias in estimation of the predicted distributions is compensated for by the reduction in sampling variation compared with methods assuming unrelated covariance matrices.

The idea of reducing the number of parameters in multivariate Normal discrimination was further explored in chapter 6. The covariance structures of individual class covariance matrices and assumed common covariance structures for groups of classes were modelled using covariance selection models. Results for the sets of covariance selection models obtained for the five data sets used in chapter 5 show that these sets of models can provide candidate procedures for the trade-off of estimated error-rate against computational time.

In chapter 7 some two-stage procedures for the calculation of discriminant scores in multivariate Normal discrimination were presented. The main motivation for this was to attempt to save computational time. At the first stage, a subset of features was used to eliminate improbable classes before a second stage in which all features were used to make a final allocation. Comparison with single-stage procedures using three data sets showed that it is possible to greatly reduce the expected computational time required to allocate the chromosomes in a normal cell for no increase in the estimated error-rate.

The application of classification trees, nearest neighbour discrimination, kernel density discrimination and logistic discrimination to the automated allocation of human chromosomes was considered in chapter 8. Results were obtained for the five data sets used in chapters 5 and 6 for particular versions of the three non-parametric methods. The results obtained for the classification trees were much worse than those for the multivariate Normal discrimination procedures described in chapter 5. No attempt was made to consider the trade-off between computational time and estimated error-rate because the estimated error-rates were so much bigger than those for the multivariate Normal discrimination procedures. The nearest-neighbour and kernel density discrimination procedures also did not give lower estimated error-rates than

did the multivariate Normal discrimination procedures. Again, no attempt was made to consider explicitly the trade-off between estimated error-rate and computational time. This is because it appeared unlikely that a procedure competitive with the multivariate Normal procedures would be obtainable. Finally, the amount of c.p.u. time required to estimate the parameters in logistic discrimination on a powerful computer for even a very small number of features seems excessive for its use to be seriously considered at the present time.

In chapter 9 non-parametric and multivariate Normal models were considered for the probabilities of band transition sequences derived from the sequence of light and dark bands along a chromosome. This is an alternative approach to chromosome allocation to that used in chapters 4, 5, 6, 7 and 8 where mainly 'global' features were used. 'Global' features such as the sums of weighted density profiles measure a property obtained from the entire chromosome whilst 'local' features such as peak density of staining in a segment measure a property from a part of the chromosome. The non-parametric models in this case gave lower estimated error-rates than the multivariate Normal models. However, a re-definition of peak density of staining and the difference in density of staining between a peak and its next "valley" would make the multivariate Normal models more plausible. This is because, currently, if there is no peak density of staining in one of the fourteen segments a zero is recorded for both the peak density and the density difference. Defining a peak density and density difference with continuous measurement for each segment would give data more appropriate to the use of multivariate Normal models. The results for the non-parametric models were still worse than those for the multivariate Normal discrimination procedures which used mainly 'global' features.

None of the results of Chapters 4-9 used a re-arrangement procedure for revising the isolated allocations in order to achieve a normal karyotype. The effects of re-arrangement for the most promising procedures needs to be assessed.

12.3 Possible future work for the automated allocation of human chromosomes.

Because of errors in the pattern recognition process it is apparent that

some feature values may have large errors and that this will affect the statistical discrimination. For example, the location of the centromere of a chromosome may be incorrectly determined. A possible approach would be to try to detect when these features are incorrect by reference to values of features which are less prone to large errors. If features are detected as having incorrect values then the allocation could be done by the operator or else these feature values could be excluded from those used in the statistical discrimination.

Related to the above idea of unreliable data is that of looking at only good quality metaphases. If segmentation can be made to be completely automatic or the operator is required to look at a number of cells anyway, chromosomes in several cells could be allocated. The cells could then be presented to the operator in descending order of probability that all chromosomes in the cell were correctly allocated. This probability might be assessed by multiplying together the posterior probabilities of class membership for the individual chromosome allocations.

A different approach would be to attempt to combine statistical and knowledge-based approaches. This might be done by using the latter to rule out certain classes as impossible for a particular chromosome because of the posession of certain local features, then using statistical discrimination to decide between the remaining classes. Alternatively, the statistical discrimination could be used to rule out classes with a very low estimated probability before a knowlege-based approach is used.

So far allocation based on weighted density profiles and on band-transition sequences has been kept distinct. Lower error-rates might be achieveable by using allocation based on both sets of features.

As noted in the previous section the effects of re-arrangement procedures for revising the isolated allocations in order to achieve a normal karyotype might be studied for some of the procedures considered in this thesis.

Finally, information on the similarity of homologues (the pair of chromosomes in each of the autosomal classes and the pair of sex chromosomes for a female, in a normal cell) might be incorporated into the model used for discrimination to try to improve allocation error-rates.

12.4 Review of results for the automated allocation of cervical smears.

In chapter 10 the method proposed by Dawid and Skene (1979) to estimate observer error-rates for allocations of the same samples by a number of observers was used to estimate a consensus probability of a cervical smear being abnormal. This method was used because of the evident disagreement between four cytologists. The logit transformations of the consensus probabilities were then regressed on features available automatically from the operation of an object discrimination procedure and also on these features and further features derived after the intervention of an operator. These two multiple regression equations were used in a sequential approach such that the second equation was used to predict the logit of a consensus probability if the probability derived from the predicted logit using the first equation lay between certain thresholds. Otherwise the probability derived from the predicted logit for the first equation was used. The results indicated that machine performance with error-rates similar to those of one of the cytologists was possible for a given amount of operator intervention. Setting the estimated false-negative error-rate to be as low as those of the other three cytologists causes estimated false-positive error-rates higher than those of these cytologists.

In chapter 11 the sequential use of linear discriminant functions for distinguishing between artefacts and cells in a cervical smear specimen was considered. The empirical procedure used demonstrated the control of the additional error-rate compared with the use of a linear discriminant function based on all the features. However, the result was worse than that obtainable by using the "box" discrimination procedure currently implemented in the system.

12.5 Possible future work for the automated allocation of cervical smears.

Given that improvement in the object allocation may be expected to improve the specimen allocation it would seem worthwhile exploring the former further. In particular, it may be worth trying to find a systematic multivariate approach for the elimination of artefacts in stages. The major constraint, however, is that such an approach must be computationally very

fast. A possible approach would be to obtain kernel density estimates for the artefacts and for each type of cell using a computationally quick kernel, to reduce drastically the number of objects for the kernel density estimates (Hand, 1981b, pages 30-31) and to use a fast algorithm for finding the objects with feature vectors within a certain distance of that object to be allocated (Friedman, Bentley and Finkel, 1977). The empirical sequential approach outlined in chapter 11 might then be used to distinguish between artefacts and other types of cell in order to save feature measurement time.

At the specimen level it would seem worthwhile obtaining more data so that cytologist and machine error-rates for three rather than two categories of abnormality of cervical smear could be estimated. These three categories could be normal, moderately abnormal and severely abnormal.

12.6 Review of theoretical results derived.

In chapter 11 the computation required by Fu's dynamic programming approach to obtaining an optimal varying order of feature measurement was briefly reviewed. This approach assumes:

1. The only costs are for feature measurement and misallocation.
2. These costs are commensurable.
3. The class distributions are known.

An alternative approach to that of obtaining an optimal varying order of feature measurement, when the feature order is free to vary, was proposed. This approach, which may require less computation to obtain a solution, is to define an optimal fixed order of feature measurement. Results were obtained for two known multivariate Normal populations with equal covariance matrices for Fu's criterion and a new criterion, for early allocation of an object when the feature order is fixed. For both criteria the only costs are assumed to be those of feature measurement and misallocation. Because the evaluation of an optimal fixed order of feature measurement may be computationally too demanding sub-optimal approaches to obtaining a fixed order of feature measurement were also considered for these two criteria. The evaluation of an optimal fixed order of feature measurement was considered when the cost of the calculation of the discriminant scores after a feature has been measured is included in the measurement cost of a feature. The two criteria for early allocation of an object need to be re-defined and it may not be optimal to calculate the discriminant

scores after the value for each feature has been obtained. It may also be the case that a one-stage procedure gives a lower cost procedure than a sequential procedure. Finally, an empirical approach for the use of the new criterion for early allocation of an object was advocated and illustrated on the cervical smear data.

12.7 General conclusion.

The methods used in this thesis could have applications to other real-time discrimination problems and to discrimination problems with no real-time component. For example, the reduction in number of parameters for multivariate Normal discrimination described in chapters 5 and 6 may improve the trade-off between computational time and error-rate in other real-time problems. The reduction in number of parameters may also, however, improve error rates in discrimination problems where computational time is not an important consideration.

REFERENCES

ABRAMSON, I.S. (1982) On bandwidth variation in kernel estimates - a square-root law.
The Annals of Statistics, 10, 1217-1223.

AITCHISON, J. and BEGG, C.B. (1976) Statistical diagnosis when basic cases are not classified with certainty.
Biometrika, 63, 1-12.

AITCHISON, J. and DUNSMORE, I. (1975) Statistical prediction analysis
Cambridge University Press.

AITCHISON, J., HABBEMA, J.D.F. and KAY, J.W. (1977) A critical comparison of two methods of statistical discrimination.
Applied Statistics, 26, 15-25.

ALBERT, A. and HARRIS, E.K. (1987) Multivariate interpretation of clinical laboratory data
New York: Marcel Dekker, Inc.

ALBERT, A. and LESAFFRE, E. (1986) Multiple group logistic discrimination. in
Statistical methods of discrimination and classification: Advances in theory and applications edited by S.C. Choi, 209-224.
New York; Oxford: Pergamon Press.

ANDERSON, J.A. (1972) Separate sample logistic discrimination.
Biometrika, 59, 19-35.

ANDERSON, T.W. (1984) An introduction to multivariate statistical analysis
Second edition, New York; Chichester: John Wiley.

BOOK, J.A. and 16 others. (1960) A proposed standard system of nomenclature of human mitotic chromosomes.
Lancet, 1, 1063-1065.

BREIMAN, L., MEISEL, W. and PURCELL, E. (1977) Variable kernel estimates of multivariate densities.
Technometrics, 19, 135-144.

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. and STONE, C.J. (1984) Classification and regression trees
Belmont, California: Wadsworth International.

CALIFORNIA STATISTICAL SOFTWARE INCORPORATED (1985)
Computer program CART
Lafayette, California.

CAROTHERS, A.D. (1987) Report on Charing Cross Experiment 4.
M.R.C. internal report, M.R.C. Human Genetics Unit, Edinburgh.

CAROTHERS, A.D. (1988) Re-analysis of Charing Cross Experiment 4 with classifier trained on reference diagnosis.
M.R.C. internal report, M.R.C. Human Genetics Unit, Edinburgh.

CAROTHERS, A.D., RUTOVITZ, D. and GRANUM, E. (1983)
An efficient multiple-cell approach to automatic aneuploidy screening.
Analytical and Quantitative Cytology, 5, 194-200.

CLARKSON, D.B. (1988a) A remark on Algorithm AS 211: The F-G diagonalization algorithm.
Applied Statistics, 37, 147-151.

CLARKSON, D.B. (1988b) A least squares version of Algorithm AS 211: The F-G diagonalization algorithm.
Applied Statistics, 37, 317-321.

DAWID, A.P. and SKENE, A.M. (1979) Maximum likelihood estimation of observer error-rates using the EM algorithm.
Applied Statistics, 28, 20-28.

DEGROOT, M.H. (1970) Optimal statistical decisions.
New York: McGraw-Hill.

DEMPSTER, A.P. (1972) Covariance selection.
Biometrics, 28, 157-175.

DEMPSTER, A.P., LAIRD, N.M. and RUBIN, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm.
Journal of the Royal Statistical Society, Series B, 39, 1-38.

EDWARDS, D. (1987) A guide to MIM
Research Report 87/1, Statistical Research Unit, Copenhagen University.

EFRON, B. (1979) Bootstrap methods: Another look at the jackknife.
The Annals of Statistics, 7, 1-26.

ERIKSEN, P.S. (1987) Proportionality of covariance matrices.
The Annals of Statistics, 15, 732-748.

FATTI, L.P. and HAWKINS, D.M. (1986) Variable selection in heteroscedastic discriminant analysis.
Journal of the American Statistical Association, 81, 494-500.

FLURY, B.N. (1984) Common principal components in k groups.
Journal of the American Statistical Association, 79, 892-8.

FLURY, B. (1988) Common principal components and related multivariate models
New York; Chichester: John Wiley.

FLURY, B.N. and CONSTANTINE, G. (1985) Algorithm AS211: The F-G diagonalization algorithm.
Applied Statistics, 34, 177-183.

FRIEDMAN, J.H., BENTLEY, J.L. and FINKEL, R.A. (1977) An algorithm for finding best matches in logarithmic expected time.
Transactions on mathematical software, 3, 209-226.

FRYDENBERG, M. and EDWARDS, D. (1989) A modified iterative proportional scaling algorithm for estimation in regular exponential families.
Computational Statistics and Data Analysis, 8, 143-153.

FU,K.S. (1968) Sequential methods in pattern recognition and machine learning
New York: Academic Press.

GERDES, T. (1979) Methods for normalisation of chromosome data
MSc dissertation, The Technical University of Denmark.

GERDES, T., MAAHR, J. and LUNDSTEEN, C. (1989) Automatic classification of human chromosomes. in
Proceedings of 11th European Workshop on automated cytogenetics, 34-47. Besse en Chandesse.

GNANADESIKAN, R. (1977) Methods for statistical data analysis of multivariate observations
New York: John Wiley.

GOODMAN, L. (1974) Exploratory latent structure analysis using both identifiable and unidentifiable models.
Biometrika, 61, 215-231.

GRANUM, E. (1982) Application of statistical and syntactical methods of analysis and classification to chromosome data. in
NATO ASI series: C81 Pattern recognition theory and applications edited by J.Kittler et al, 373-398.
Dordrecht: D. Reidel.

HABBEMA, J.D.F. (1979) Statistical methods for classification of human chromosomes.
Biometrics, 35, 103-118.

HAND, D.J. (1981a) Branch amd bound in statistical data analysis.
The Statistician, 30, 1-13.

HAND, D.J. (1981b) Discrimination and classification
Chichester; New York: John Wiley.

HAWKINS, D.M. and RAATH, E.L. (1982) An extension of Geisser's discrimination model to proportional covariance matrices.
The Canadian Journal of Statistics, 10, 261-270.

HILDITCH, C.J. and RUTOVITZ, D. (1972) Normalization of chromosome measurements.
Computers in Biology and Medicine, 2, 167-179.

HORA, S.C. (1980) Sequential discrimination.
Communications in Statistics Theory and Methods, A9(9), 905-916.

JAIN, A.K., DUBES, R.C. and CHEN, C.C. (1987) Bootstrap techniques for error rate estimation.
IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-9, 628-633.

KENWARD, M.G. (1987) A method for comparing profiles of repeated measurements.
Applied Statistics, 36, 296-308.

KIRBY, S.P.J., THEOBALD, C.M., PIPER, J. and CAROTHERS, A.D. (accepted for publication) Some methods of combining class information in multivariate Normal discrimination for the classification of human chromosomes.
Statistics in Medicine

KLEINSCHMIDT, P., LEE, C.W. and SCHANNATH, H. (1987) Transportation problems which can be solved by the use of Hirsch-paths for the dual problems.
Mathematical Programming, 37, 153-168.

KRZANOWSKI, W.J. (1988) Principles of multivariate analysis
Oxford: Clarendon Press.

LESAFFRE, E. and ALBERT, A. (1989) Partial separation in logistic discrimination.
Journal of the Royal Statistical Society, B, 51, 109-116.

LÖRCH, T., WITTLER, C., STEPHAN, G. and BILLE, J. (1989)
An automated chromosome aberration scoring system. in
Automation of cytogenetics edited by C.Lundsteen and J.Piper, 19-30.
Berlin: Springer-Verlag.

LUNDSTEEN, C. and GRANUM, E. (1979) Description of chromosome banding patterns by band transition sequences: A new basis for automated chromosome analysis.
Clinical Genetics, 15, 418-429.

LUNDSTEEN, C., GERDES, T., GRANUM, E., PHILIP, J.
and PHILIP, K. (1981) Automatic chromosome analysis:
II. Karyotyping of banded human chromosomes using band transition sequences.
Clinical Genetics, 19, 26-36.

LUNDSTEEN, C., GERDES, T. and MAAHR, J. (1986) Automatic classification of chromosomes as part of a routine system for clinical analysis.
Cytometry, 7, 1-7.

MANLY, B.F.J. and RAYNER, J.C.W. (1987) The comparison of sample covariance matrices using likelihood ratio tests.
Biometrika, 74, 841-847.

MCLACHLAN, G.J. (1986) Assessing the performance of an allocation rule in
Statistical methods of discrimination and classification: Advances in theory and applications. edited by S.C. Choi, 261-272.
New York; Oxford: Pergamon Press.

MORAN, M.A. and MURPHY, B.J. (1979) A closer look at two alternative methods of statistical discrimination.
Applied Statistics, 28, 223-232.

MORRISON, D.F. (1976) Multivariate statistical methods.
Second Edition, New York: McGraw-Hill.

NUMERICAL ALGORITHMS GROUP LIMITED (1988) N.A.G. Fortran library manual
Oxford.

PATERSON, A. and NIBLETT, T. (1982) ACLS user manual
Glasgow: Intelligent Terminals Ltd.

PIPER, J. (1986) Classification of chromosomes constrained by expected cell class size.
Pattern recognition letters, 4, 391-395.

PIPER, J. (1987) The effect of zero feature correlation assumption on maximum likelihood based classification of chromosomes.
Signal Processing, 12, 49-57.

PIPER, J. and GRANUM, E. (1989) On fully automatic feature measurement for banded chromosome classification.
Cytometry, 10, 242-255.

PIPER, J., GRANUM, E., RUTOVITZ, D. and RUTTLEDGE, H. (1980) Automation of chromosome analysis.
Signal Processing, 2, 203-221.

PIPER, J., TOWERS, S., GORDON, J., IRELAND, J. and MCDOUGALL, D. (1988) Hypothesis combination and context sensitive classification for chromosome aberration scoring. in
Pattern recognition and artificial intelligence.
edited by E.S. Gelsema and L.N. Kanal, 449-460.
North-Holland: Elsevier Science.

PORTEOUS, B.T. (1985) Improved likelihood ratio statistics for covariance selection models.
Biometrika, 72, 97-101.

SHEPHERD, B., PIPER, J. and RUTOVITZ, D. (1987) Comparison of ACLS and classical linear methods in a biological application.
in
Machine intelligence 11 edited by J.E. Hayes, D. Michie and J. Richards, 423-434.
Oxford University Press.

SILVERMAN, B.W. (1986) Density estimation for statistics and data analysis.
London: Chapman and Hall.

SPEED, T.P. and KIIVERI, H.T. (1986) Gaussian Markov distributions over finite graphs.
The Annals of Statistics, 14, 138-150.

THOMASON, M.G. and GRANUM, E. (1986) Dynamic programming inference of Markov networks from finite sets of sample strings.
I.E.E.E. Transactions on Pattern Analysis and Machine Intelligence, PAMI-8, 491-501.

TIMMERS, T. (1987) Pattern recognition of cytological specimens
Ph.D Thesis, University of Amsterdam.

TSO, M.K.S. (1989) Evaluation of the transportation approach to automatic karyotyping. in
Proceedings of 11th European workshop on automated cytogenetics, 92. Besse en Chandesse.

TSO, M.K.S. and GRAHAM, J. (1983) The transportation algorithm as an aid to chromosome classification.
Pattern recognition letters, 1, 489-496.

TUCKER, J.H. (1979) An image analysis system for cervical cytology automation using nuclear DNA content.
The Journal of Histochemistry and Cytochemistry, 27, 613-620.

UEBERSAX, J.S. and GROVE, W.M. (1990) Latent class analysis of diagnostic agreement.
Statistics in Medicine, 9, 559-572.

VAN DEN BERG, H.T.C.M., DE FRANCE, H.F., HABBEMA, J.D.F. and RAATGEVER, J.W. (1981) Automated selection of metaphase cells by quality.
Cytometry, 1, 363-368.

VAN VLIET, L.J., YOUNG, I.T., TEN KATE, T.K., MAYALL, B.H., GROEN, F.C.A. and ROOS, R. (1989) Athena: A Mackintosh-based interactive karyotyping system. in
Automation of cytogenetics edited by C. Lundsteen and J.Piper, 47-66.
Berlin: Springer-Verlag.

WERMUTH, N. and SCHEIDT, E. (1977) Algorithm AS105: Fitting a covariance selection model to a matrix.
Applied Statistics, 26, 88-92.

ZIELEZNY, M. and DUNN, O.J. (1975) Cost evaluation of a two-stage classification procedure.
Biometrics, 31, 37-47.